Cluster Analysis

Statistics 407, ISU

Definition

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.

Distance Measures

Let $\mathbf{X} = (X_1 \ X_2 \ \ldots X_p)'$ and $\mathbf{Y} = (Y_1 \ Y_2 \ \ldots Y_p)'$ be two points in p-space (two rows of a data matrix).

Euclidean Distance:

$$d(\mathbf{X}, \mathbf{Y}) = \sqrt{(\mathbf{X} - \mathbf{Y})'(\mathbf{X} - \mathbf{Y})} = \sqrt{(X_1 - Y_1)^2 + \ldots + (X_p - Y_p)^2}$$

Statistical Distance:

$$d(\mathbf{X}, \mathbf{Y}) = \sqrt{(\mathbf{X} - \mathbf{Y})' S^{-1} (\mathbf{X} - \mathbf{Y})}$$

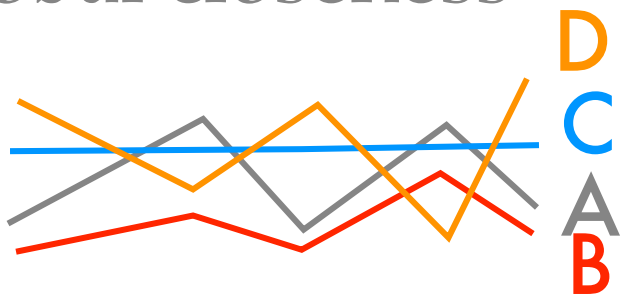
Both of these distance measures benefit from standardizing the variables first.

Distance Metrics

- Kendall tau distance: Order each variable. For all pairs of elements of the two points, count 1 for each pair which the ranks are in the same relationship (low, low; high, high) and 0 otherwise.
- Measures the association between two points, eg height value is often as highly ranked as weight value says that the two variables are positively correlated. Not as affected by outliers as raw data values.

Distance Metrics

- Pearson correlation: d=1-r
  - d=0 when r=1
- Pearson square correlation: $d=1-r^2$
  - d=1 when r=0, d=0 when r=1 or -1
- Measures the similarity in trend, rather than global closeness



A close to C by Euclidean, but B close to A by Pearson correlation. And D close to A,B by Pearson square correlation

- Also can be considered to be angular distance

Hierarchical Clustering

- Hierarchical algorithms sequentially fuse (or split) cases to make clusters.
- Process can be viewed using a dendrogram.
- The vertical heights of the dendrogram are used to decide how many clusters.

Linkage

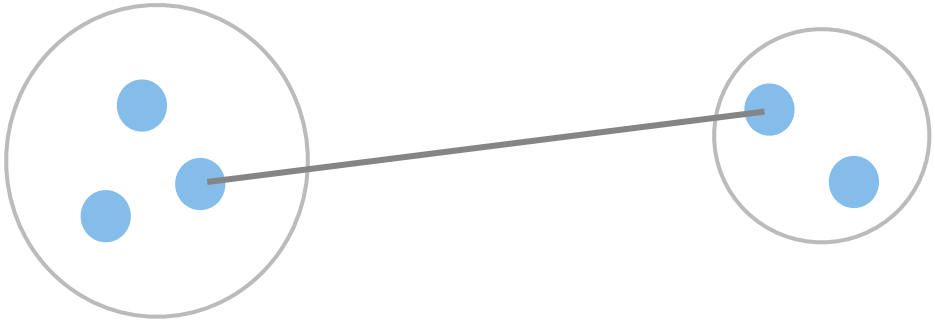
When a cluster is formed, containing two or more cases, there are now multiple ways to define the distance from the cluster to other clusters or cases. For example, we could define the distance from one cluster to another as the minimum interpoint distance, or the maximum interpoint distance or the average interpoint distance. These are called **linkage methods**. Each method changes the results of the cluster analysis.

Common linkage methods

The intercluster distance is described by:

- **Single:** the distance between the two **closest** points.
- **Complete:** the distance between the two **farthest** points.
- **Average:** the average of all the interpoint distance.
- **Centroid:** the distance between the two means.
- **Wards:** the smallest increase in the error sum of squares after fusing two clusters, like ANOVA.

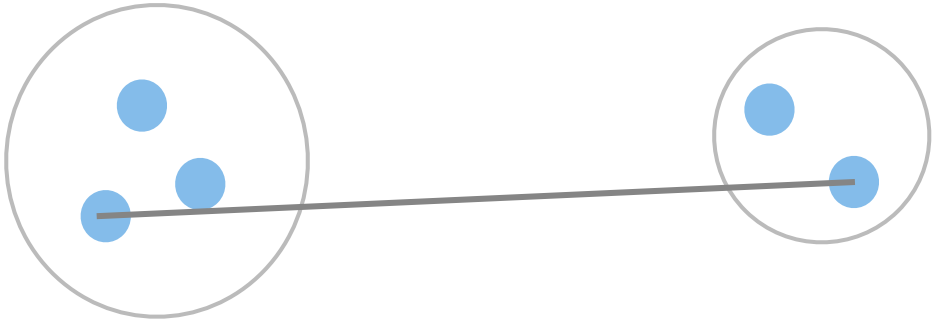
Single Linkage



Cluster 1 Cluster 2

Closest points define the intercluster distance

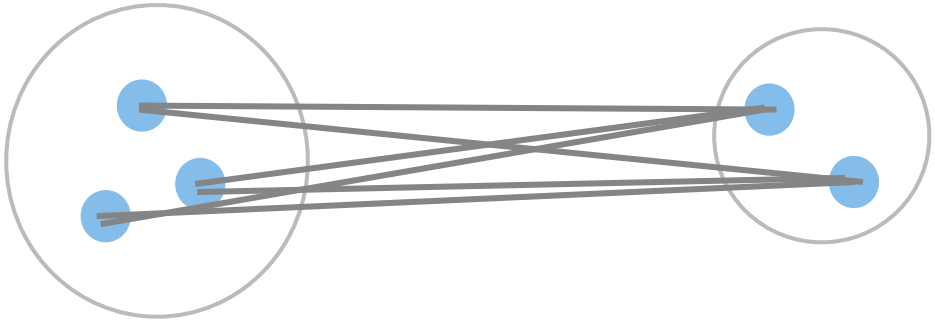
Complete Linkage



Cluster 1 Cluster 2

Farthest points define the intercluster distance

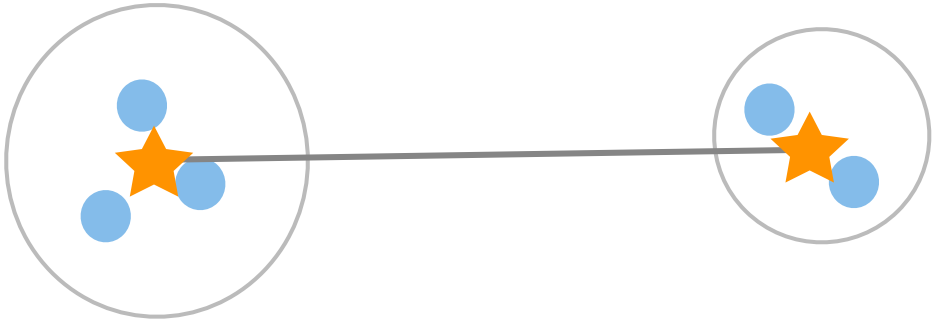
Average Linkage



Cluster 1 Cluster 2

Average of all of the distances defines the intercluster distance

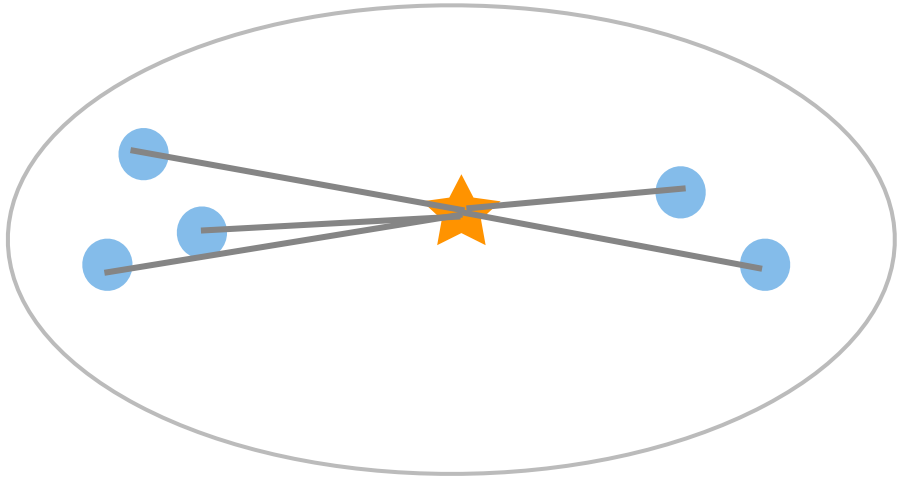
Centroid Linkage



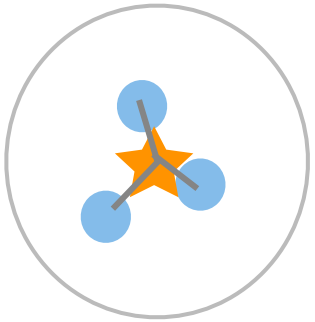
Cluster 1 Cluster 2

Distance between the cluster means defines the intercluster distance

Ward Linkage



One Cluster



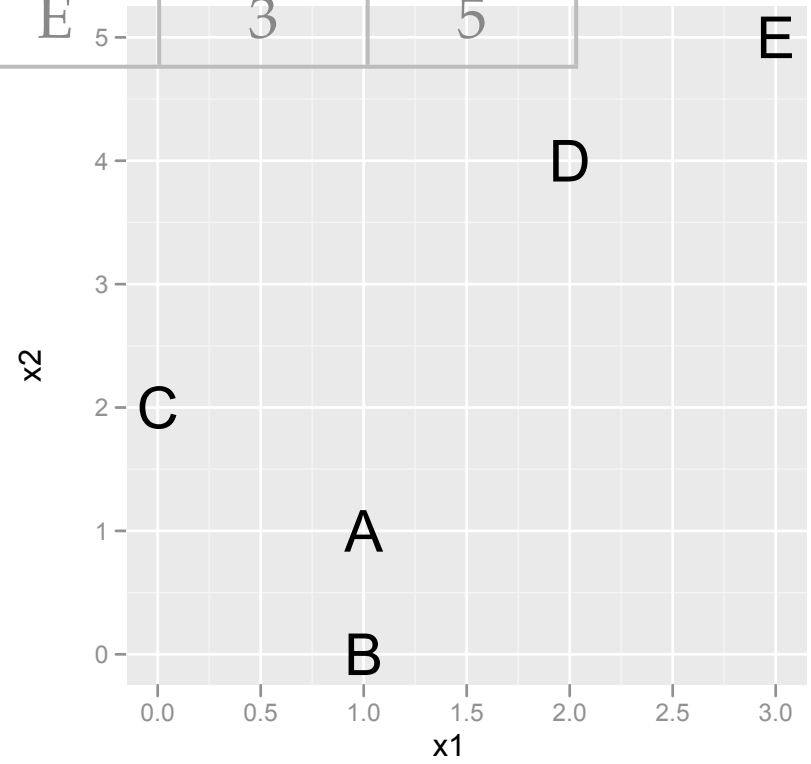
Cluster 1

Cluster 2

Ratio of sum of squared distance from means, between one cluster, and the two clusters defines the intercluster distance

Example

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5



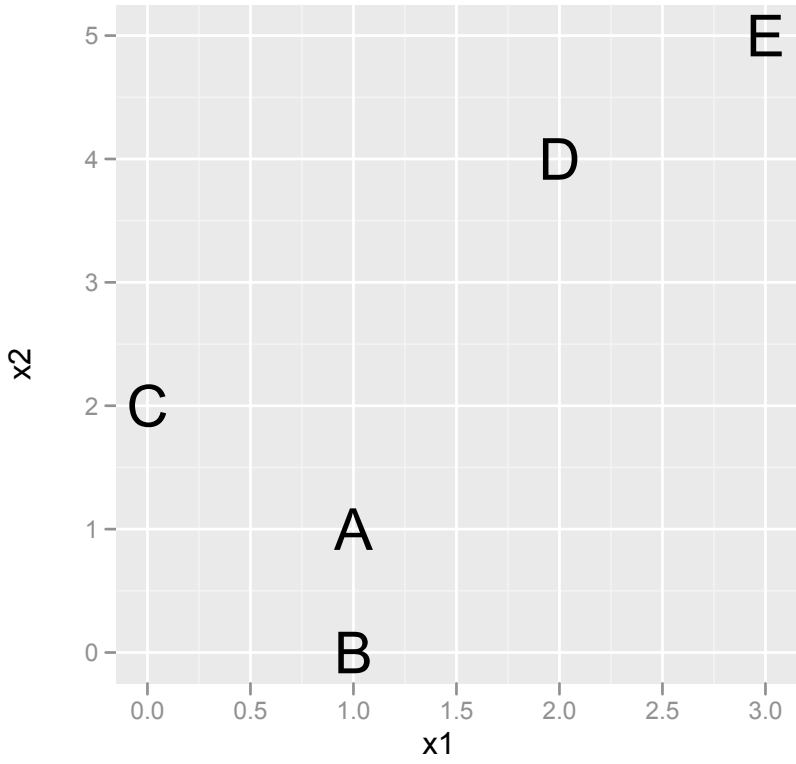
Euclidean distances

	A	B	C	D	E
A	0	1	1.4	3.2	4.5
B	1	0	2.2	4.1	5.4
C	1.4	2.2	0	2.8	4.2
D	3.2	4.1	2.8	0	1.4
E	4.5	5.4	4.2	1.4	0

Step 1.1

Join the two closest points into a cluster.

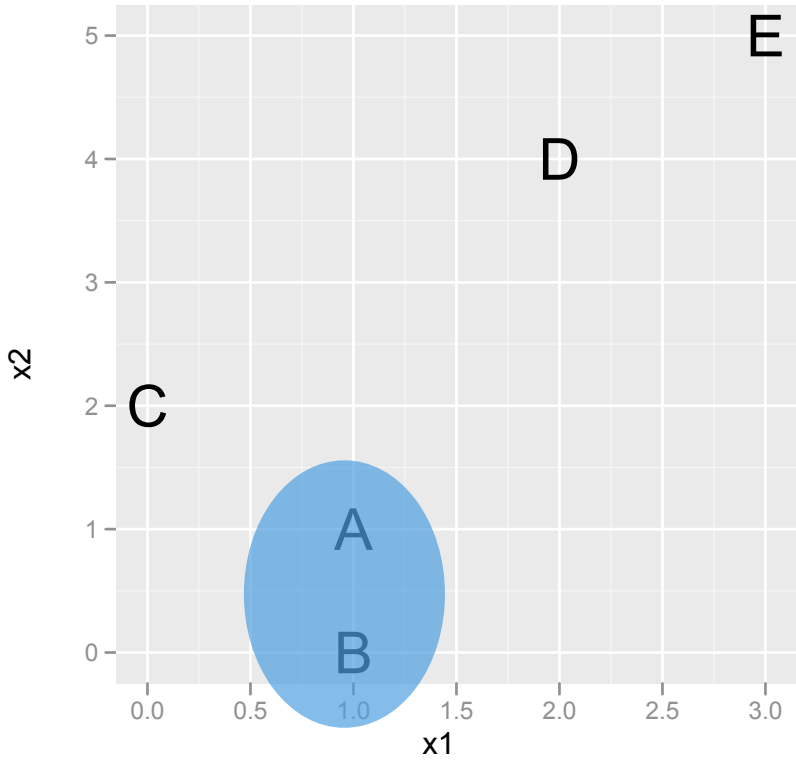
	A	B	C	D	E
A	0	1	1.4	3.2	4.5
B	1	0	2.2	4.1	5.4
C	1.4	2.2	0	2.8	4.2
D	3.2	4.1	2.8	0	1.4
E	4.5	5.4	4.2	1.4	0



Step 1.1

Join the two closest points into a cluster.

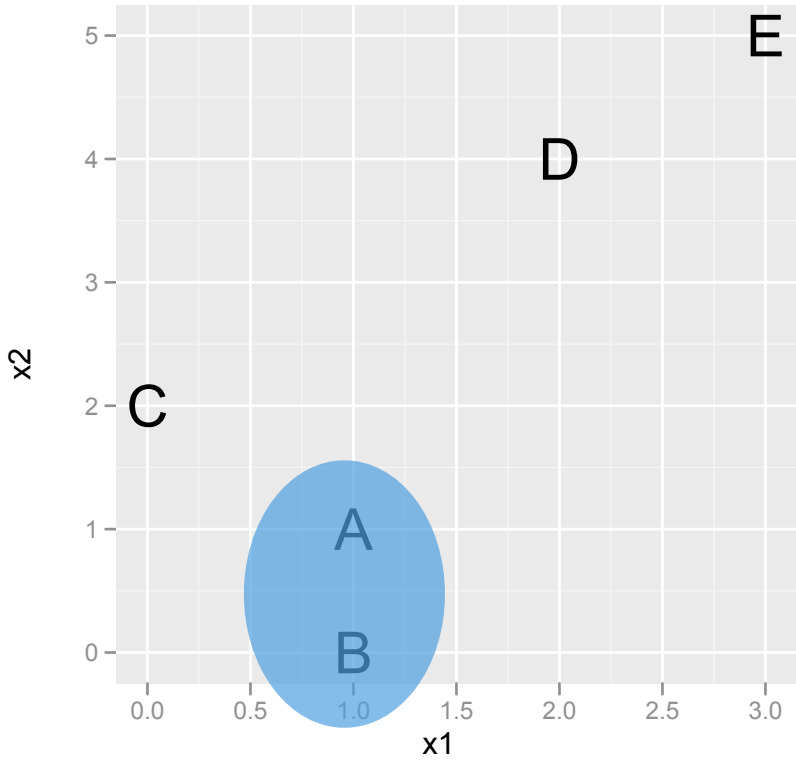
	A	B	C	D	E
A	0	1	1.4	3.2	4.5
B	1	0	2.2	4.1	5.4
C	1.4	2.2	0	2.8	4.2
D	3.2	4.1	2.8	0	1.4
E	4.5	5.4	4.2	1.4	0



Step 1.1

Join the two closest points into a cluster.

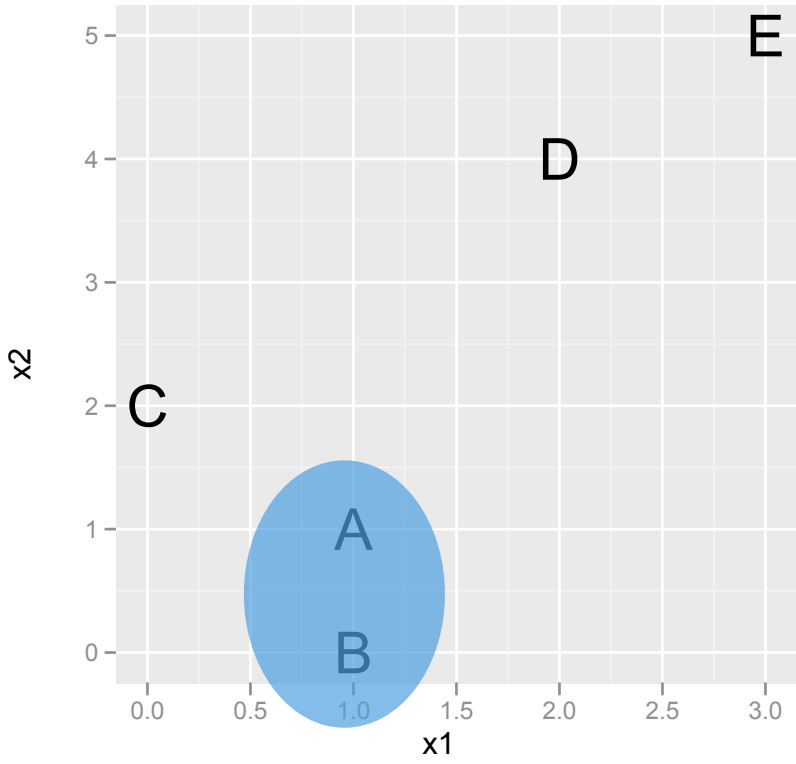
	A	B	C	D	E
A	0	1	1.4	3.2	4.5
B	1	0	2.2	4.1	5.4
C	1.4	2.2	0	2.8	4.2
D	3.2	4.1	2.8	0	1.4
E	4.5	5.4	4.2	1.4	0



Step 1.1

Join the two closest points into a cluster.

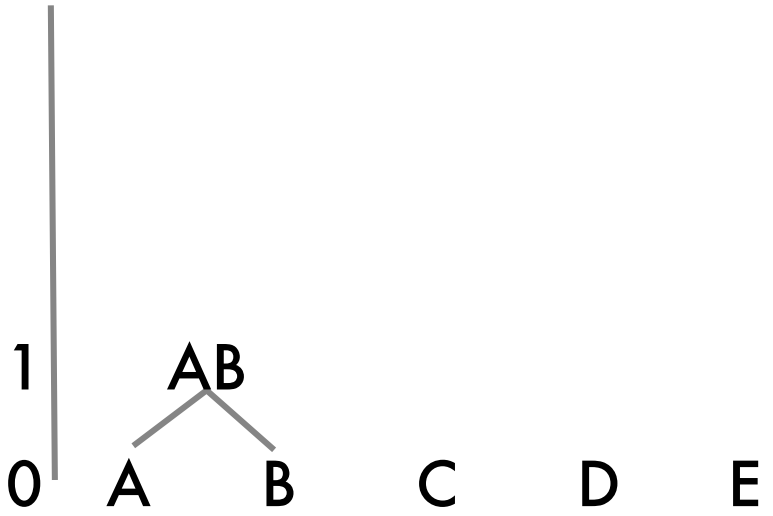
	A	B	C	D	E
A	0	1	1.4	3.2	4.5
B	1	0	2.2	4.1	5.4
C	1.4	2.2	0	2.8	4.2
D	3.2	4.1	2.8	0	1.4
E	4.5	5.4	4.2	1.4	0



Step 1.2

Reduce the distance matrix, using the linkage methods. Draw the dendrogram.

	AB	C	D	E
AB	0	1.8	3.6	4.9
C	1.8	0	2.8	4.2
D	3.6	2.8	0	1.4
E	4.9	4.2	1.4	0

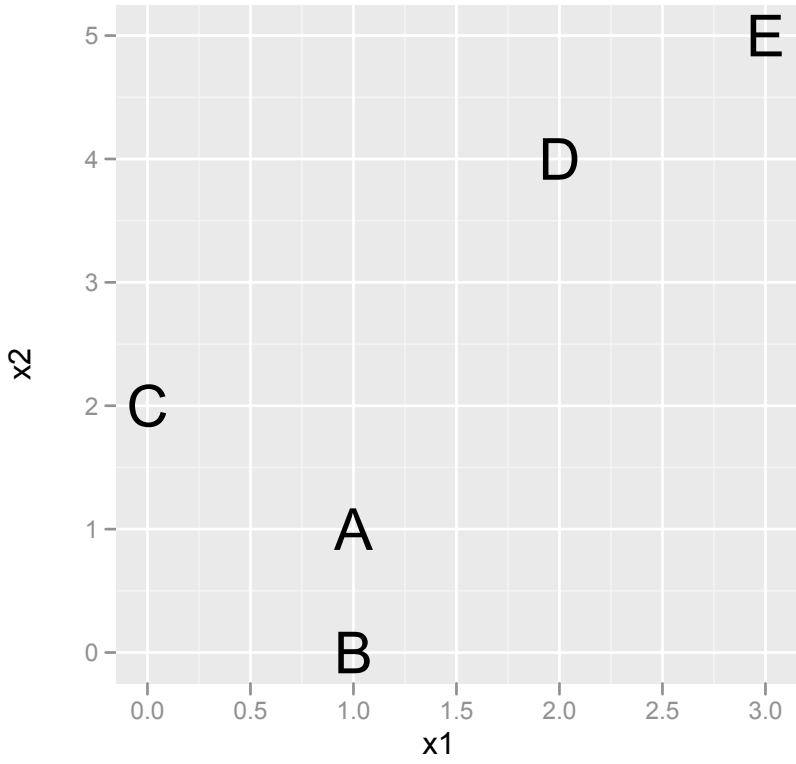


Average linkage used.

Step 2.1

Join the two closest points into a cluster.

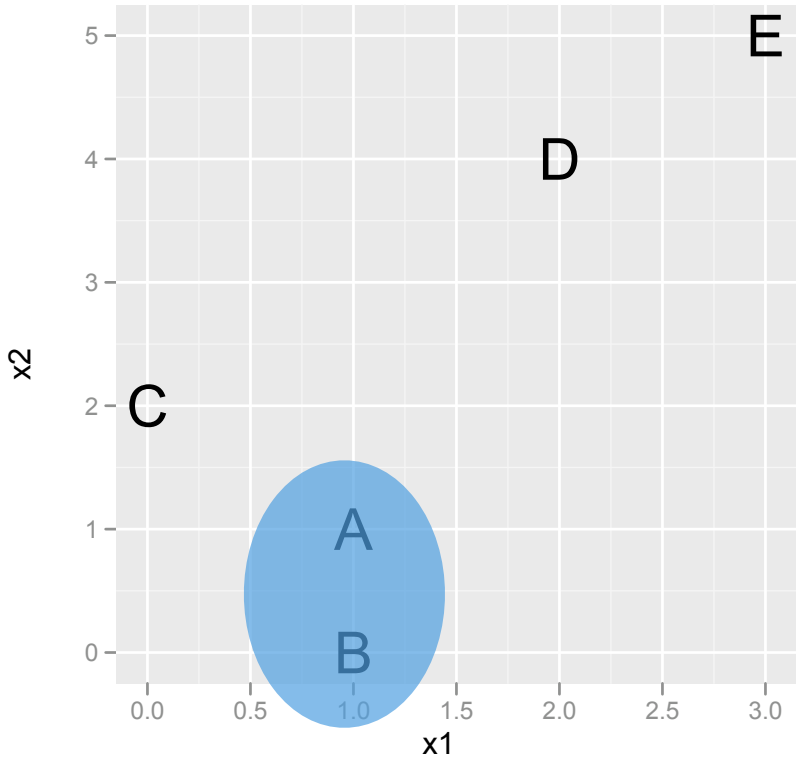
	AB	C	D	E
AB	0	1.8	3.6	4.9
C	1.8	0	2.8	4.2
D	3.6	2.8	0	1.4
E	4.9	4.2	1.4	0



Step 2.1

Join the two closest points into a cluster.

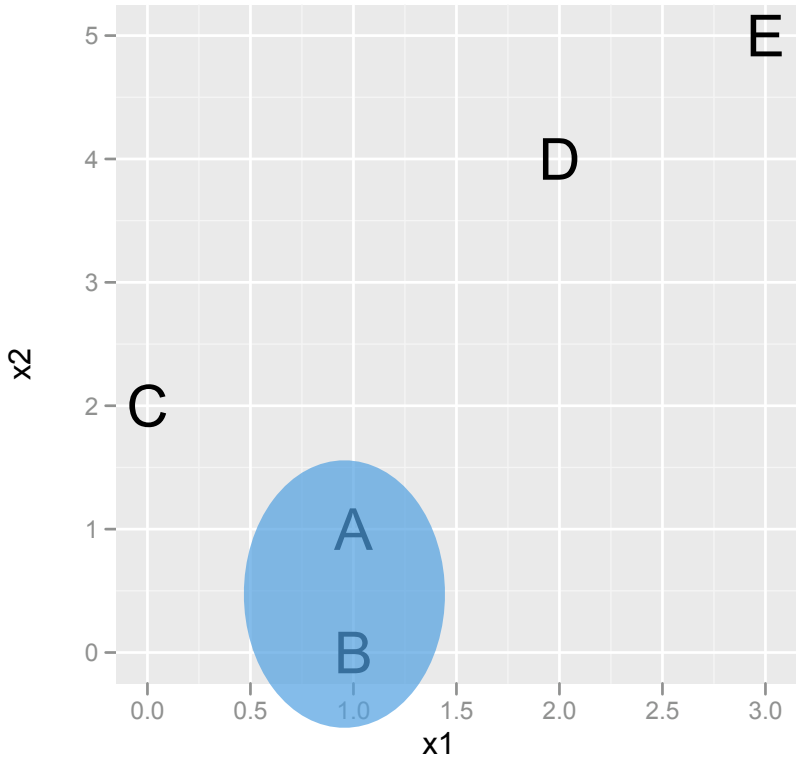
	AB	C	D	E
AB	0	1.8	3.6	4.9
C	1.8	0	2.8	4.2
D	3.6	2.8	0	1.4
E	4.9	4.2	1.4	0



Step 2.1

Join the two closest points into a cluster.

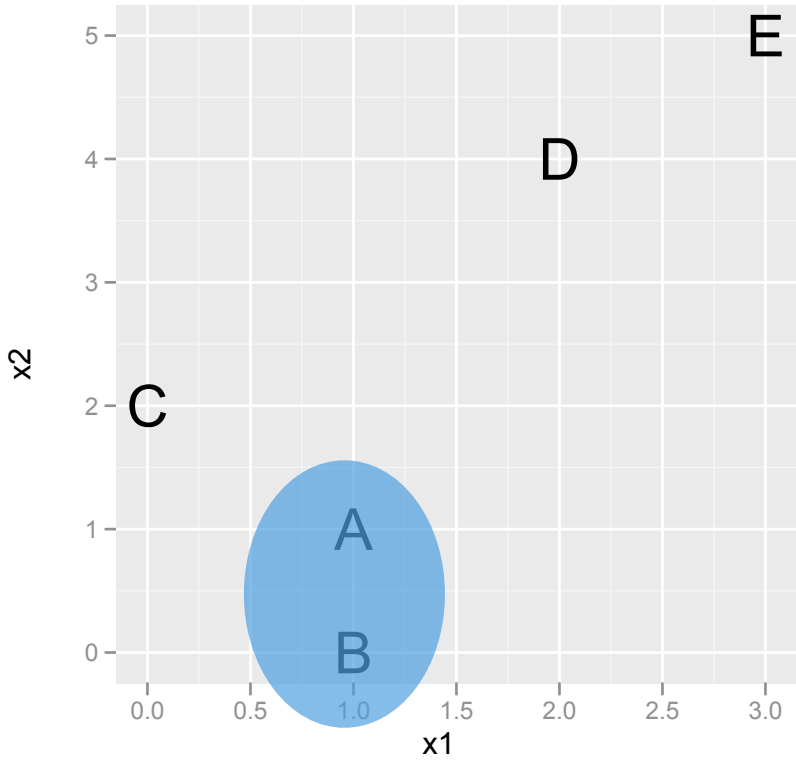
	AB	C	D	E
AB	0	1.8	3.6	4.9
C	1.8	0	2.8	4.2
D	3.6	2.8	0	1.4
E	4.9	4.2	1.4	0



Step 2.1

Join the two closest points into a cluster.

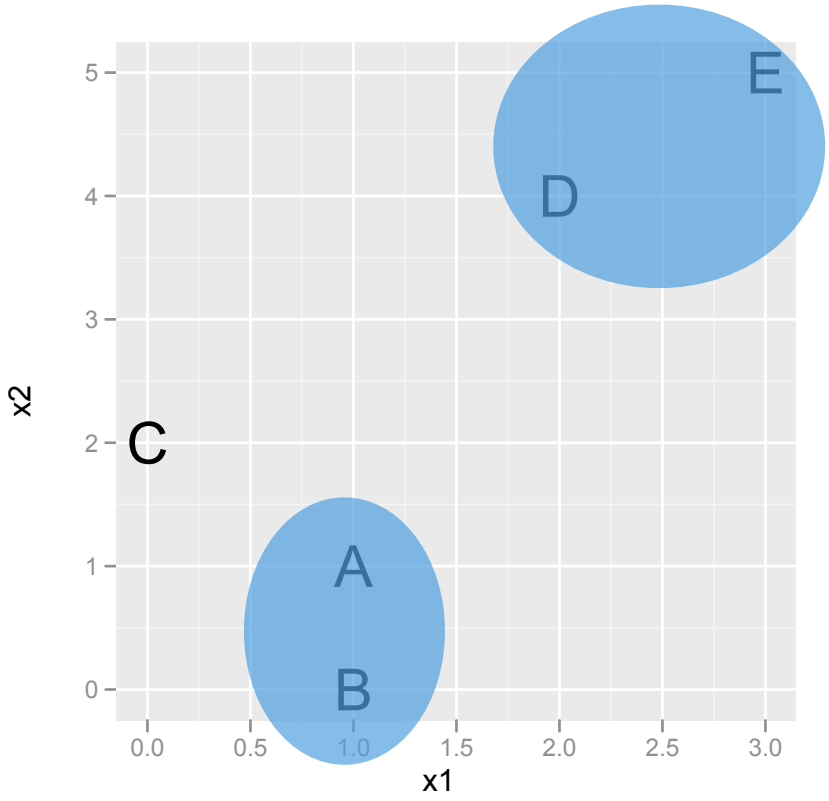
	AB	C	D	E
AB	0	1.8	3.6	4.9
C	1.8	0	2.8	4.2
D	3.6	2.8	0	1.4
E	4.9	4.2	1.4	0



Step 2.1

Join the two closest points into a cluster.

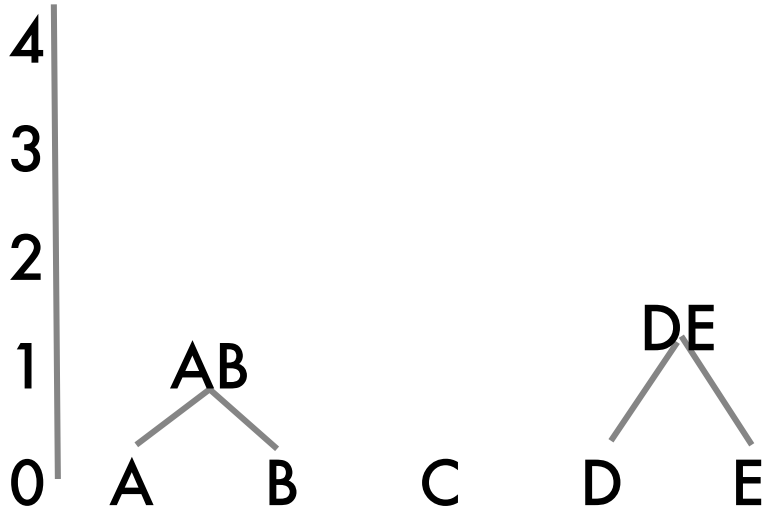
	AB	C	D	E
AB	0	1.8	3.6	4.9
C	1.8	0	2.8	4.2
D	3.6	2.8	0	1.4
E	4.9	4.2	1.4	0



Step 2.2

Reduce the distance matrix, using the linkage methods. Draw the dendrogram.

	AB	C	DE
AB	0	1.8	4.3
C	1.8	0	3.5
DE	4.3	3.5	0

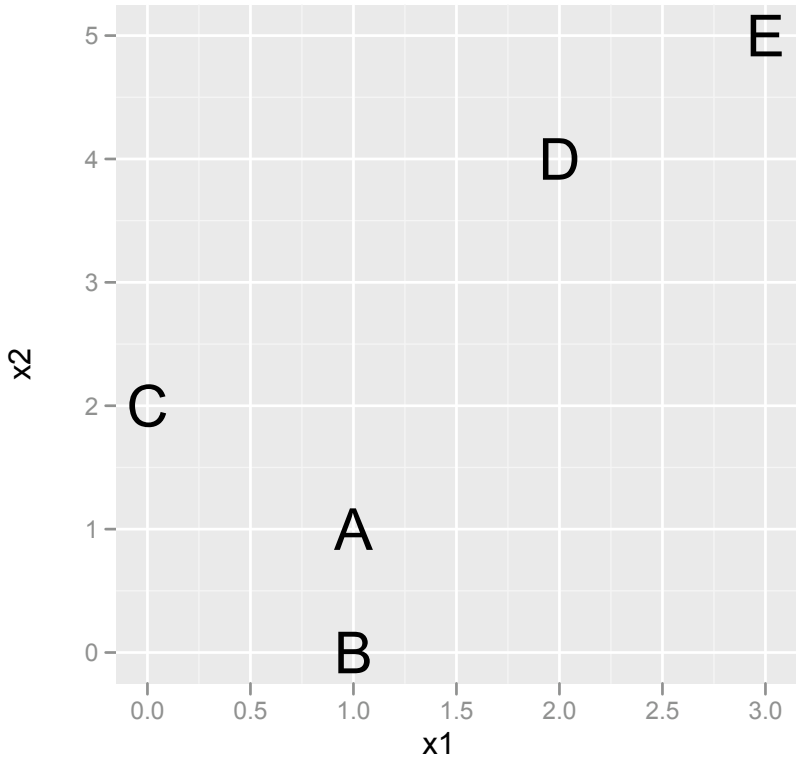


Average linkage used.

Step 3.1

Join the two closest points into a cluster.

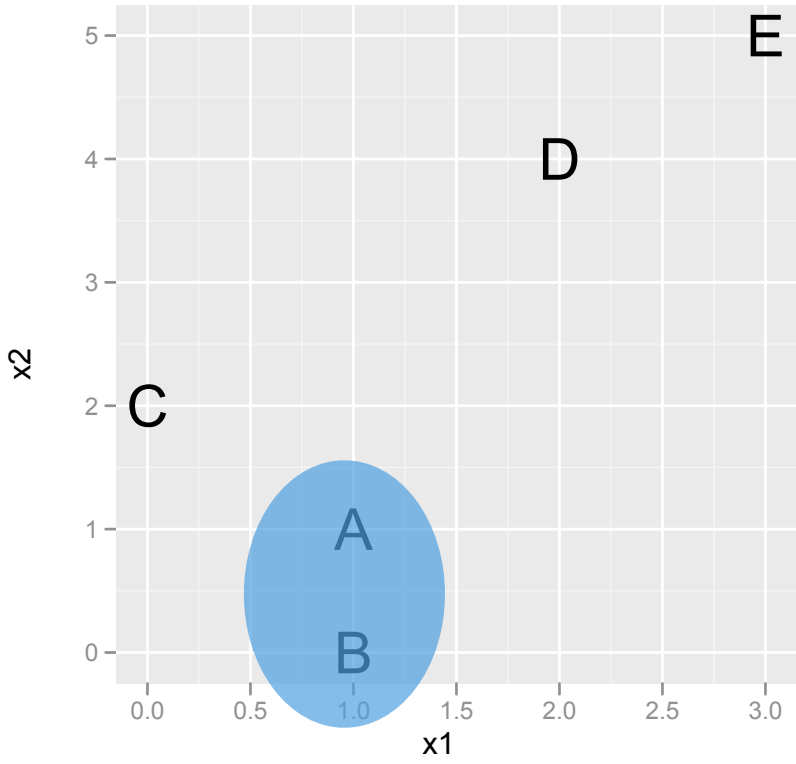
	AB	C	DE
AB	0	1.8	4.3
C	1.8	0	3.5
DE	4.3	3.5	0



Step 3.1

Join the two closest points into a cluster.

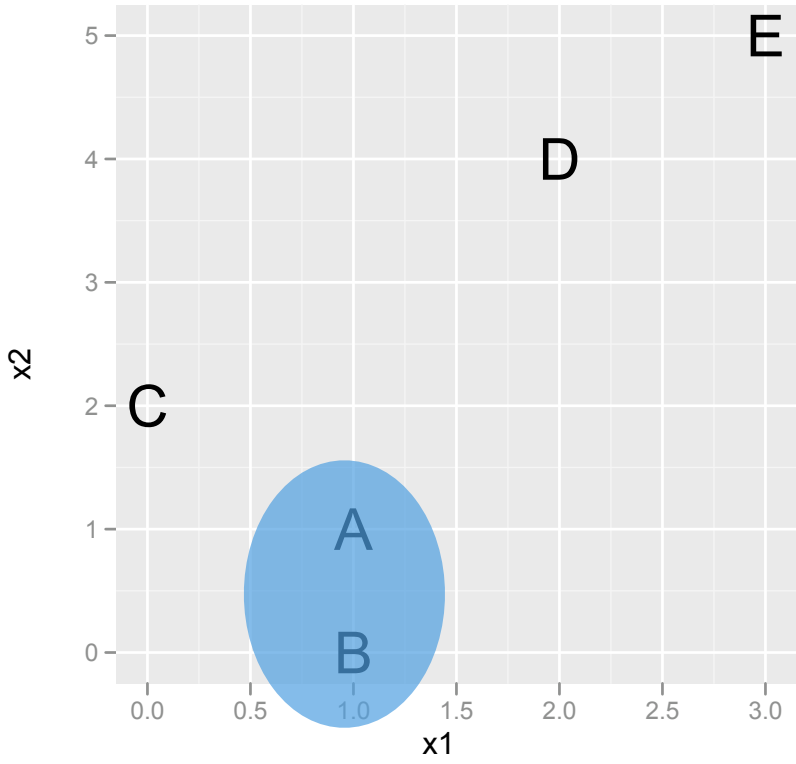
	AB	C	DE
AB	0	1.8	4.3
C	1.8	0	3.5
DE	4.3	3.5	0



Step 3.1

Join the two closest points into a cluster.

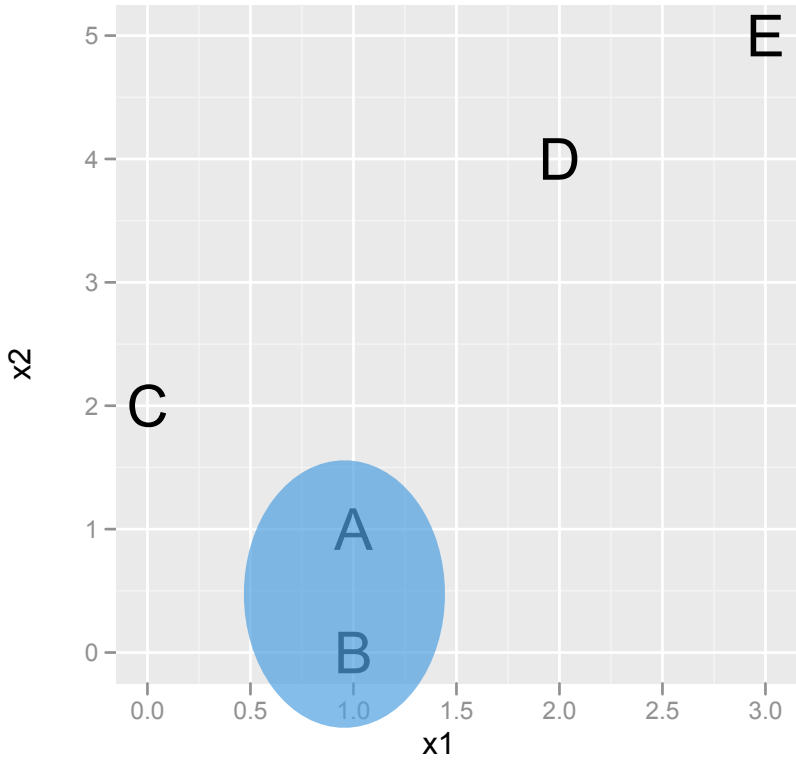
	AB	C	DE
AB	0	1.8	4.3
C	1.8	0	3.5
DE	4.3	3.5	0



Step 3.1

Join the two closest points into a cluster.

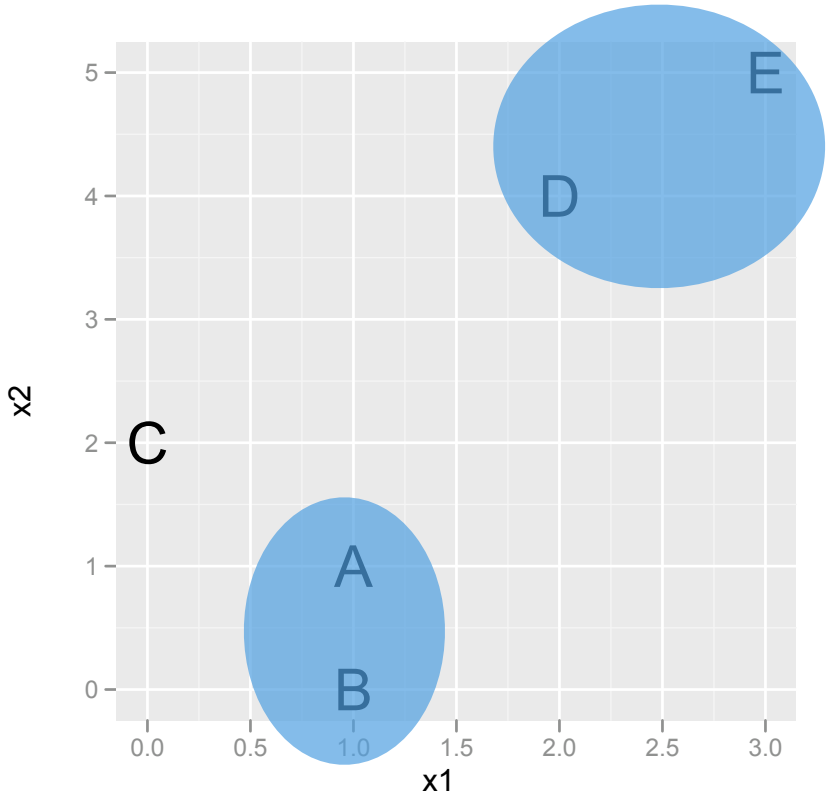
	AB	C	DE
AB	0	1.8	4.3
C	1.8	0	3.5
DE	4.3	3.5	0



Step 3.1

Join the two closest points into a cluster.

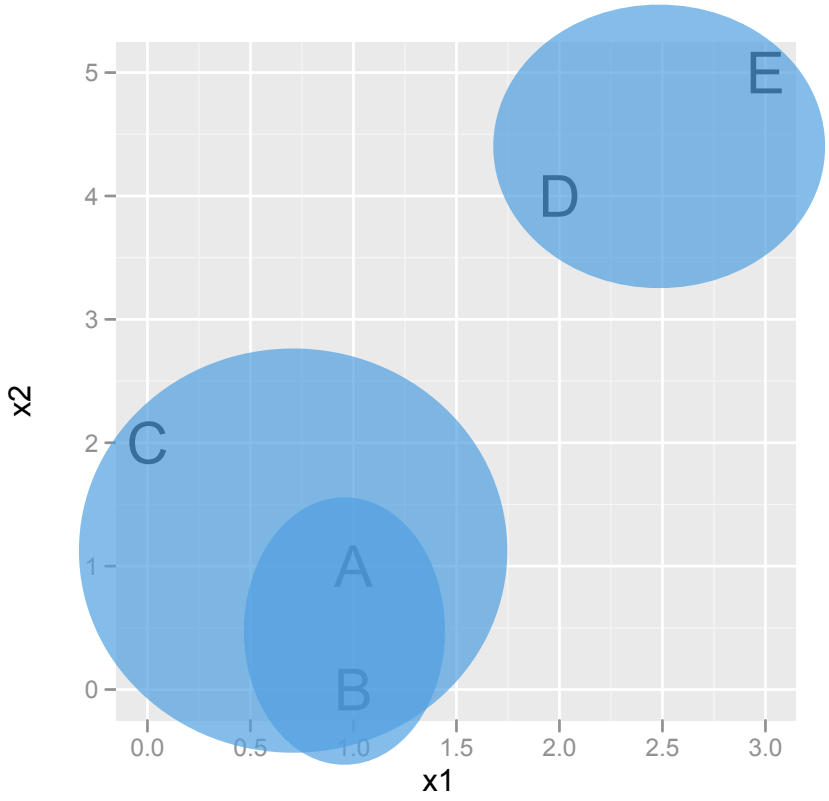
	AB	C	DE
AB	0	1.8	4.3
C	1.8	0	3.5
DE	4.3	3.5	0



Step 3.1

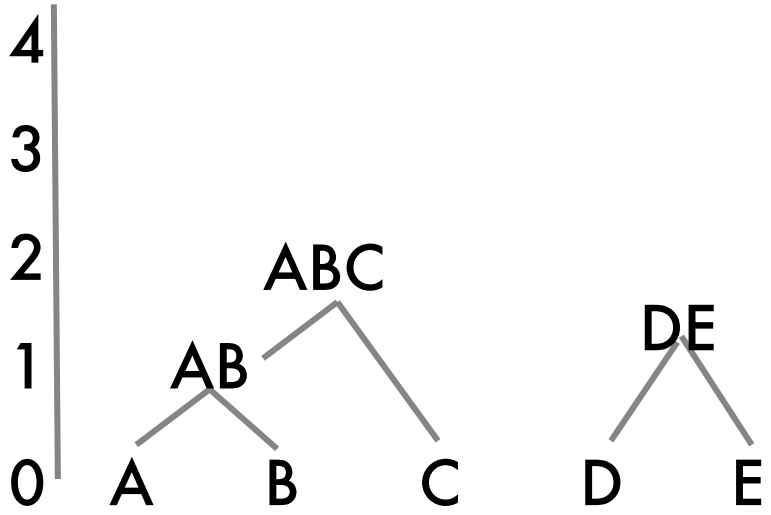
Join the two closest points into a cluster.

	AB	C	DE
AB	0	1.8	4.3
C	1.8	0	3.5
DE	4.3	3.5	0



Step 2.2

Reduce the distance matrix, using the linkage methods. Draw the dendrogram.

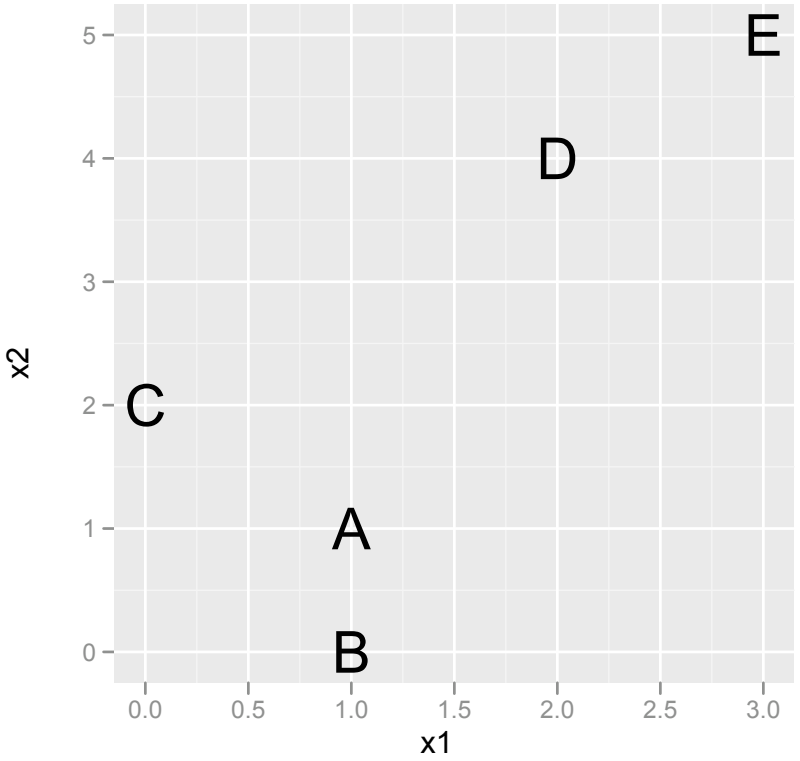


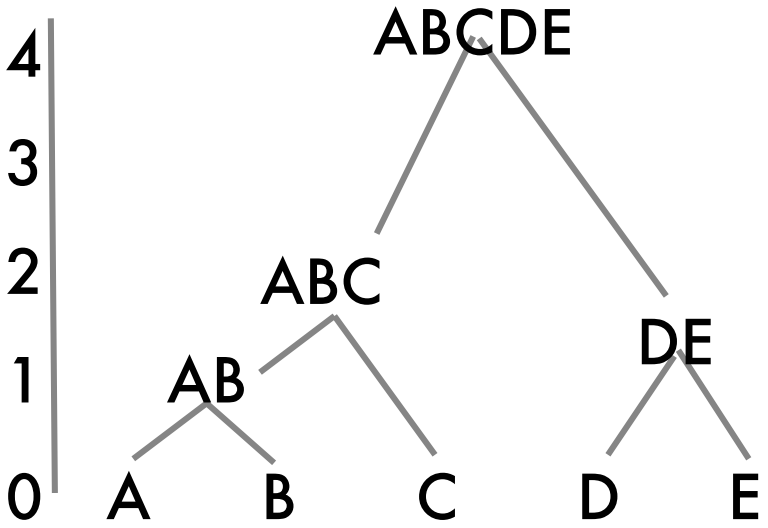
	ABC	DE
ABC	0	4.0
DE	4.0	0

Average linkage used.

Step 3

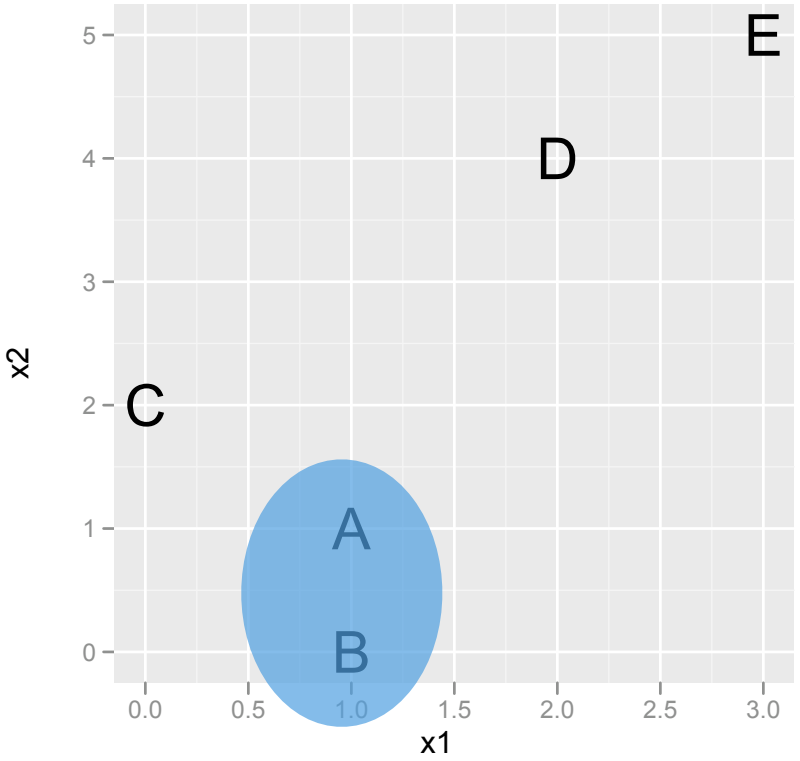
Join last two clusters

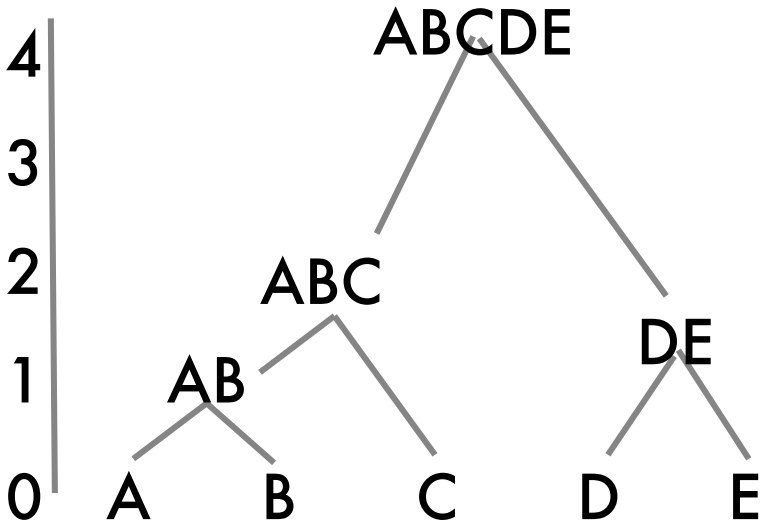




Step 3

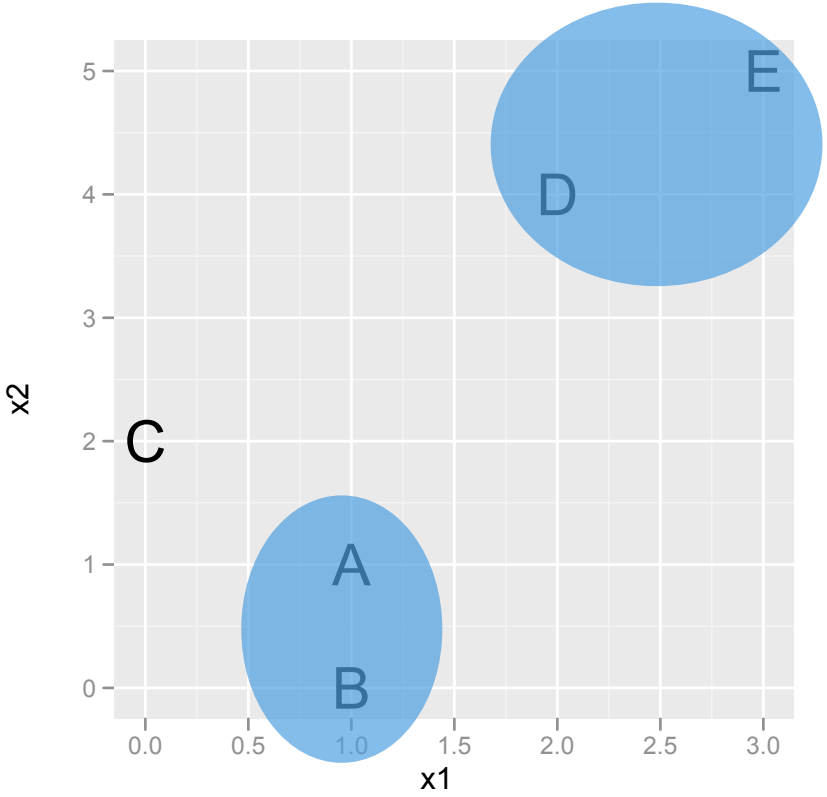
Join last two clusters

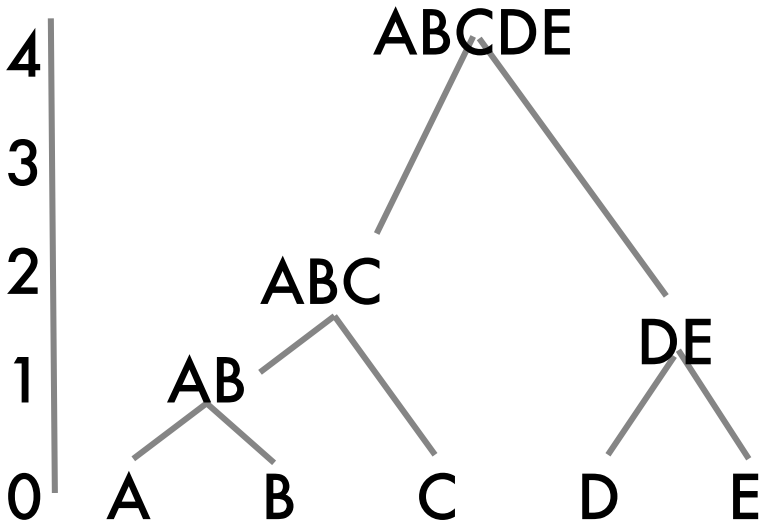




Step 3

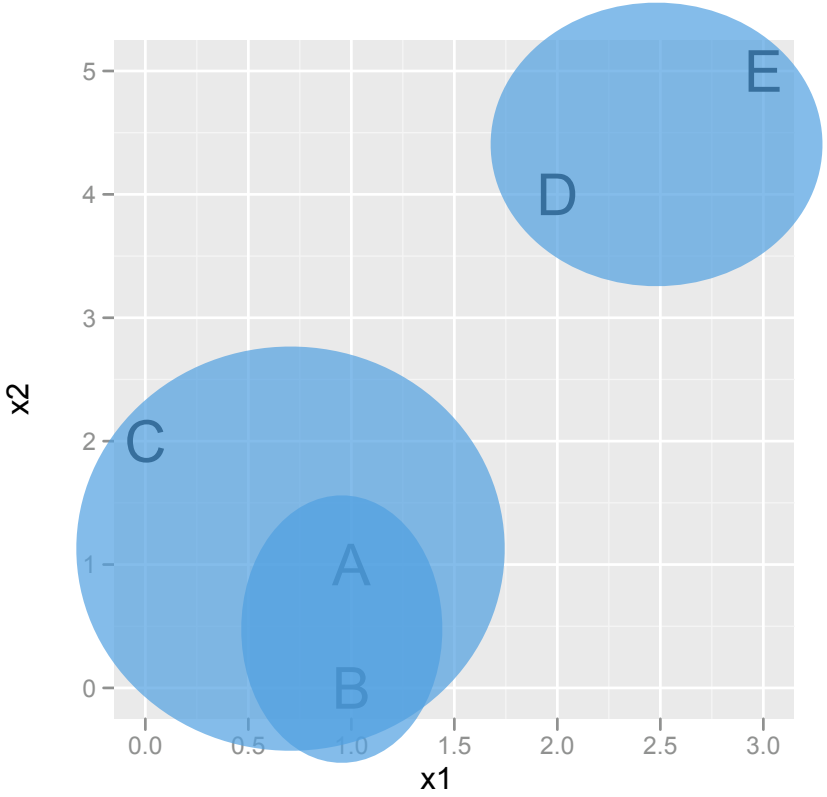
Join last two clusters

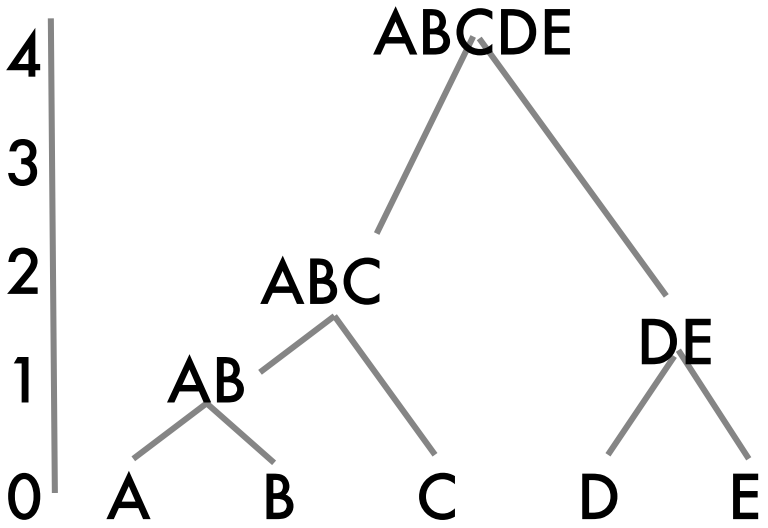




Step 3

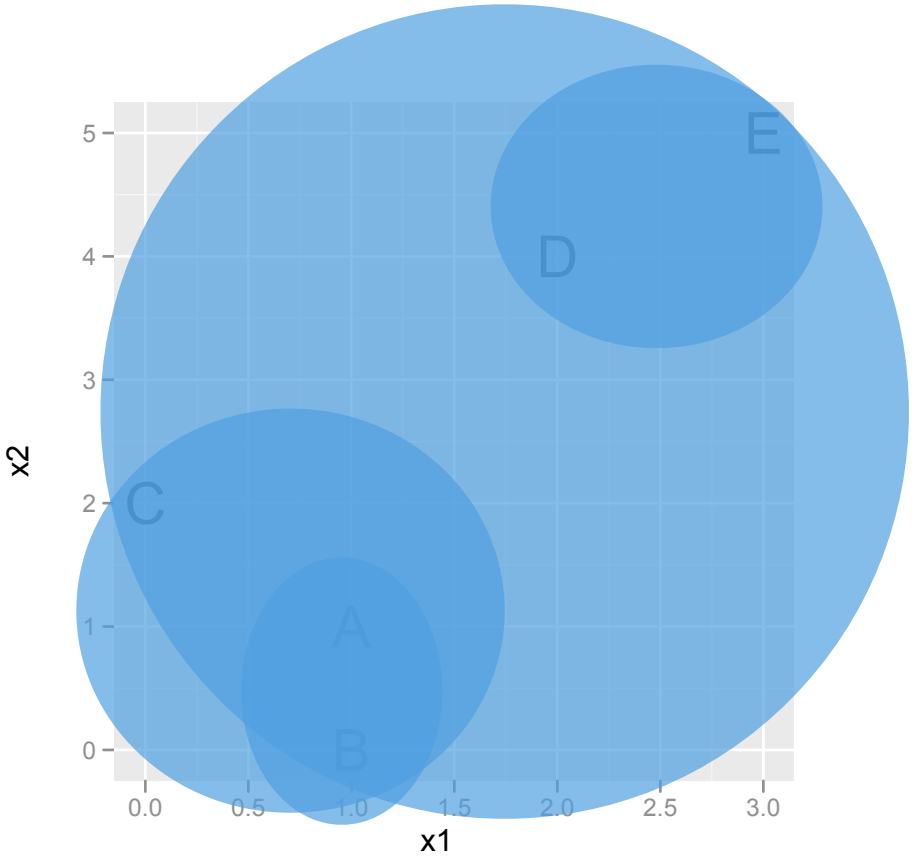
Join last two clusters

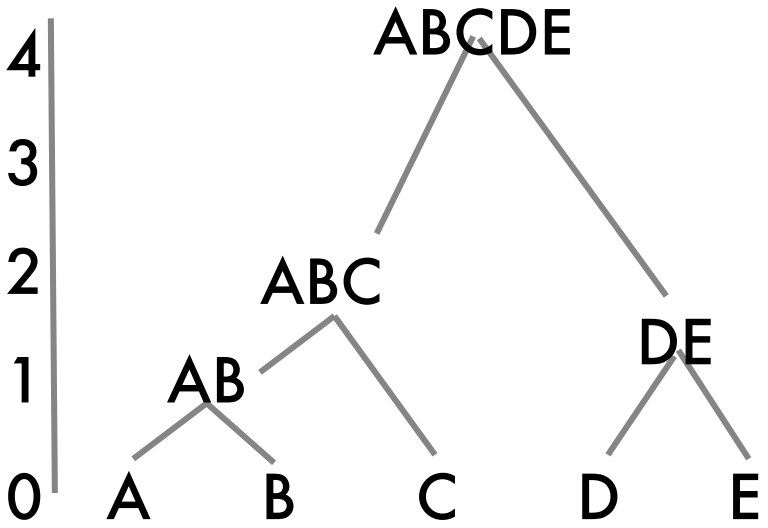




Step 3

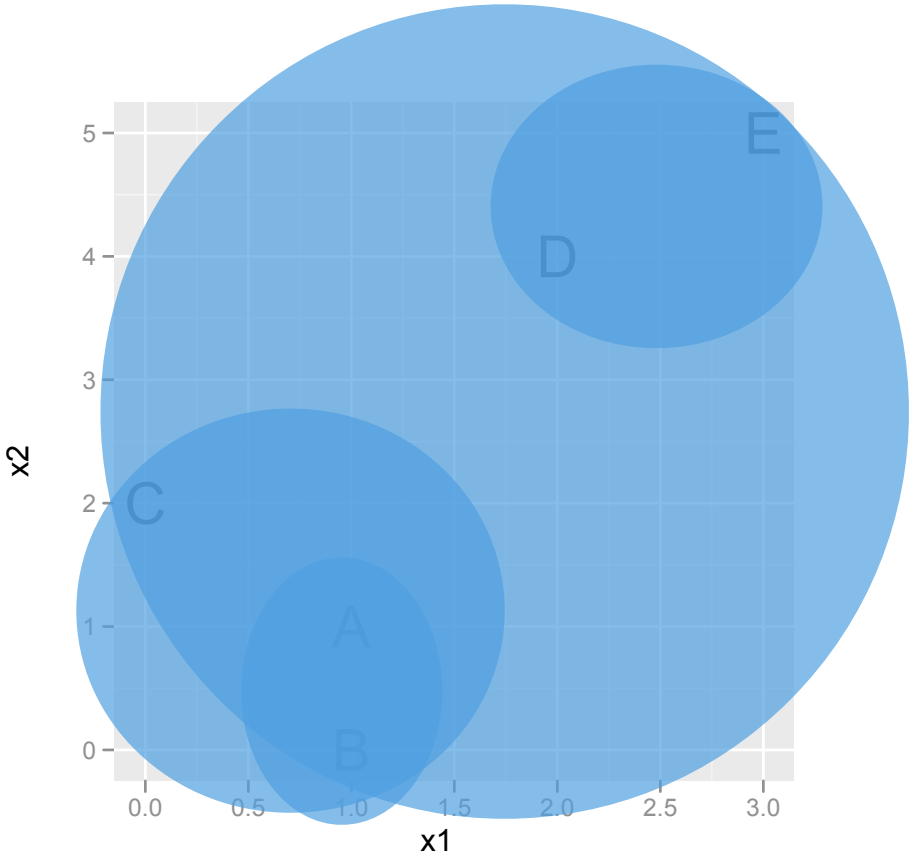
Join last two clusters

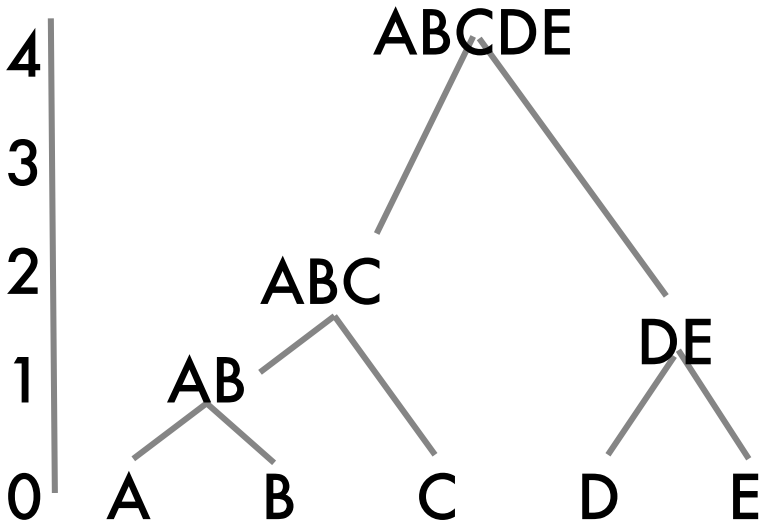




Step 3

Join last two clusters

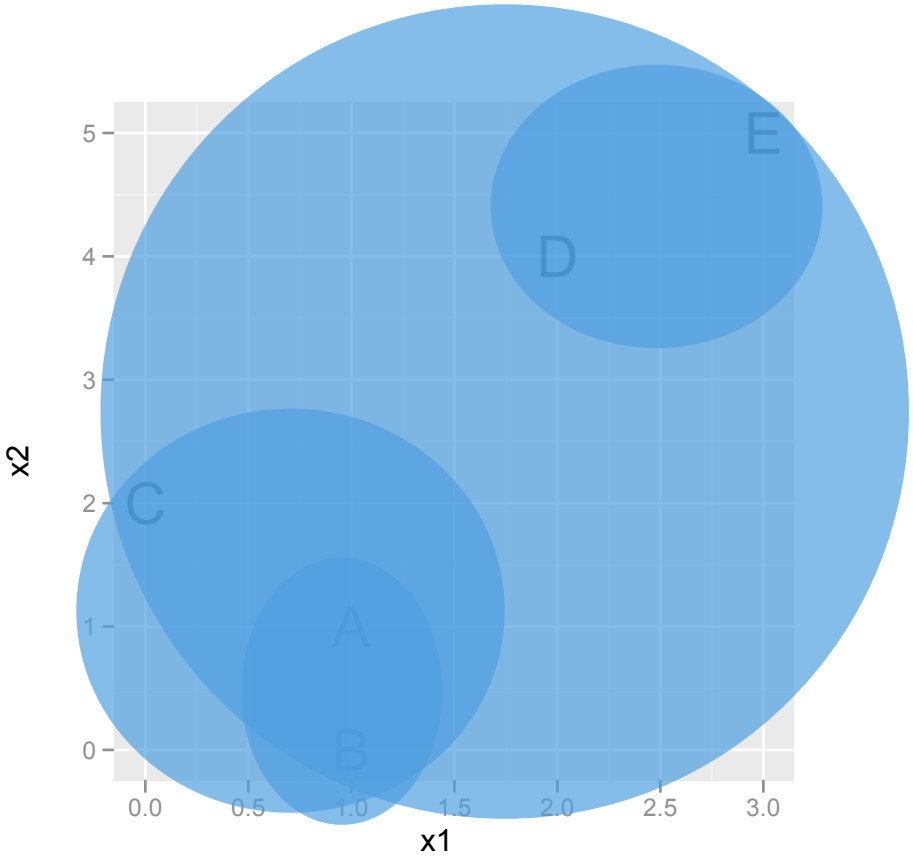


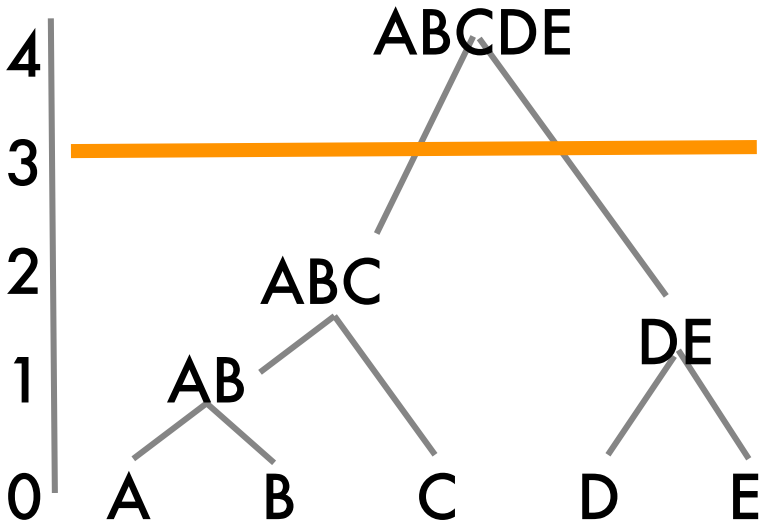


Use dendrogram to decide on the number of clusters.

Step 3

Join last two clusters

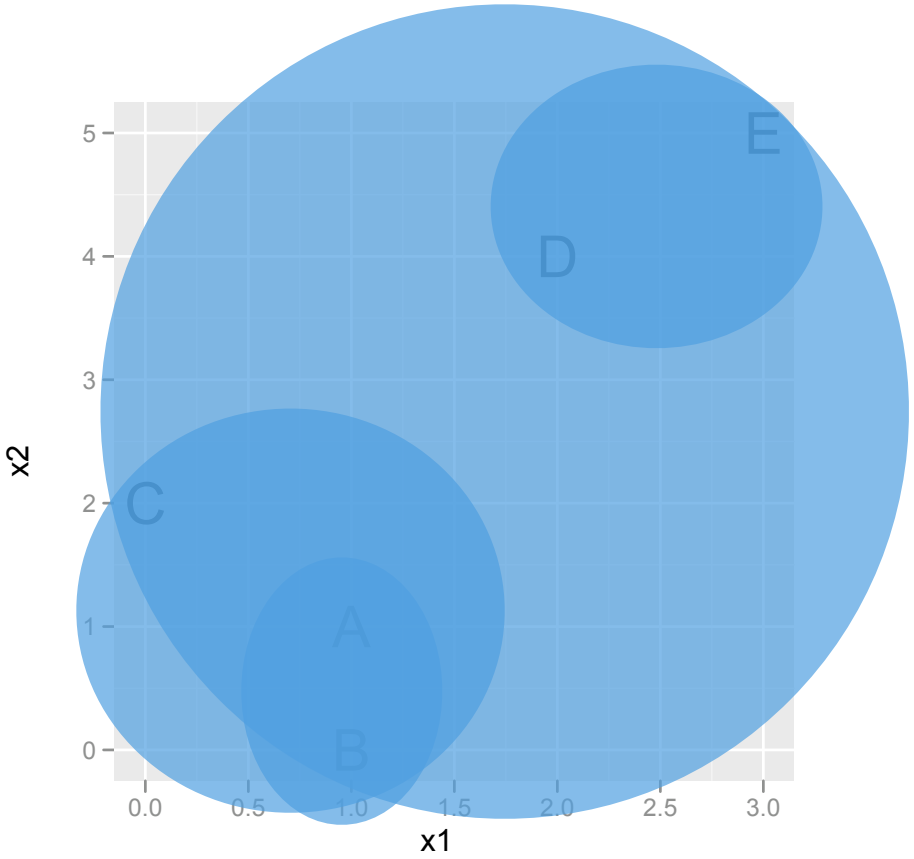


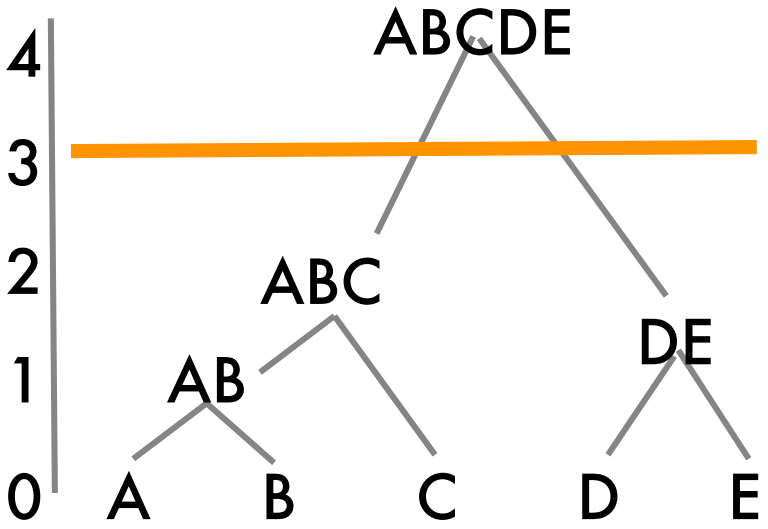


Use dendrogram to decide on the number of clusters.

Step 3

Join last two clusters

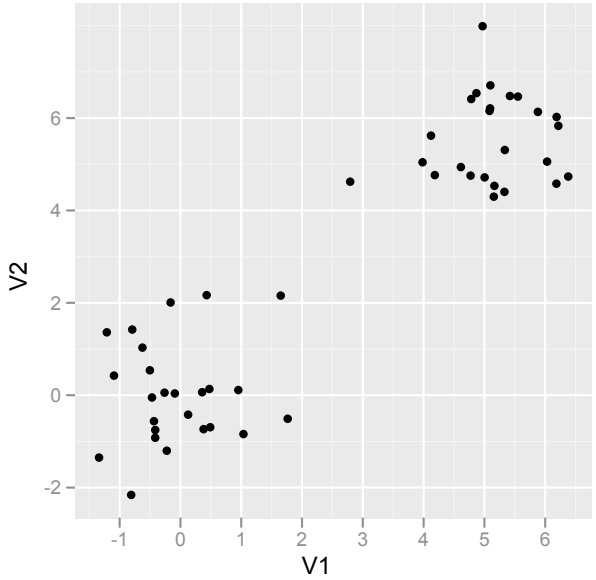




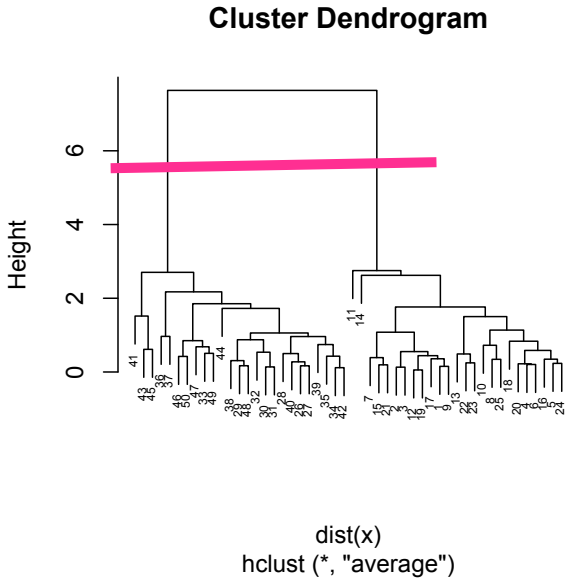
Use dendrogram to decide on the number of clusters.

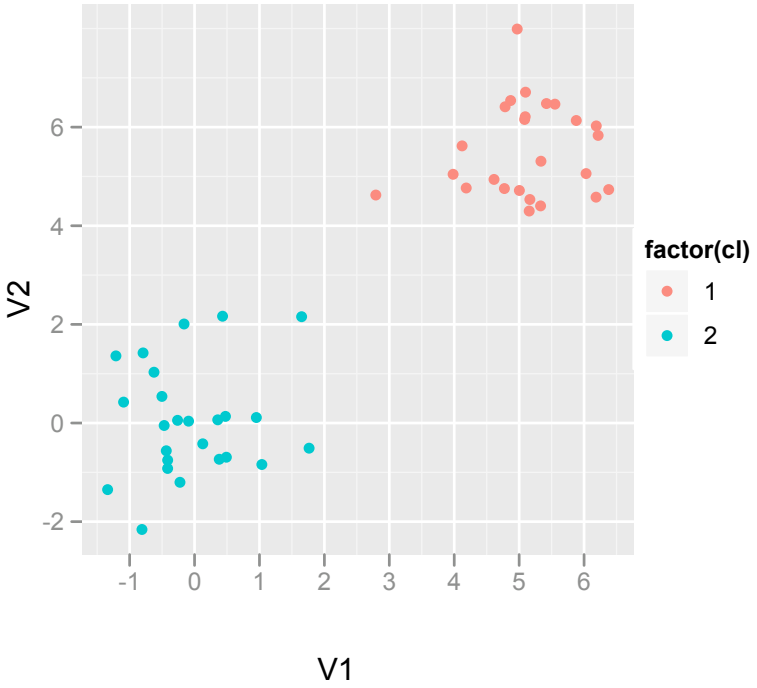
Two clusters: ABC, DE

Examples

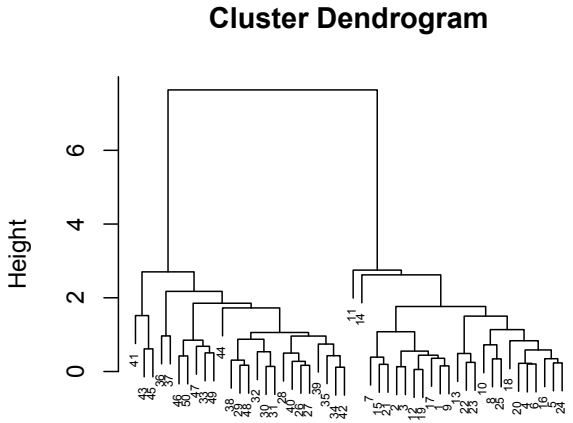


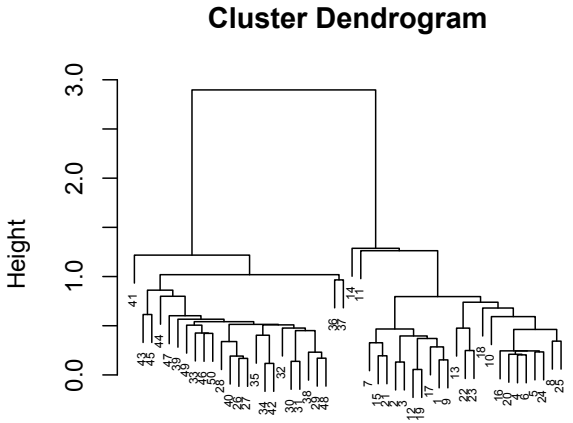
Simulated data with two clusters.

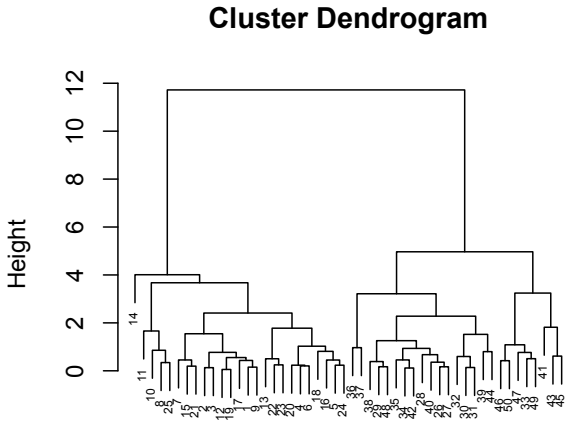


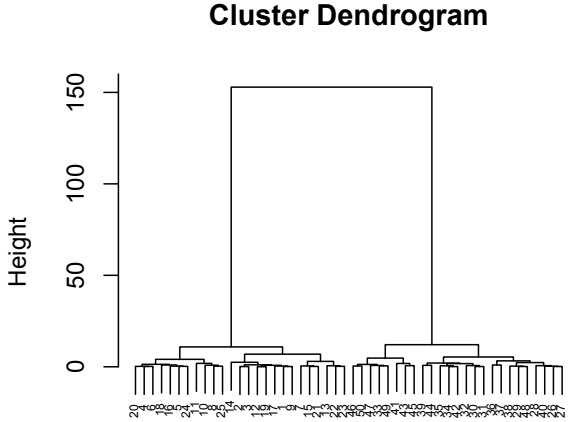


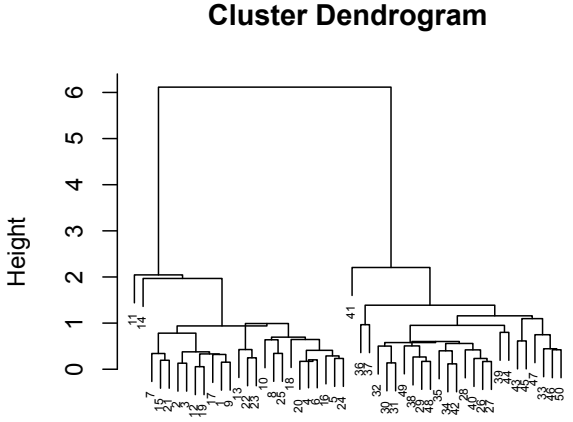
Examples







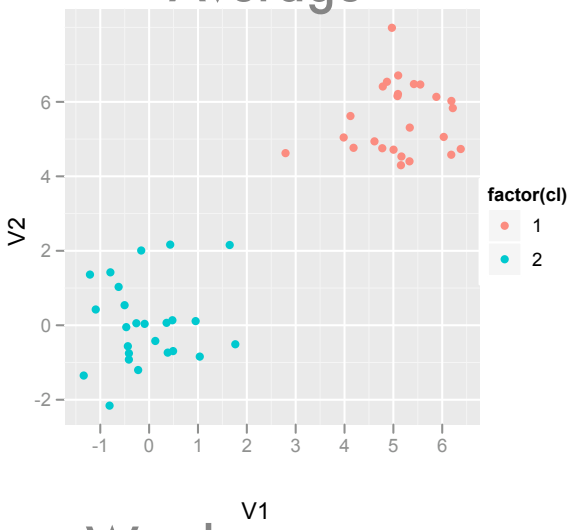




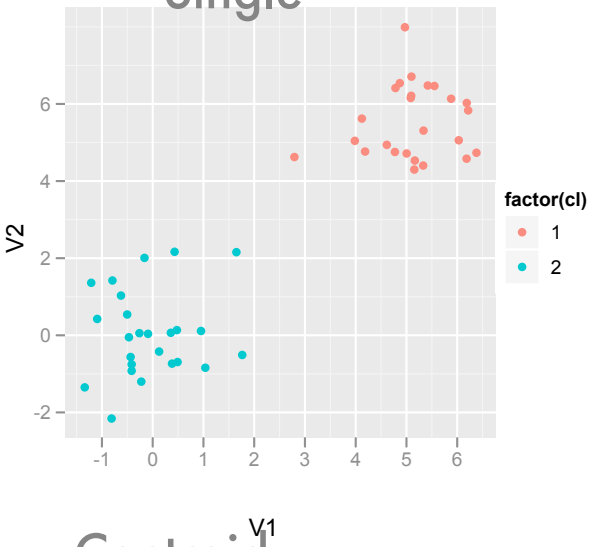
Each of the dendrograms suggests two clusters, but there are a lot of differences. Several suggest some points are outliers: 11, 14, 41.

Examples

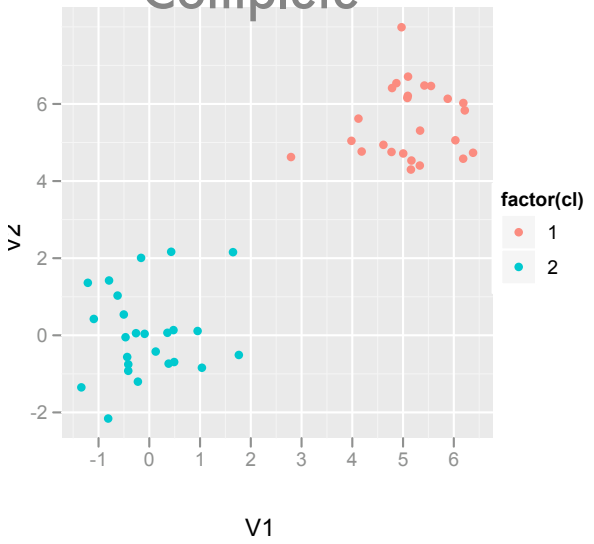
Average



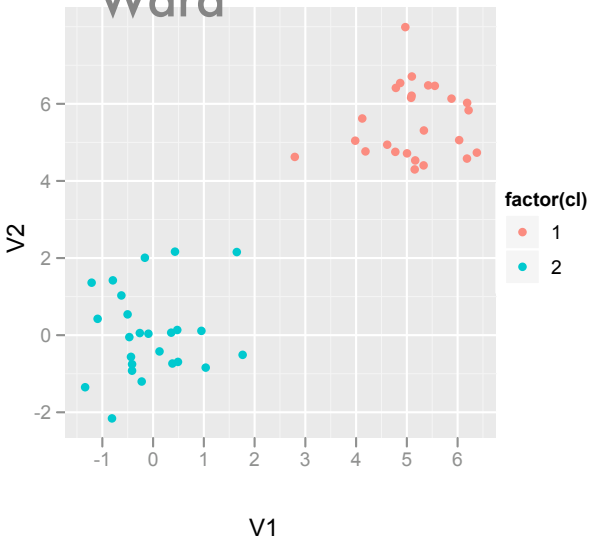
Single



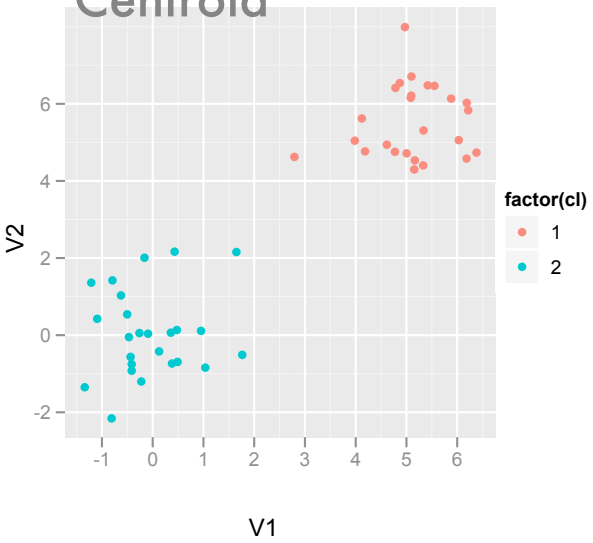
Complete



Ward

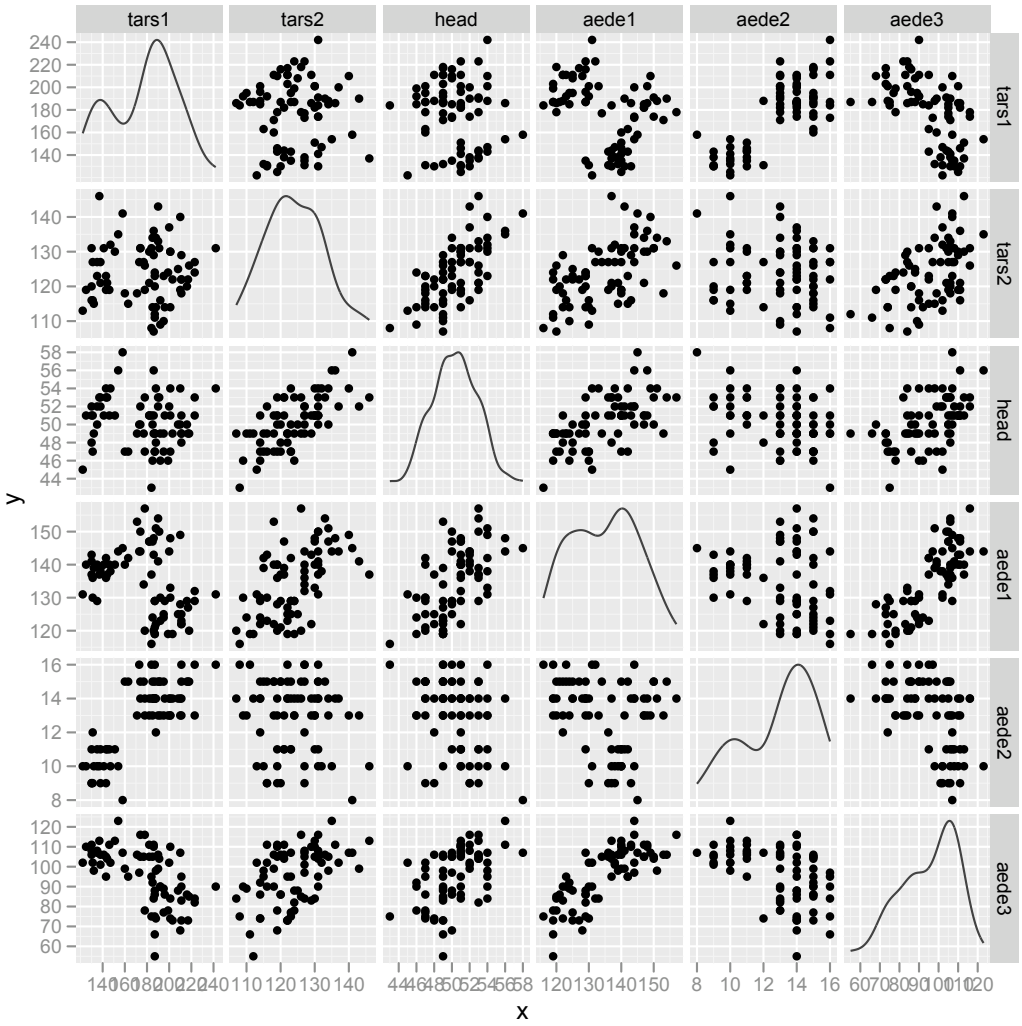


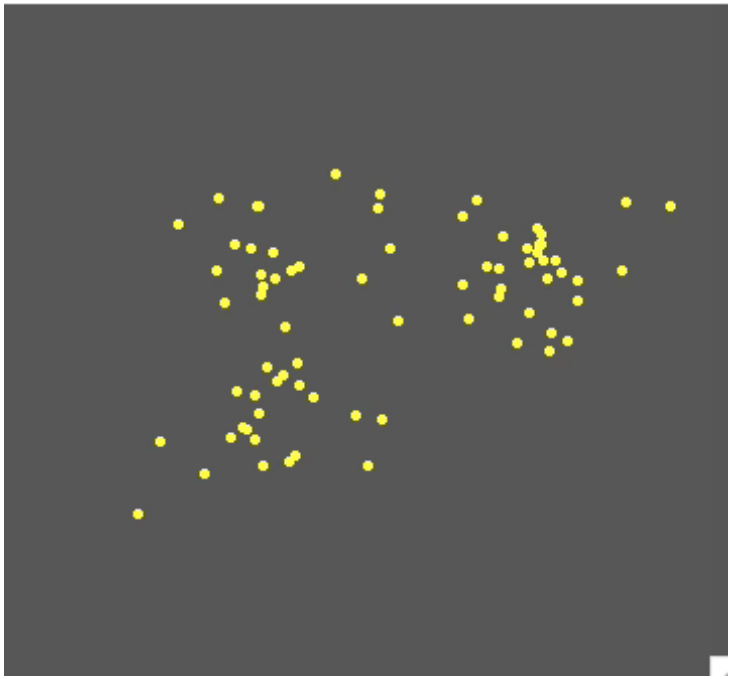
Centroid



And the two cluster solution is the same for all methods.

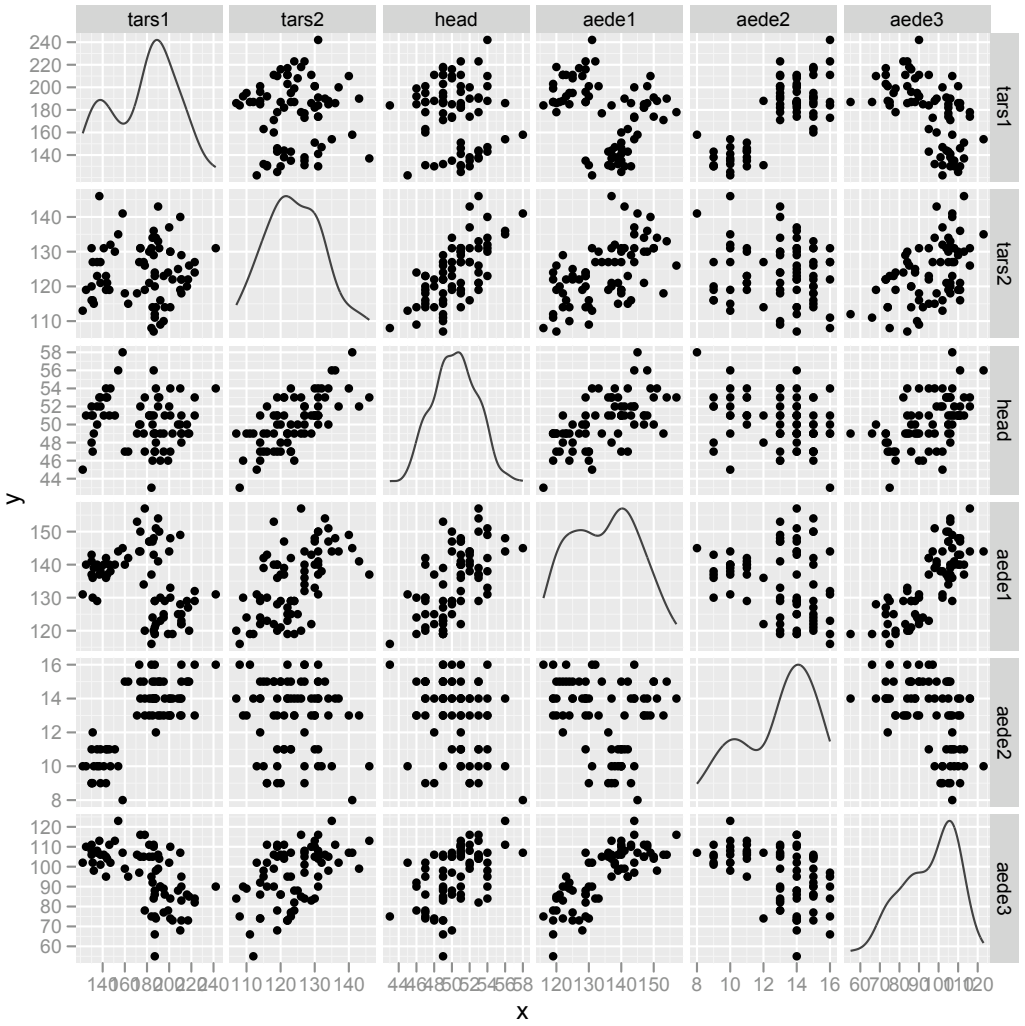
Flea beetles

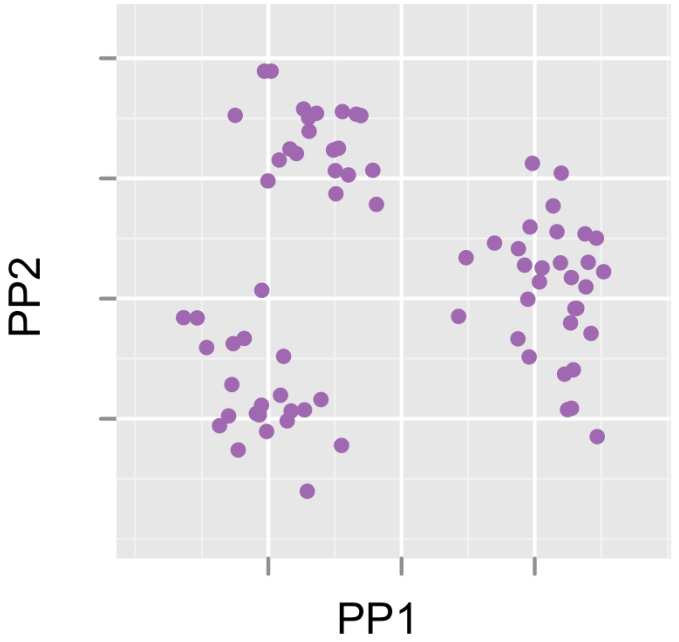




We expect clustering to produce three clusters. Data is standardized before calculating distances.

Flea beetles

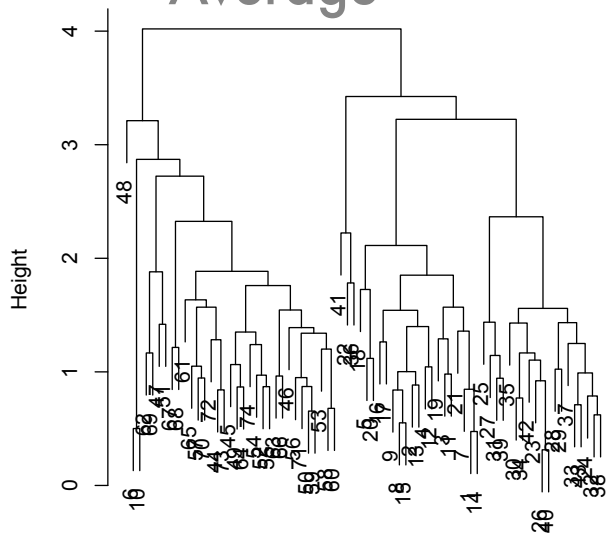




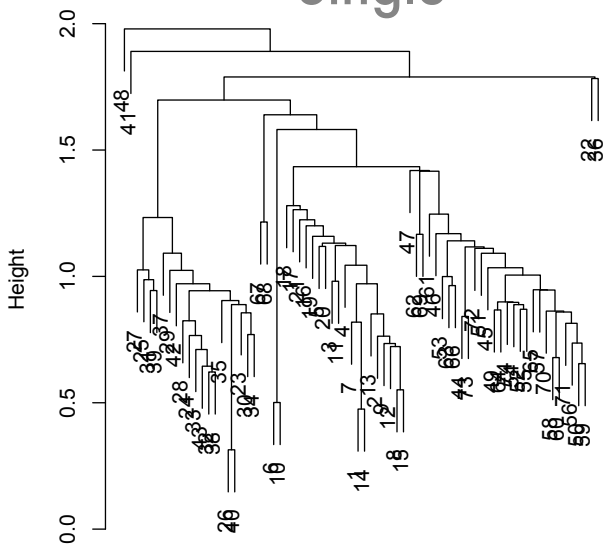
We expect clustering to produce three clusters. Data is standardized before calculating distances.

Flea beetles

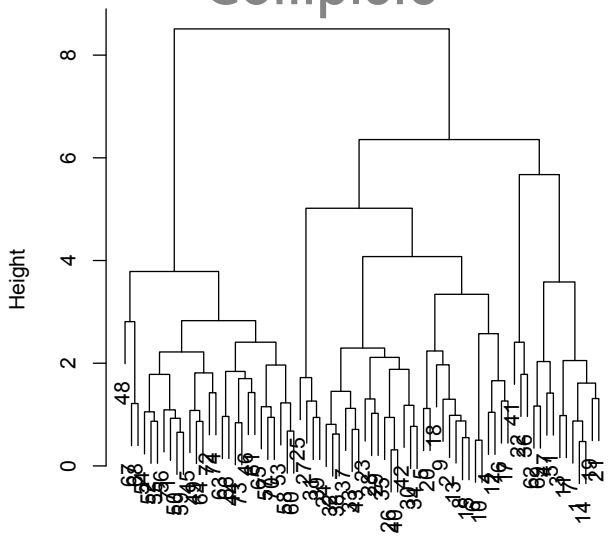
Average



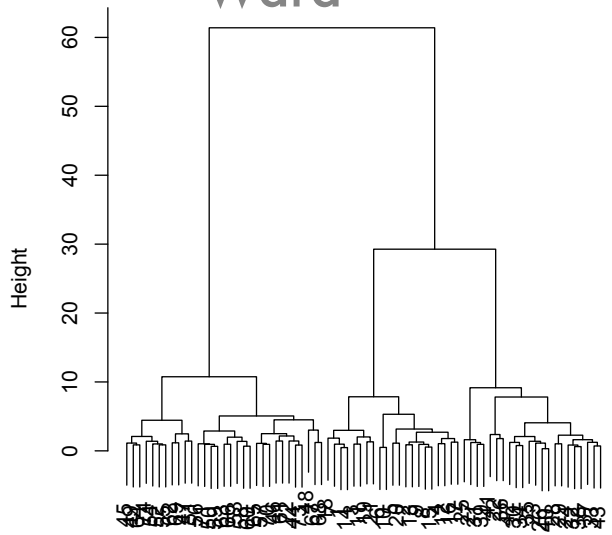
Single



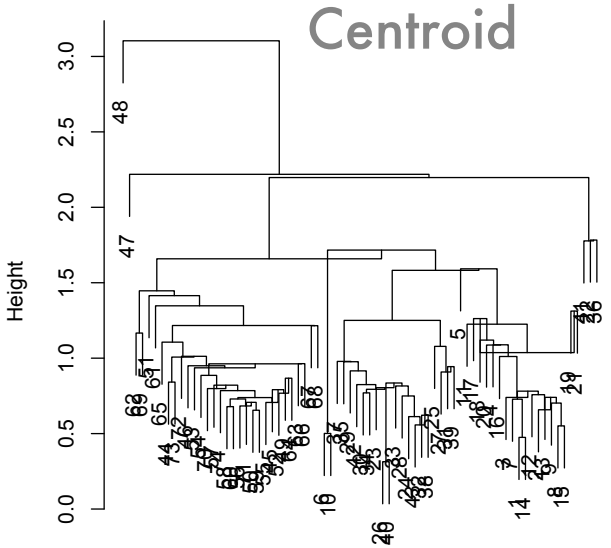
Complete



Ward



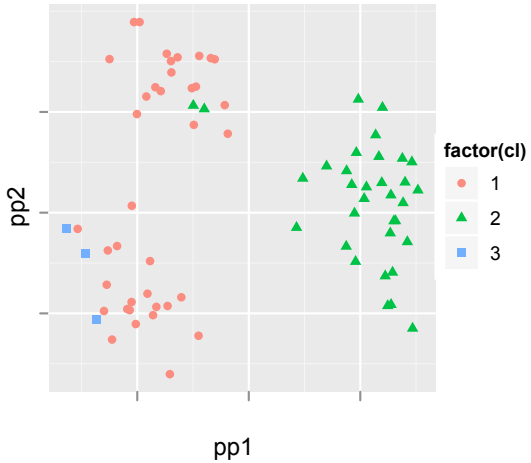
Centroid



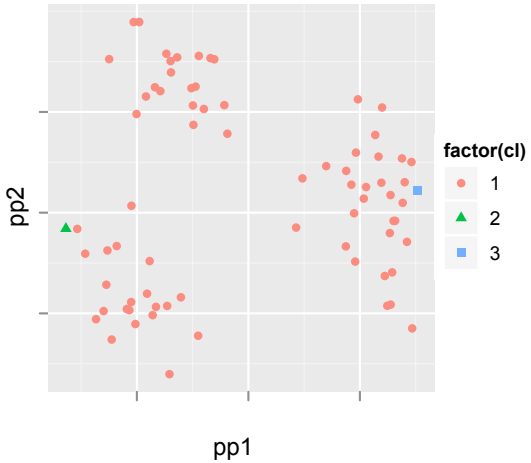
No agreement between methods. Which would you use?

Flea beetles

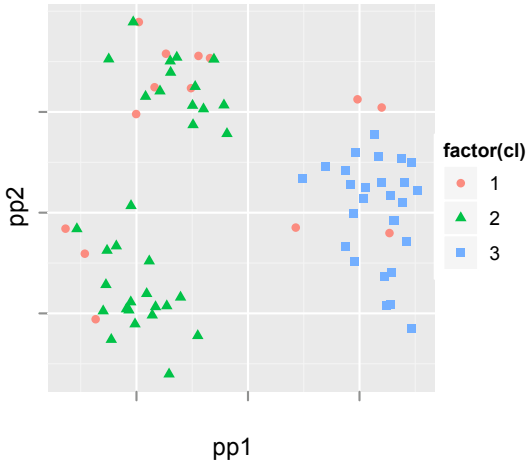
Average



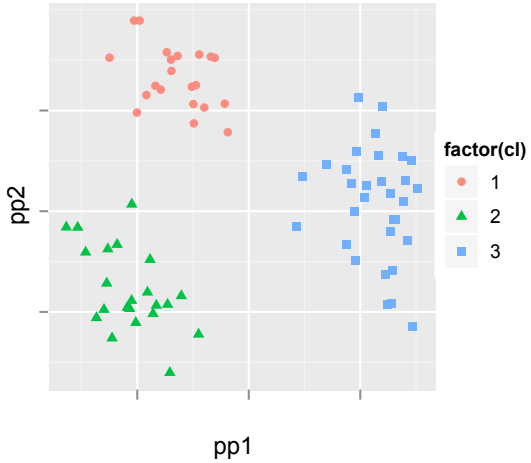
Single



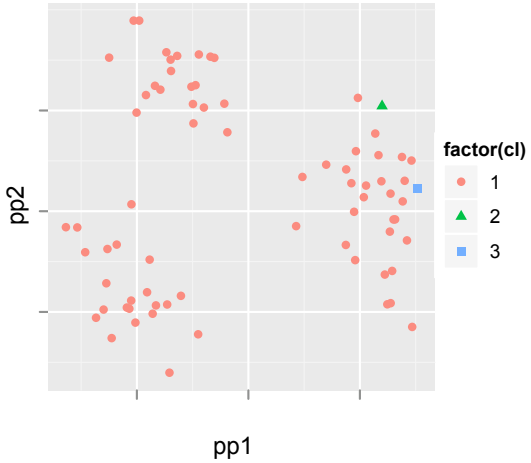
Complete



Ward

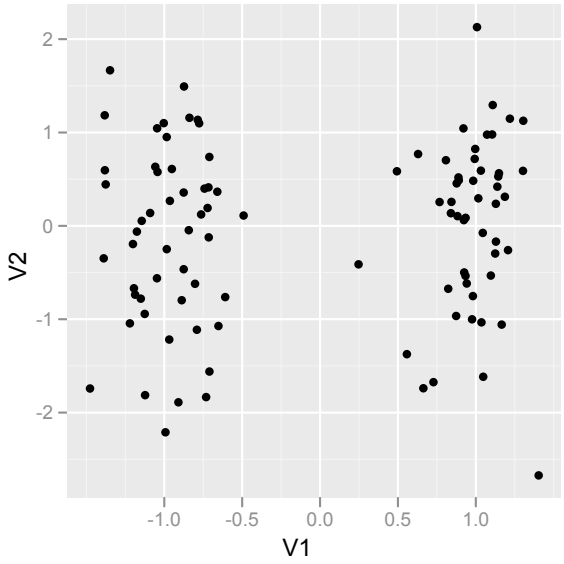


Centroid



Which would you use? How can there be so much difference?

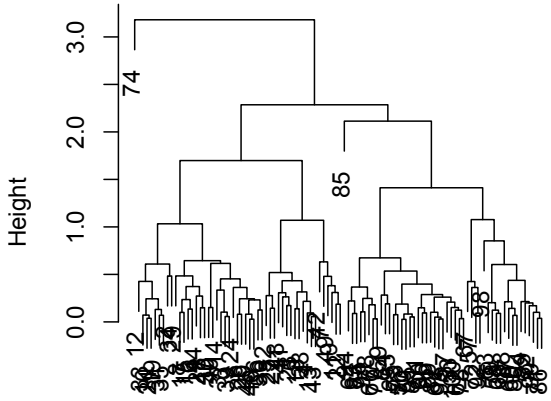
Nuisance variables



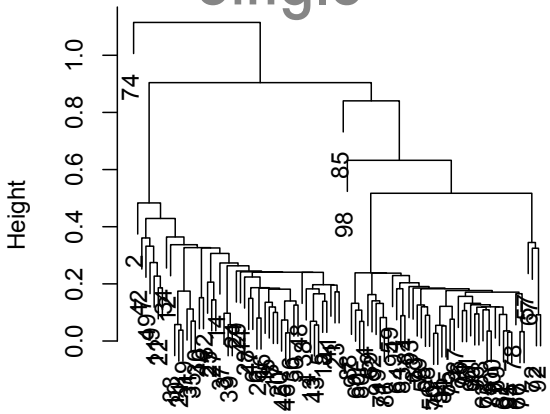
Variables that don't contribute to the clustering but are included in the distance calculations.

Nuisance variables

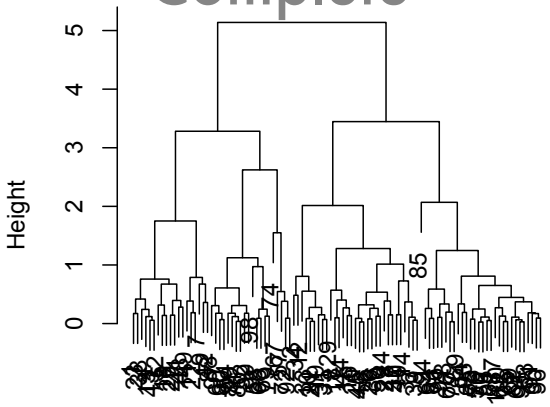
Average



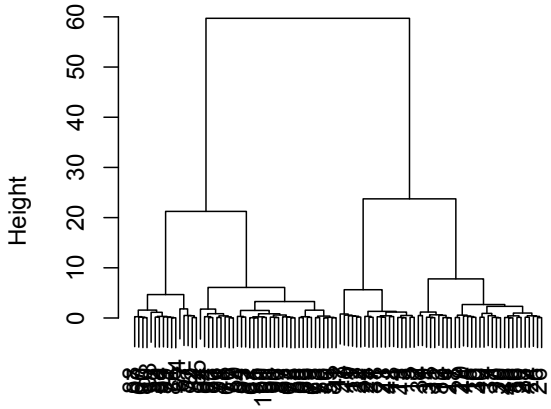
Single



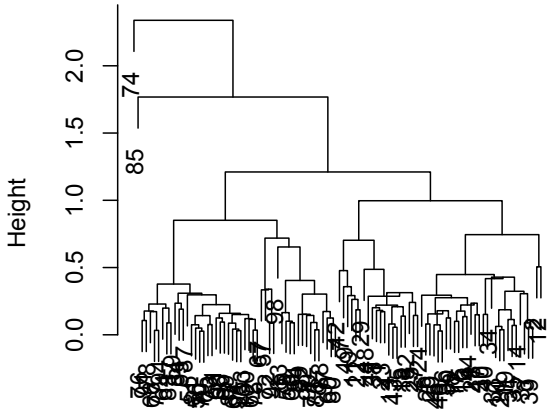
Complete



Ward



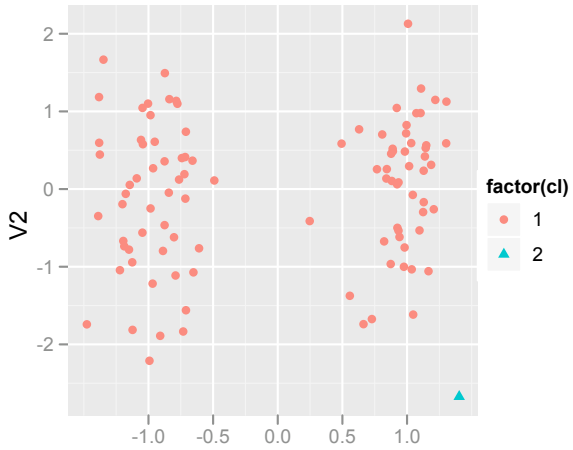
Centroid



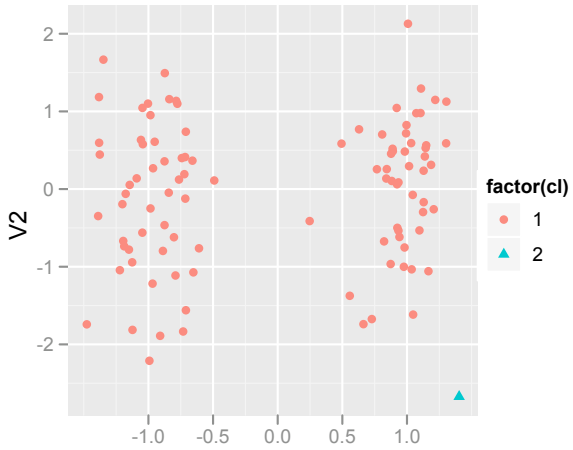
Complete and ward linkage see two clusters

Nuisance variables

Average

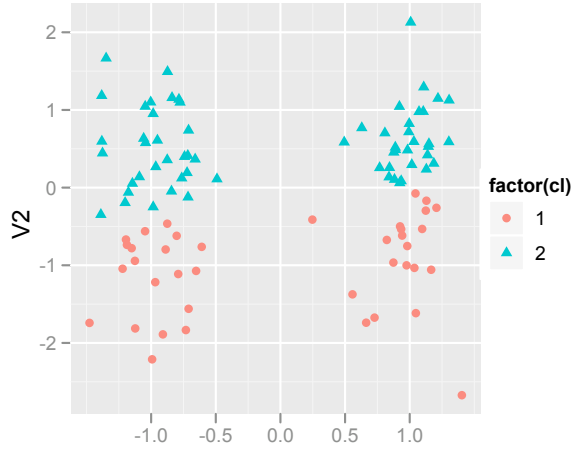


Single

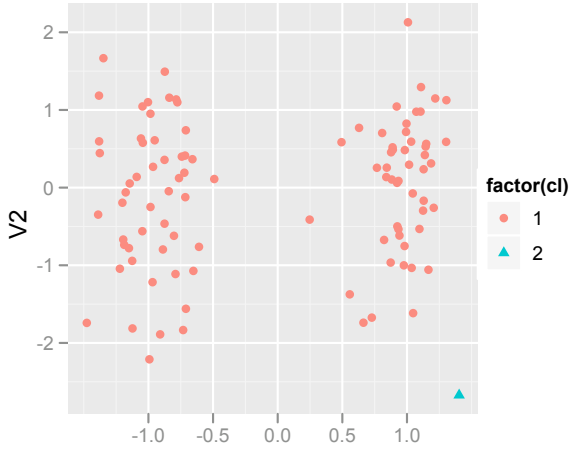


Complete

Ward



Centroid



V1

Is this what you expected of complete and ward?

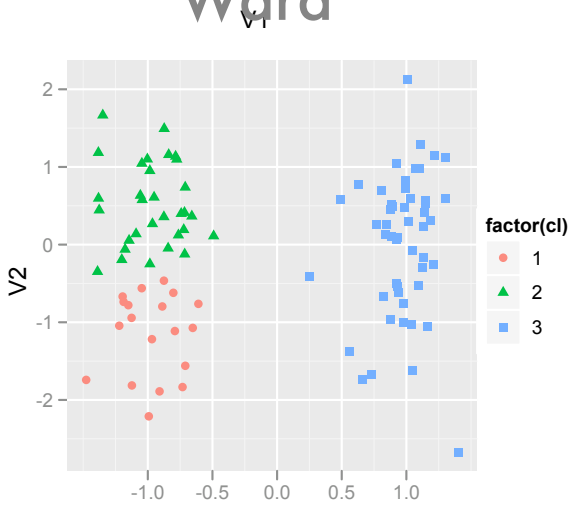
Nuisance variables

Average

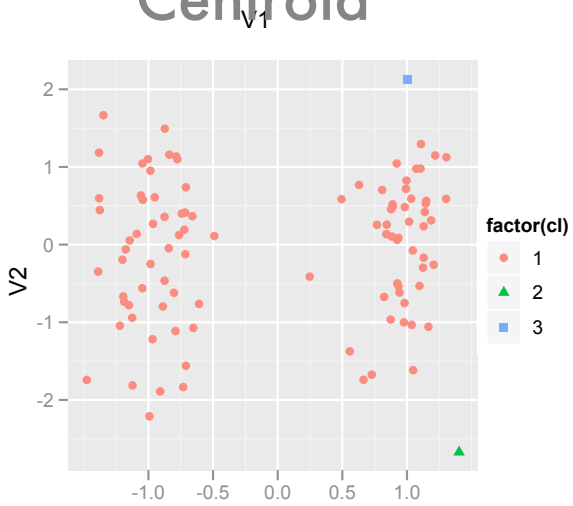
Single

Complete

Ward



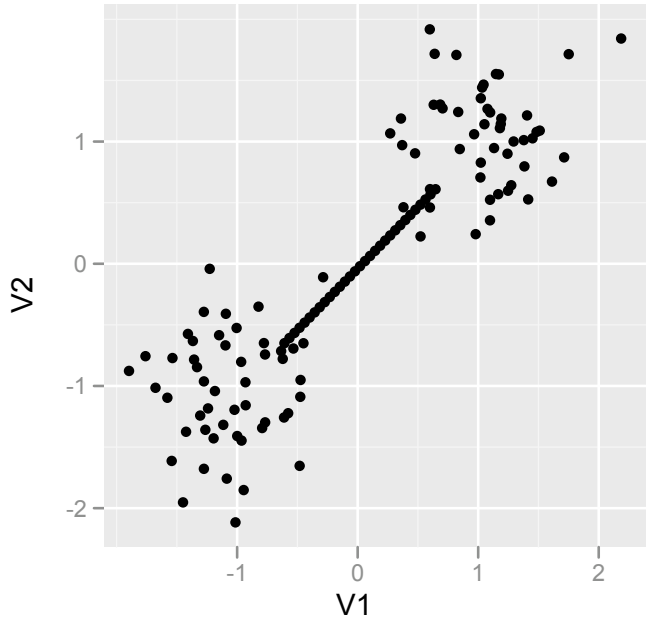
Centroid



V1

Which method is best, now?

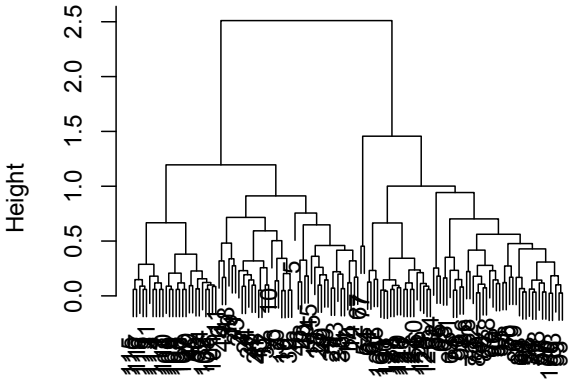
Nuisance points



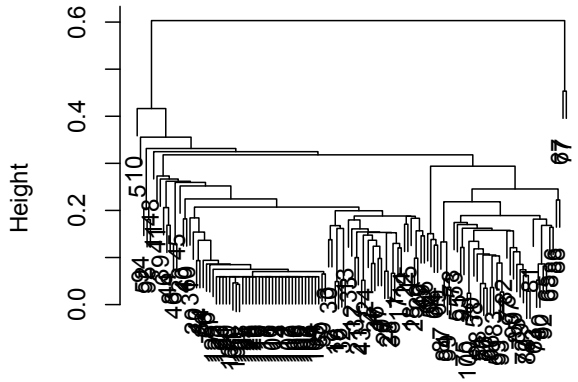
Points that are between major clusters of data. This affects some linkage methods, eg single, which will tend to "chain" through the data grouping everything together.

Nuisance points

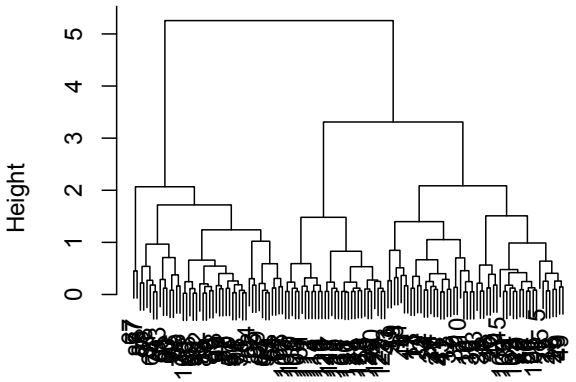
Average



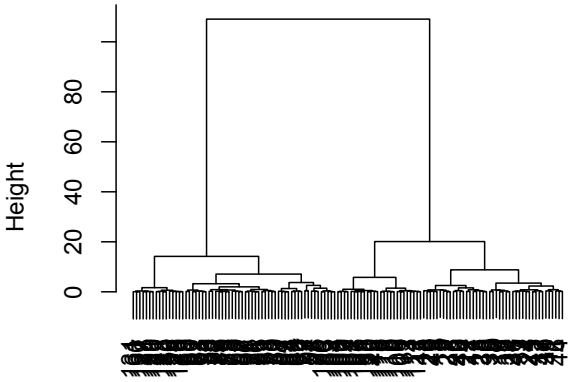
Single



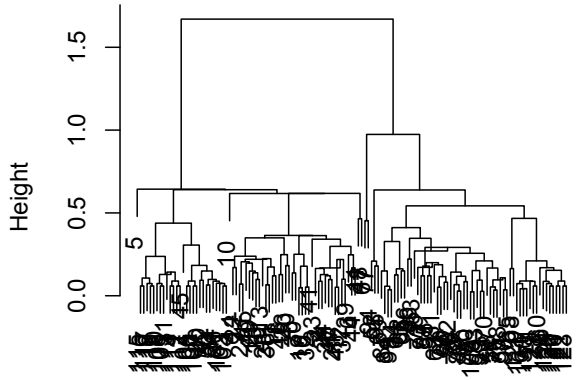
Complete



Ward



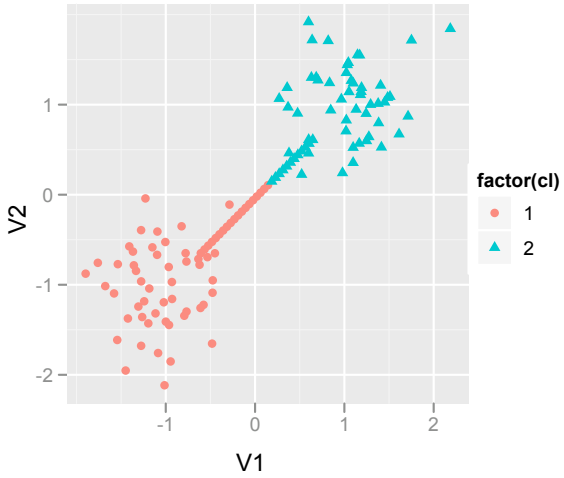
Centroid



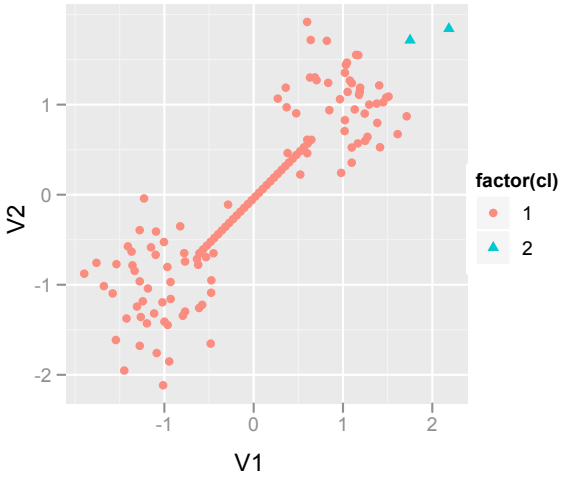
Only single linkage does NOT see two clusters

Nuisance points

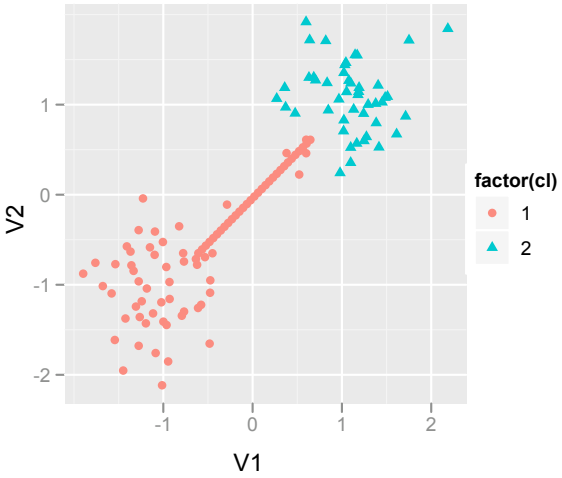
Average



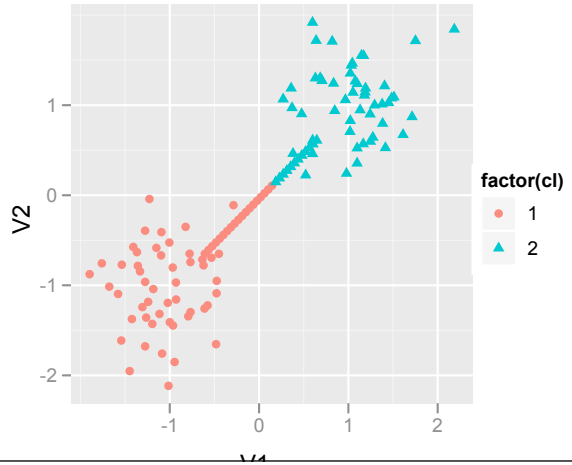
Single



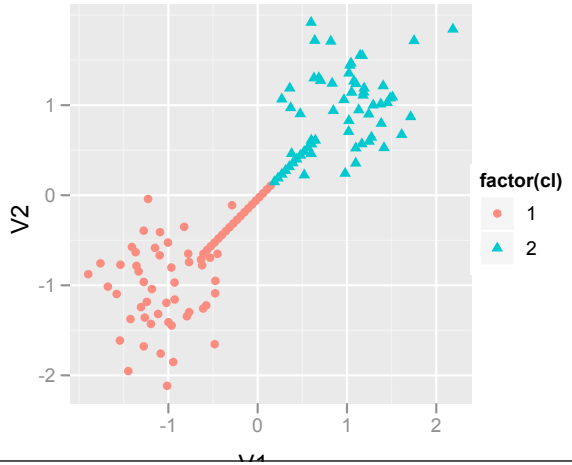
Complete



Ward



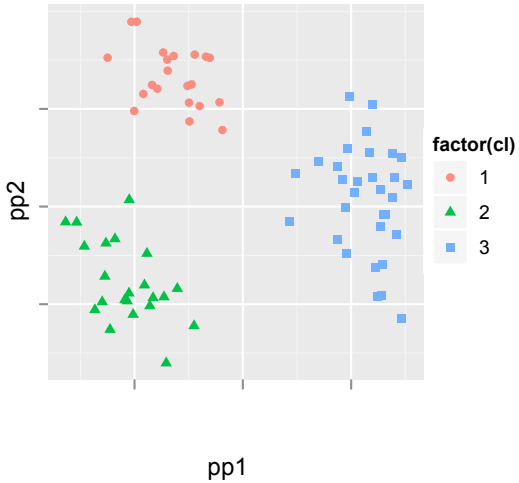
Centroid



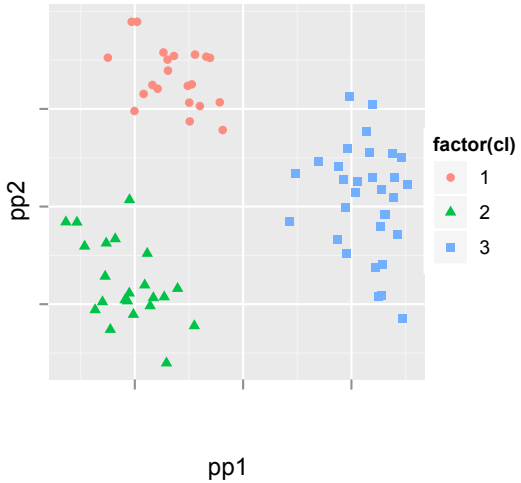
All but single linkage ignore the nuisance points.

Flea beetles (PP dim)

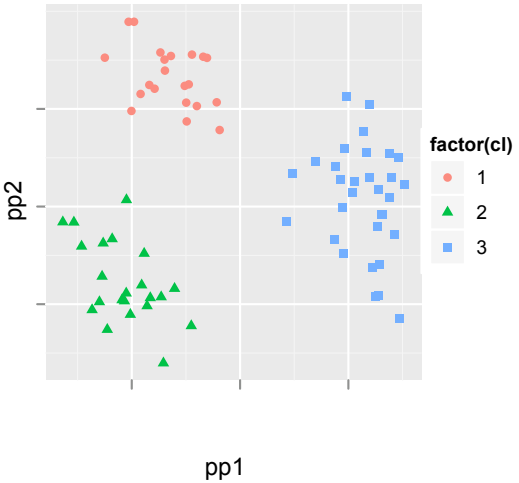
Average



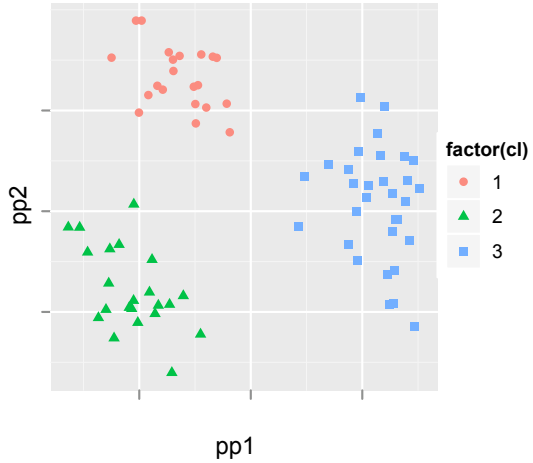
Single



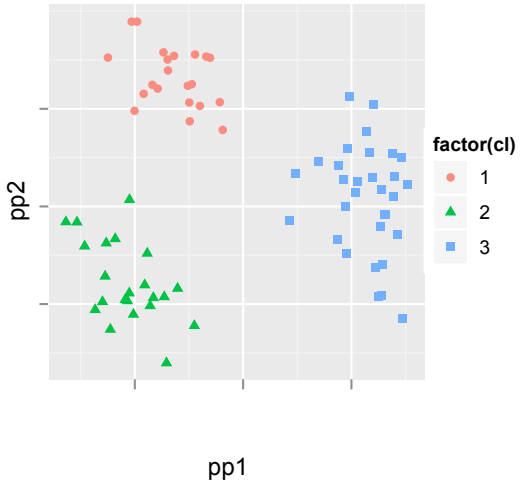
Complete



Ward



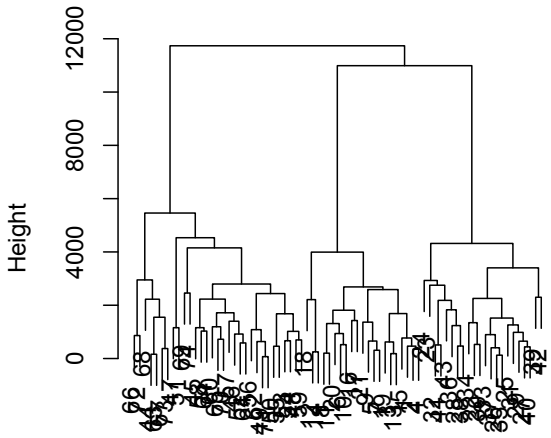
Centroid



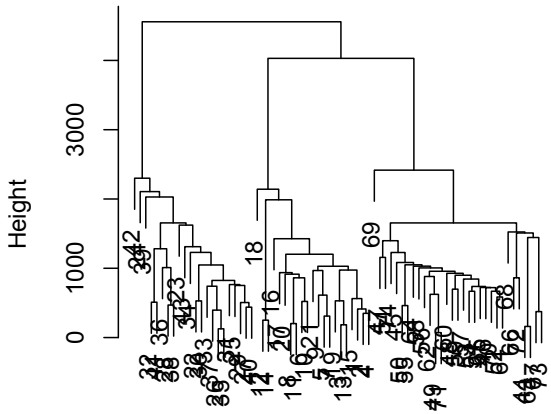
When the nuisance variables are removed all linkage methods see the clusters.

Flea beetles

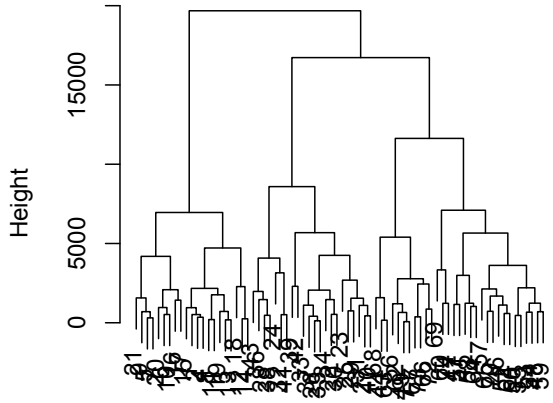
Average



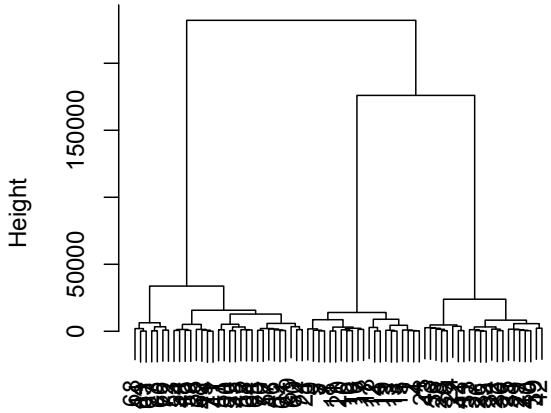
Single



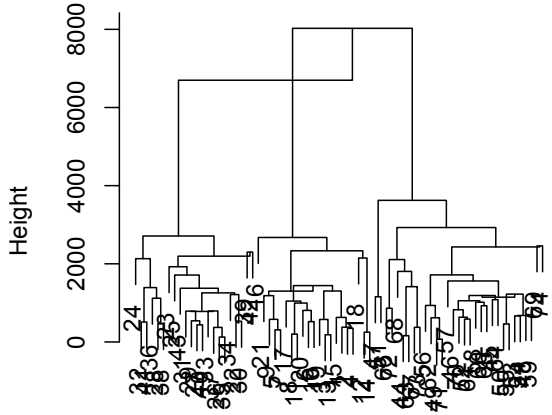
Complete



Ward



Centroid



All methods see 3 clusters. Complete still a little confused between 3 or 4.

k-Means Clustering

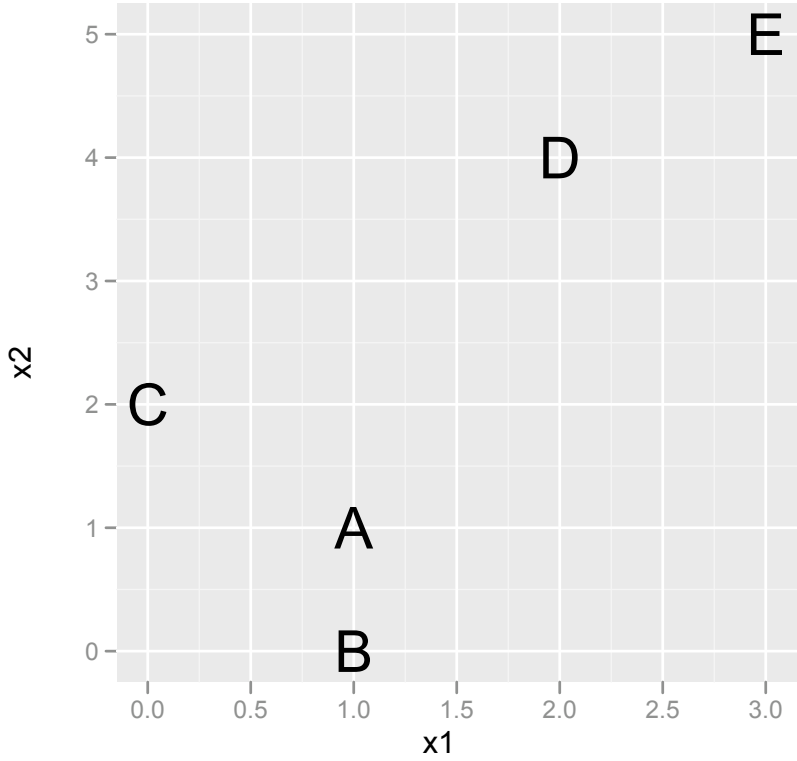
This is an iterative procedure. To use it the number of clusters, k, must be decided first. The stages of the iteration are:

1. 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).
2. Loop over all observations reassigning them to the group with the closest mean.
3. Recompute group means.

Iterate steps 2 and 3 until convergence.

Step 0

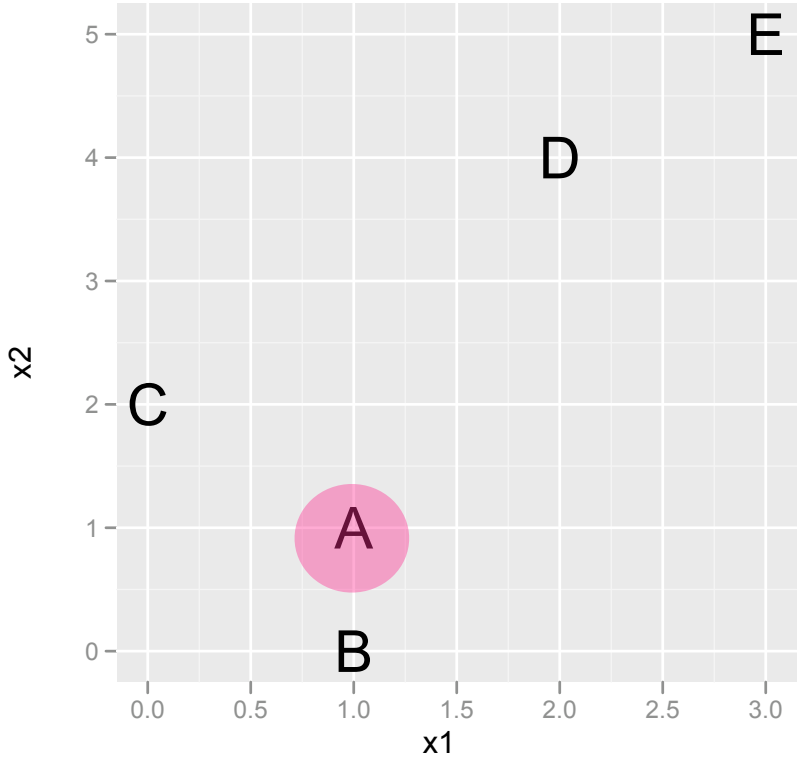
i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5



Use k=2. Suppose A and C are randomly selected as the initial means.

Step 0

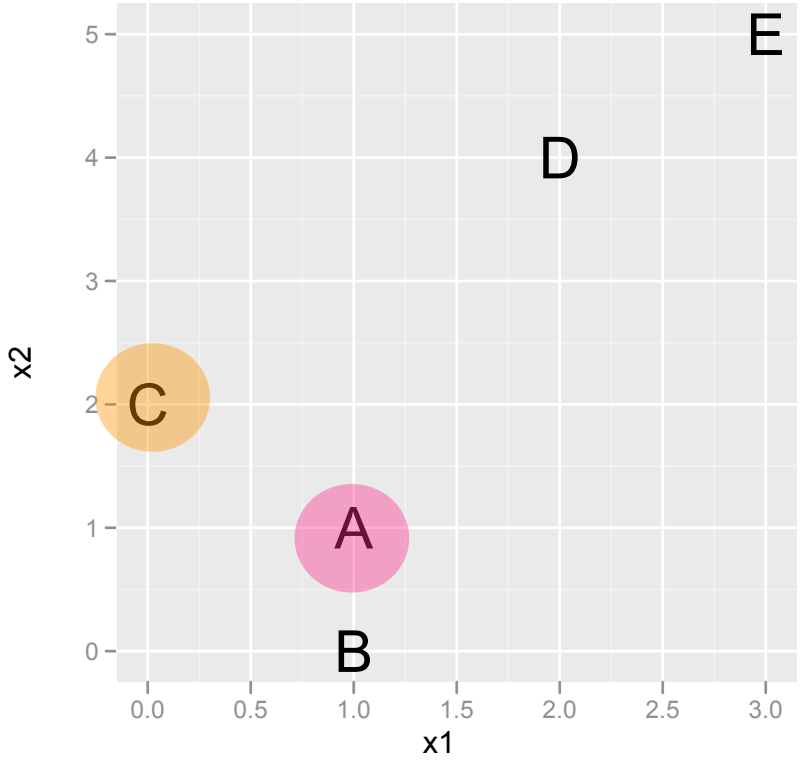
i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5



Use k=2. Suppose A and C are randomly selected as the initial means.

Step 0

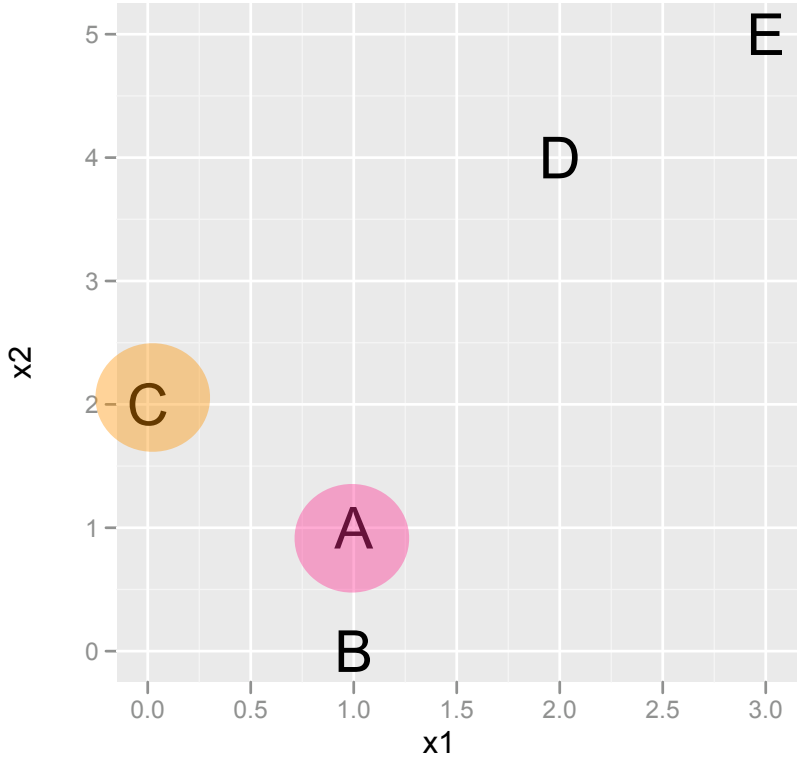
i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5



Use k=2. Suppose A and C are randomly selected as the initial means.

Step 0

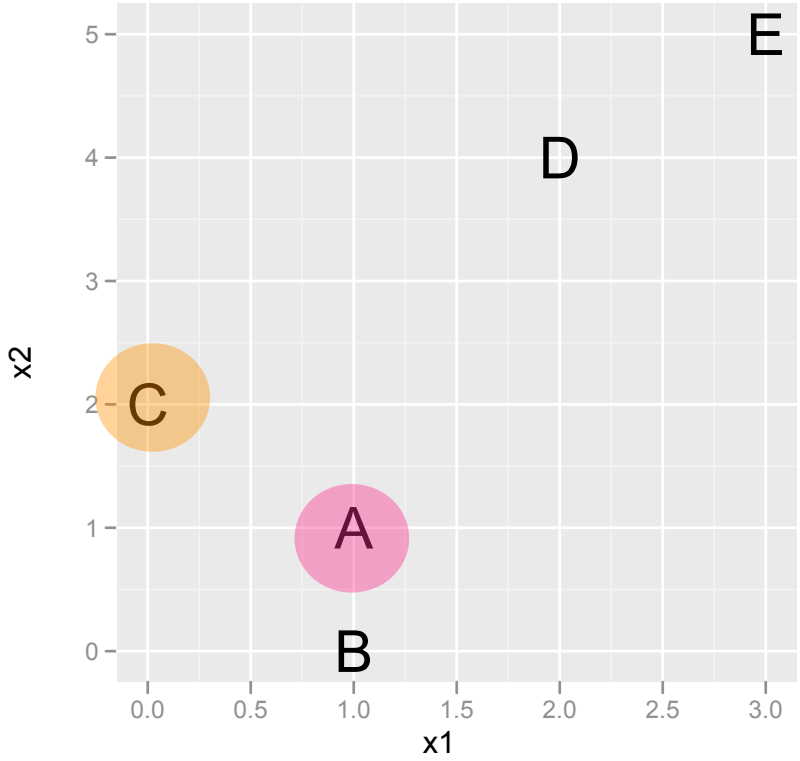
i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5



Use k=2. Suppose A and C are randomly selected as the initial means.

Step 0

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

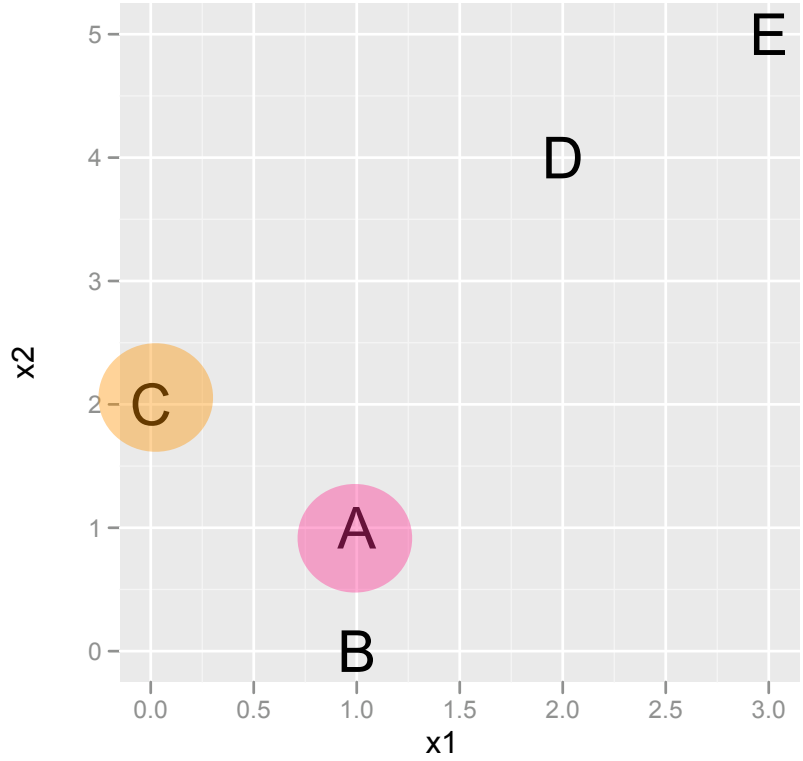


Use k=2. Suppose A and C are randomly selected as the initial means.

Step 0

$\bar{X}_1^0$

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5



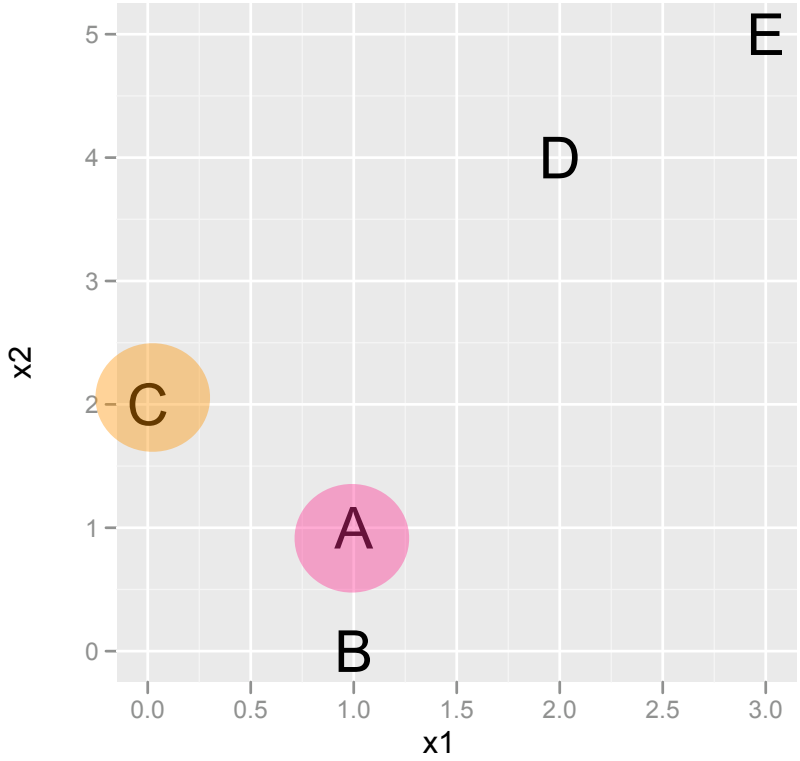
Use k=2. Suppose A and C are randomly selected as the initial means.

Step 0

$\bar{X}_1^0$

$\bar{X}_2^0$

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5



Use k=2. Suppose A and C are randomly selected as the initial means.

Step 1.1

$\bar{X}_1^0$

$\bar{X}_2^0$

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

i	1	2
A	0	1.4
B	1	2.2
C	1.4	0
D	3.2	2.8
E	4.5	4.2

Compute distances between each of the cluster means and all other points.

Step 1.1

i	1	2	Cluster
A	0	1.4	1
B	1	2.2	1
C	1.4	0	2
D	3.2	2.8	2
E	4.5	4.2	2

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

$\bar{X}_1^1$

$\bar{X}_2^1$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 1.1

i	1	2	Cluster
A	0	1.4	1
B	1	2.2	1
C	1.4	0	2
D	3.2	2.8	2
E	4.5	4.2	2

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

$\bar{X}_1^1$

$\bar{X}_2^1$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 1.1

i	1	2	Cluster
A	0	1.4	1
B	1	2.2	1
C	1.4	0	2
D	3.2	2.8	2
E	4.5	4.2	2

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

$\bar{X}_1^1$

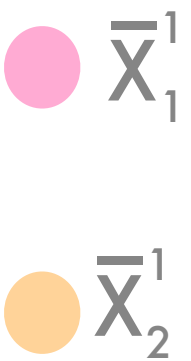
$\bar{X}_2^1$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 1.1

i	1	2	Cluster
A	0	1.4	1
B	1	2.2	1
C	1.4	0	2
D	3.2	2.8	2
E	4.5	4.2	2

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

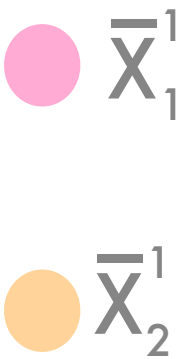


Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 1.1

i	1	2	Cluster
A	0	1.4	1
B	1	2.2	1
C	1.4	0	2
D	3.2	2.8	2
E	4.5	4.2	2

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5



Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 1.1

i	1	2	Cluster
A	0	1.4	1
B	1	2.2	1
C	1.4	0	2
D	3.2	2.8	2
E	4.5	4.2	2

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

$\bar{X}_1^1$

$\bar{X}_2^1$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 1.1

i	1	2	Cluster
A	0	1.4	1
B	1	2.2	1
C	1.4	0	2
D	3.2	2.8	2
E	4.5	4.2	2

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

$\bar{X}_1^1 = (1, 0.5)$

$\bar{X}_2^1$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 1.1

i	1	2	Cluster
A	0	1.4	1
B	1	2.2	1
C	1.4	0	2
D	3.2	2.8	2
E	4.5	4.2	2

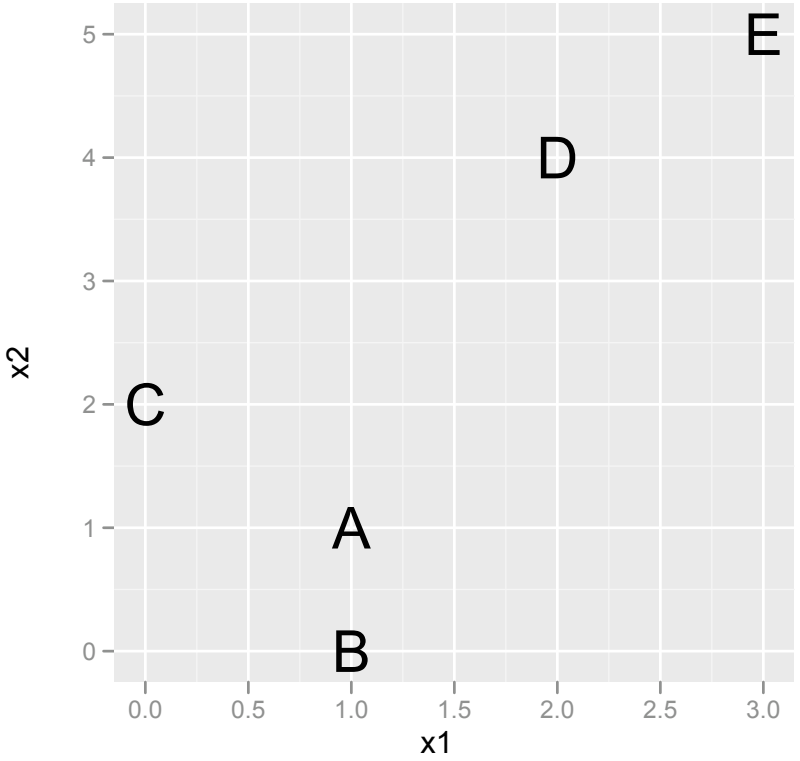
i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

$\bar{X}_1^1 = (1, 0.5)$

$\bar{X}_2^1 = (1.7, 3.7)$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 1.1 - Plots

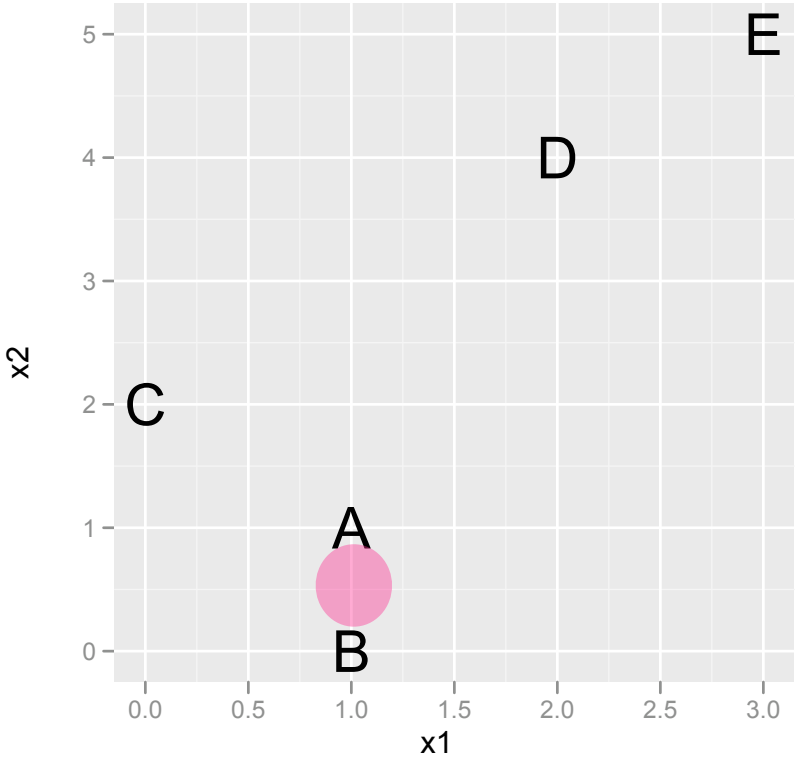


$\bar{X}_1^1 = (1, 0.5)$

$\bar{X}_2^1 = (1.7, 3.7)$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 1.1 - Plots

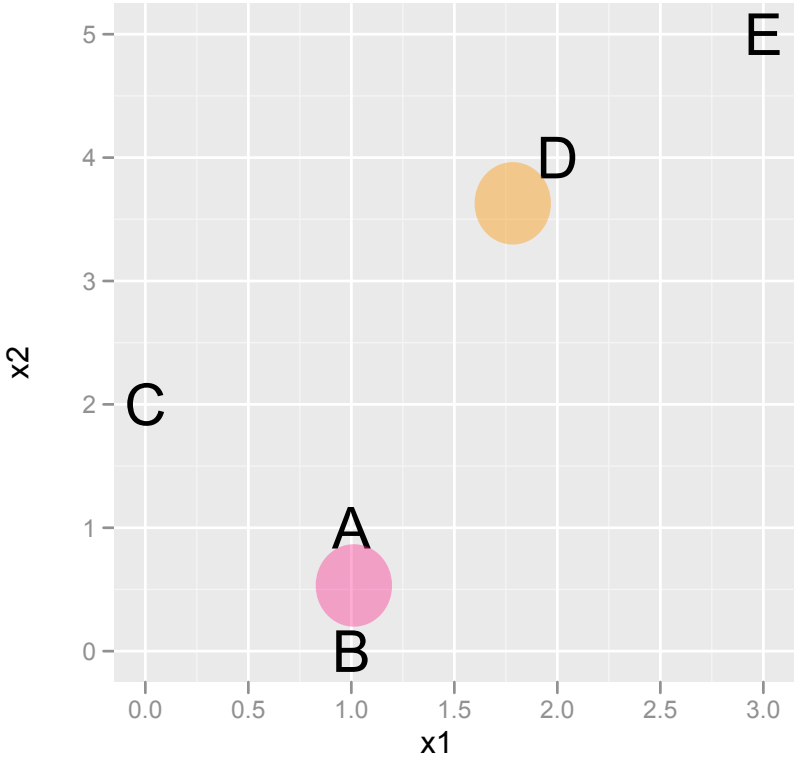


$\bar{X}_1^1 = (1, 0.5)$

$\bar{X}_2^1 = (1.7, 3.7)$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 1.1 - Plots

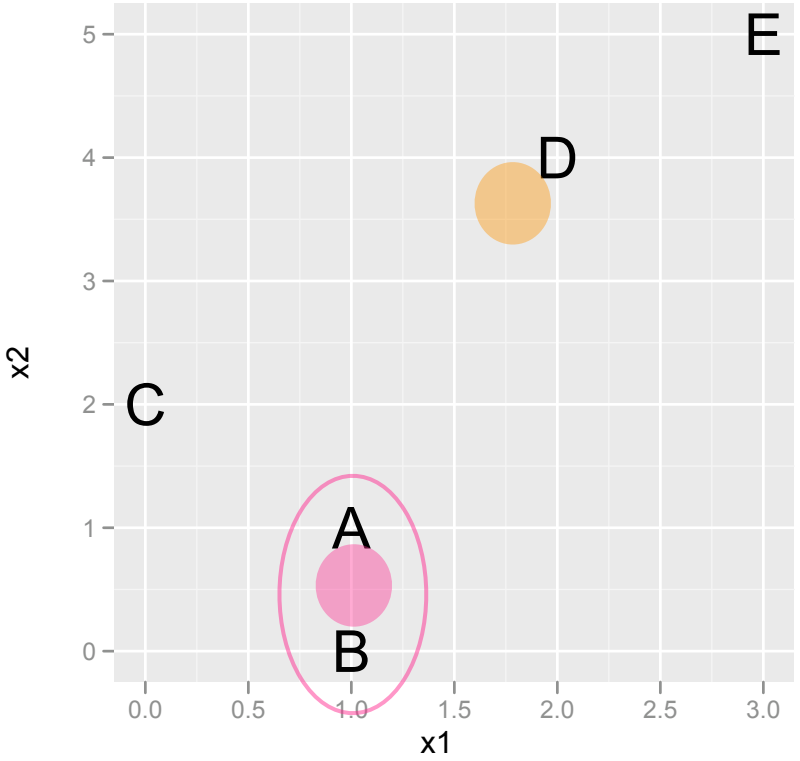


$\bar{X}_1^1 = (1, 0.5)$

$\bar{X}_2^1 = (1.7, 3.7)$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 1.1 - Plots

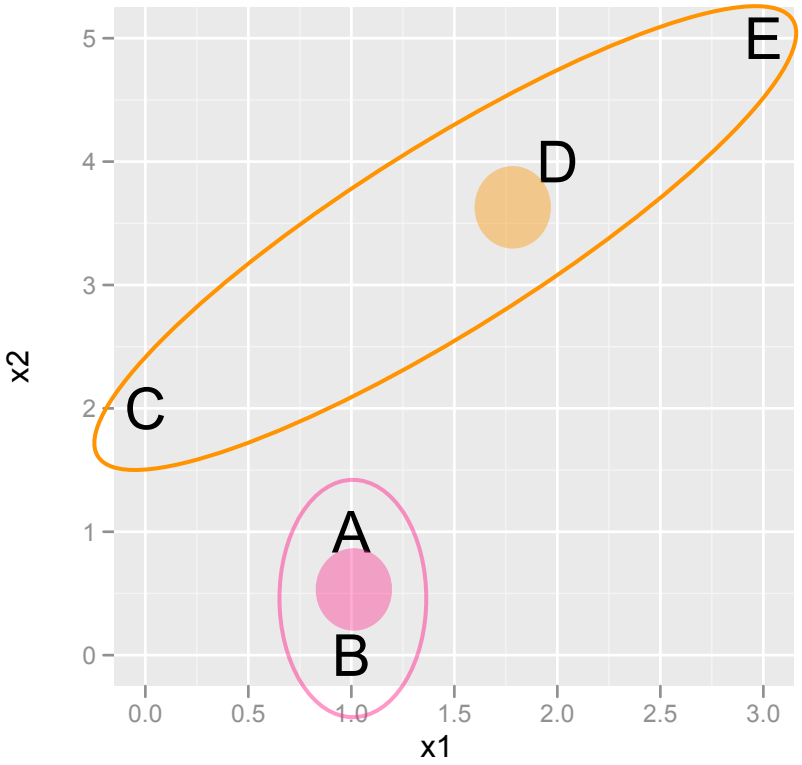


$\bar{X}_1^1 = (1, 0.5)$

$\bar{X}_2^1 = (1.7, 3.7)$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 1.1 - Plots



$\bar{X}_1^1 = (1, 0.5)$

$\bar{X}_2^1 = (1.7, 3.7)$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 2.1

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

$\bar{X}_1^1 = (1, 0.5)$

$\bar{X}_2^1 = (1.7, 3.7)$

i	1	2
A	0.5	2.7
B	0.5	3.7
C	1.8	2.4
D	3.6	0.5
E	4.9	1.9

Compute distances between each of the cluster means and all other points.

Step 2.1

i	1	2	Cluster
A	0.5	2.7	1
B	0.5	3.7	1
C	1.8	2.4	1
D	3.6	0.5	2
E	4.9	1.9	2

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

$\bar{X}_1^2$

$\bar{X}_2^2$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 2.1

i	1	2	Cluster
A	0.5	2.7	1
B	0.5	3.7	1
C	1.8	2.4	1
D	3.6	0.5	2
E	4.9	1.9	2

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

$\bar{X}_1^2$

$\bar{X}_2^2$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 2.1

i	1	2	Cluster
A	0.5	2.7	1
B	0.5	3.7	1
C	1.8	2.4	1
D	3.6	0.5	2
E	4.9	1.9	2

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

$\bar{X}_1^2$

$\bar{X}_2^2$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 2.1

i	1	2	Cluster
A	0.5	2.7	1
B	0.5	3.7	1
C	1.8	2.4	1
D	3.6	0.5	2
E	4.9	1.9	2

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

$\bar{X}_1^2$

$\bar{X}_2^2$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 2.1

i	1	2	Cluster
A	0.5	2.7	1
B	0.5	3.7	1
C	1.8	2.4	1
D	3.6	0.5	2
E	4.9	1.9	2

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

$\bar{X}_1^2$

$\bar{X}_2^2$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 2.1

i	1	2	Cluster
A	0.5	2.7	1
B	0.5	3.7	1
C	1.8	2.4	1
D	3.6	0.5	2
E	4.9	1.9	2

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

$\bar{X}_1^2$

$\bar{X}_2^2$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 2.1

i	1	2	Cluster
A	0.5	2.7	1
B	0.5	3.7	1
C	1.8	2.4	1
D	3.6	0.5	2
E	4.9	1.9	2

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

$\bar{X}_1^2 = (0.7,1)$

$\bar{X}_2^2$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 2.1

i	1	2	Cluster
A	0.5	2.7	1
B	0.5	3.7	1
C	1.8	2.4	1
D	3.6	0.5	2
E	4.9	1.9	2

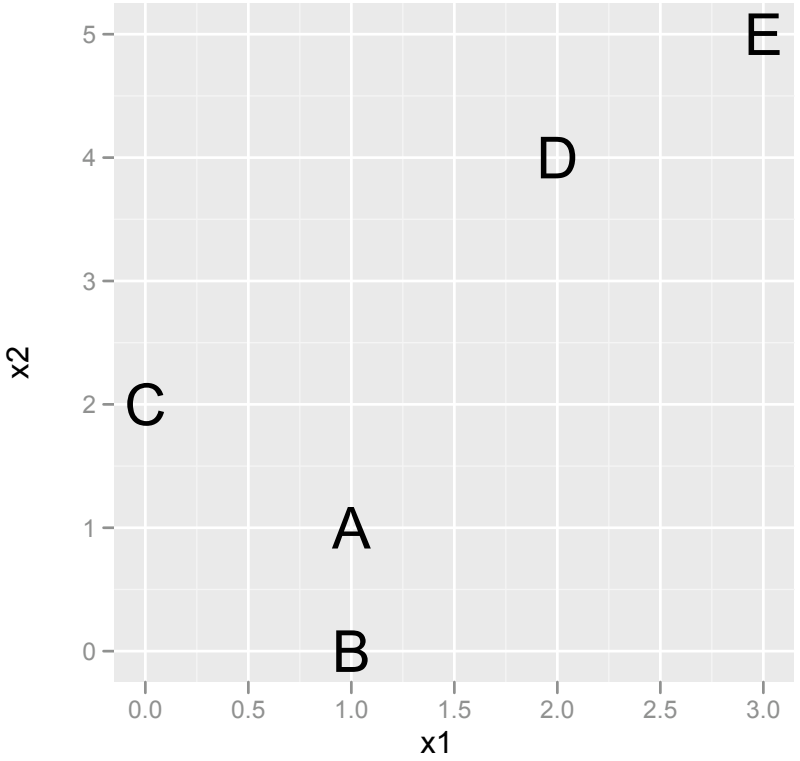
i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

$\bar{X}_1^2 = (0.7, 1)$

$\bar{X}_2^2 = (2.5, 4.5)$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 2.1 - Plots

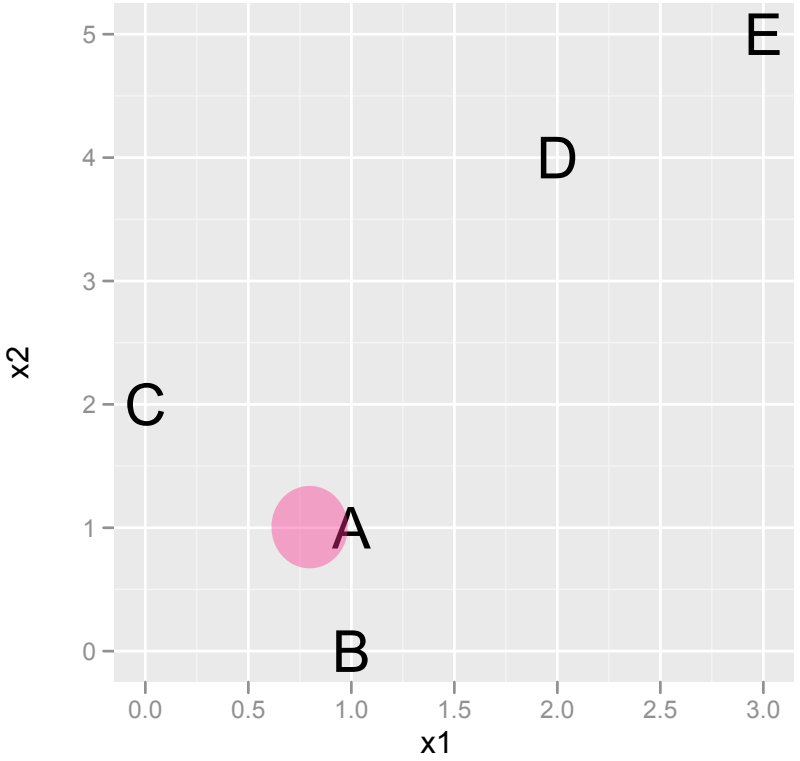


$\bar{X}_1^2 = (0.7, 1)$

$\bar{X}_2^2 = (2.5, 4.5)$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 2.1 - Plots

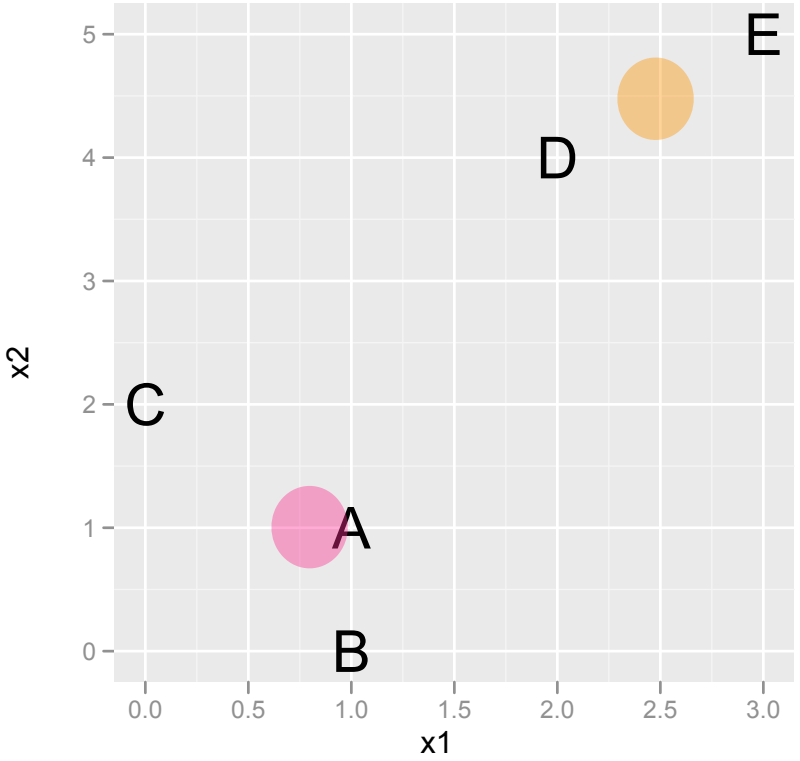


$\bar{X}_1^2 = (0.7, 1)$

$\bar{X}_2^2 = (2.5, 4.5)$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 2.1 - Plots

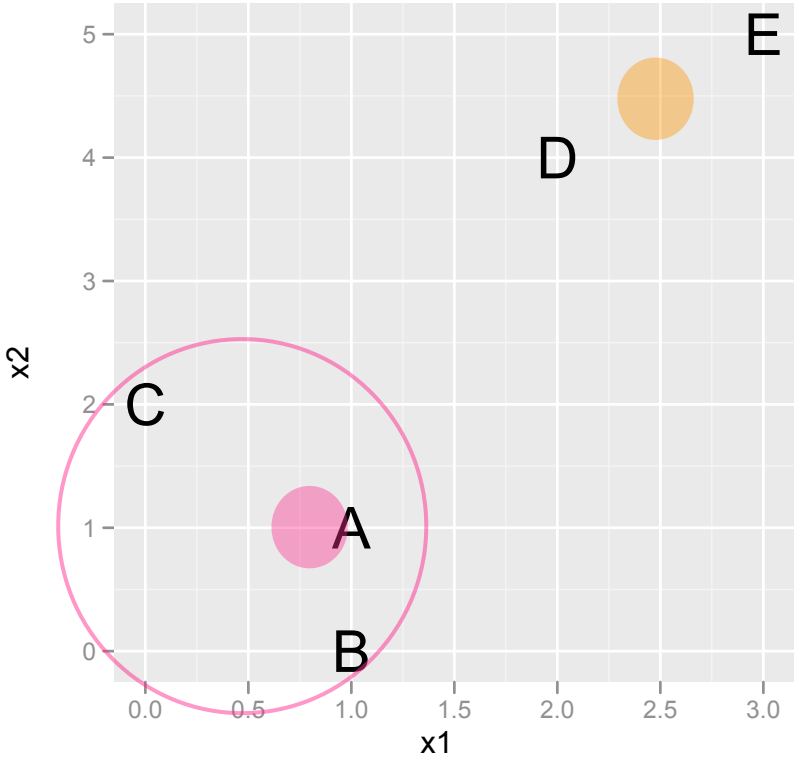


$\bar{X}_1^2 = (0.7, 1)$

$\bar{X}_2^2 = (2.5, 4.5)$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 2.1 - Plots

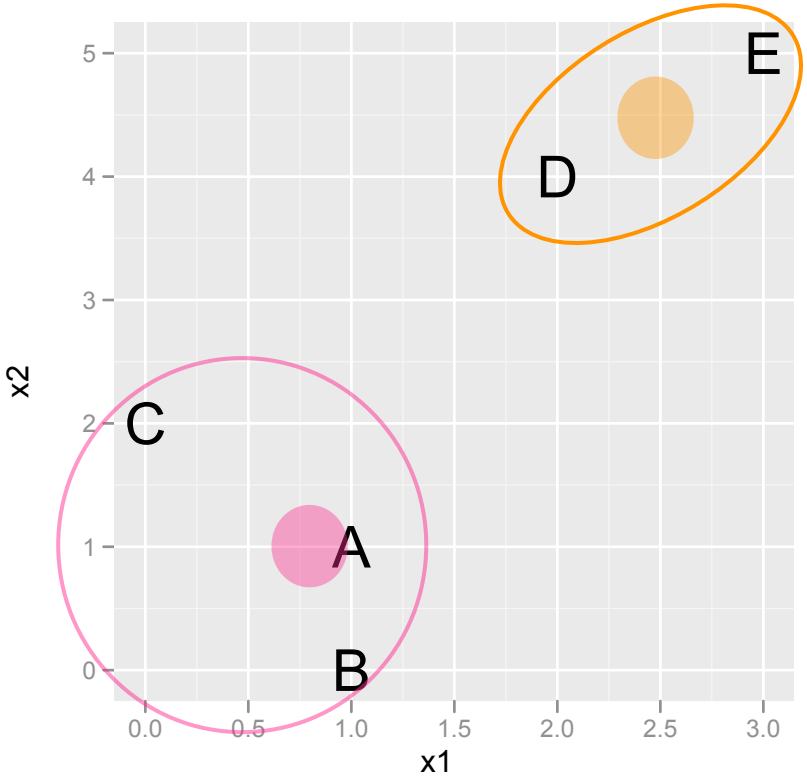


$\bar{X}_1^2 = (0.7, 1)$

$\bar{X}_2^2 = (2.5, 4.5)$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 2.1 - Plots

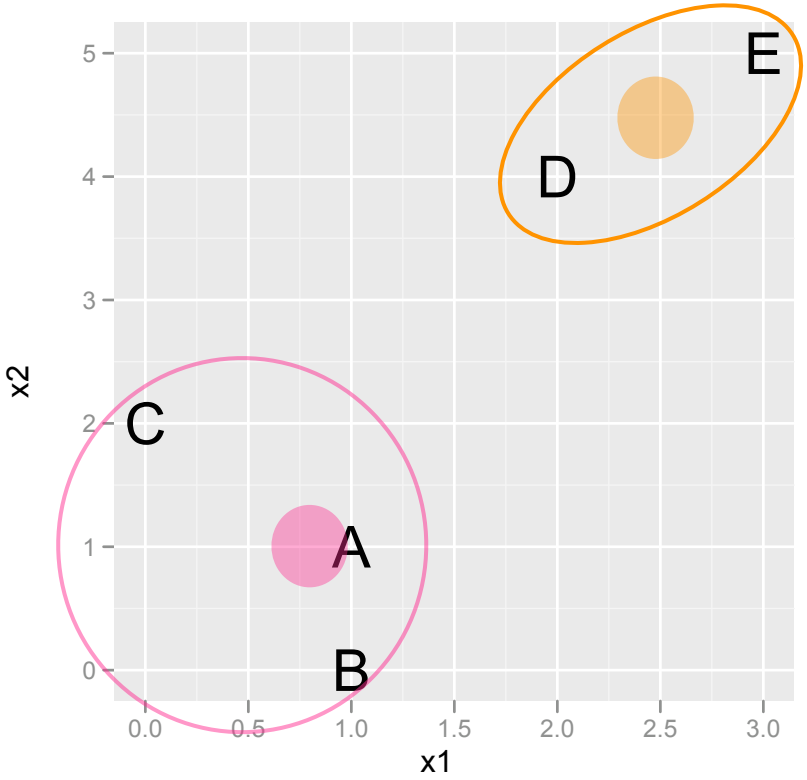


$\bar{X}_1^2 = (0.7, 1)$

$\bar{X}_2^2 = (2.5, 4.5)$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 3



$\bar{X}_1^2 = (0.7, 1)$

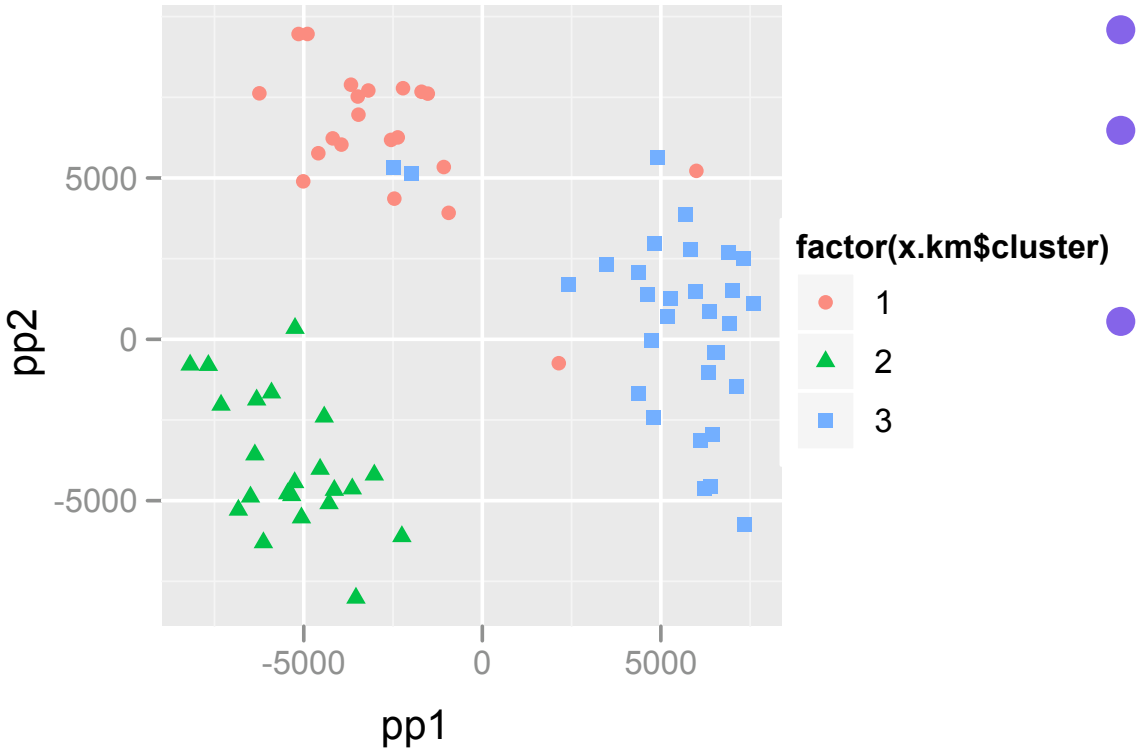
$\bar{X}_2^2 = (2.5, 4.5)$

Algorithm has converged - re-calculating distances, reassigning cases to clusters results in no change. This is the final solution.

k-Means - Initialization

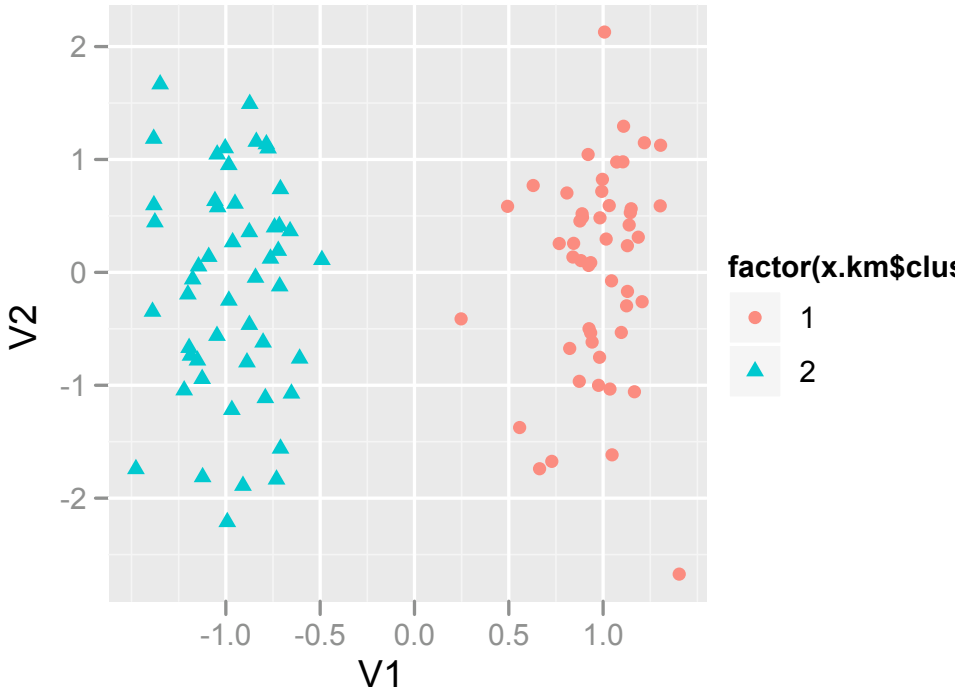
- The algorithm needs to be initialized by choosing k initial means.
- Approaches:
  1. Randomly choose k points from the data set to act as the initial means.
  2. First do hierarchical clustering, decide on k, and use the means of these clusters as the initial k-means.
- Initialization can change the final result.
- If k is not known, re-run for several different k.

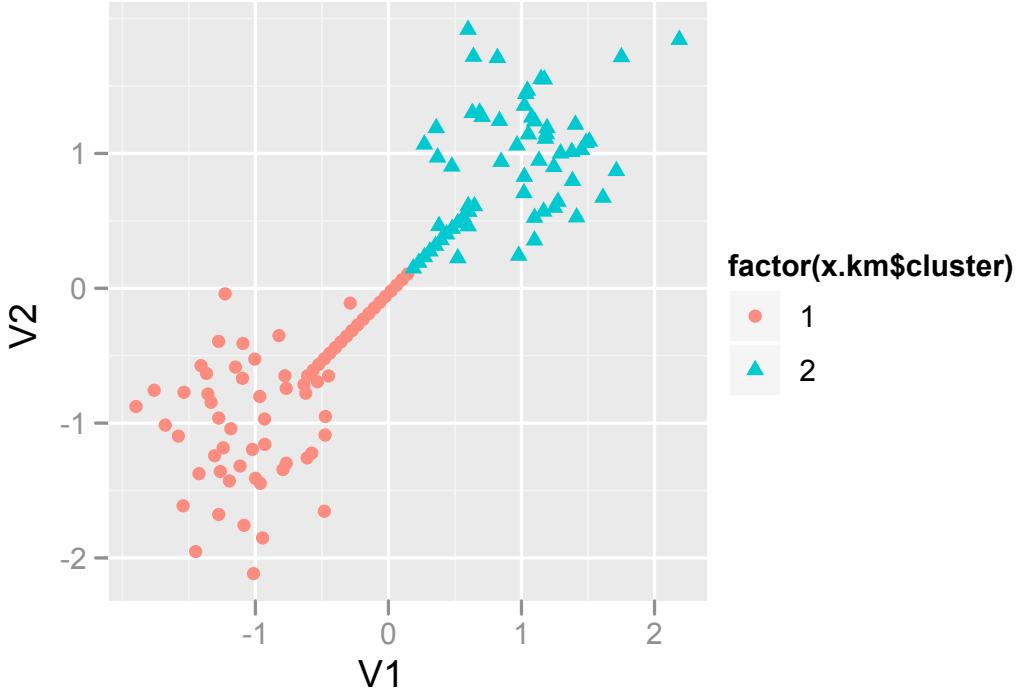
Examples



- Flea beetles
- Several cases are confused.
- Why would k-means have trouble with this data?

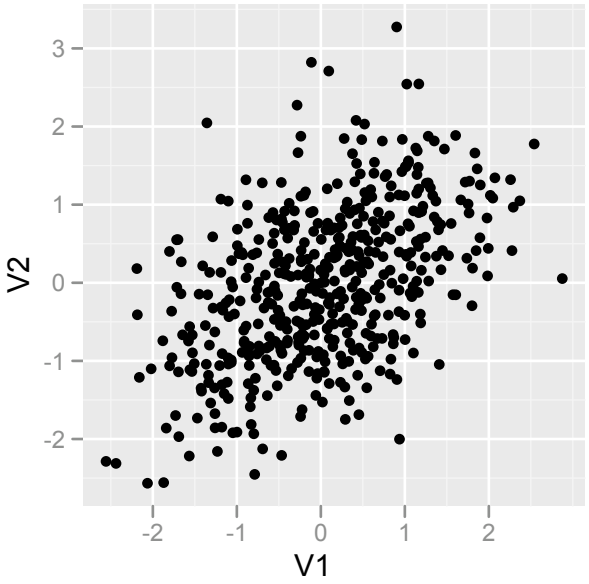
Example





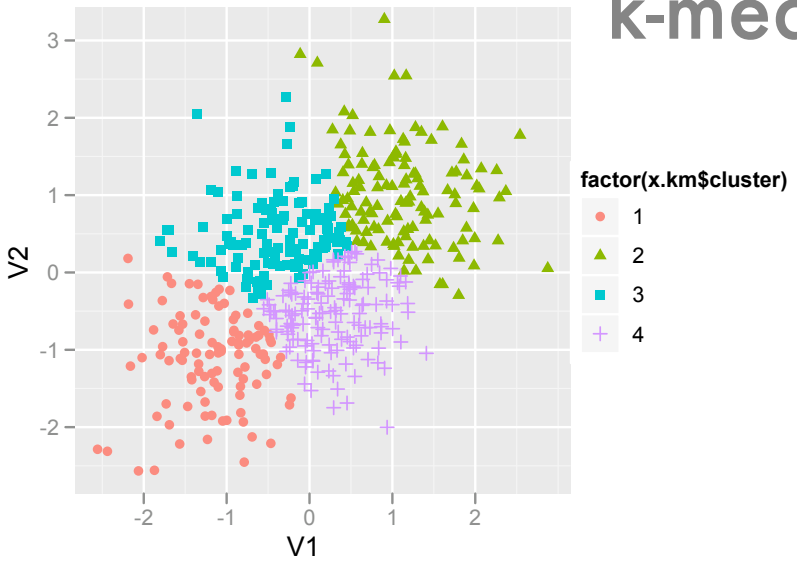
- k-means does not handle nuisance variables well, but surprisingly does well with these data sets.

Example - partitioning

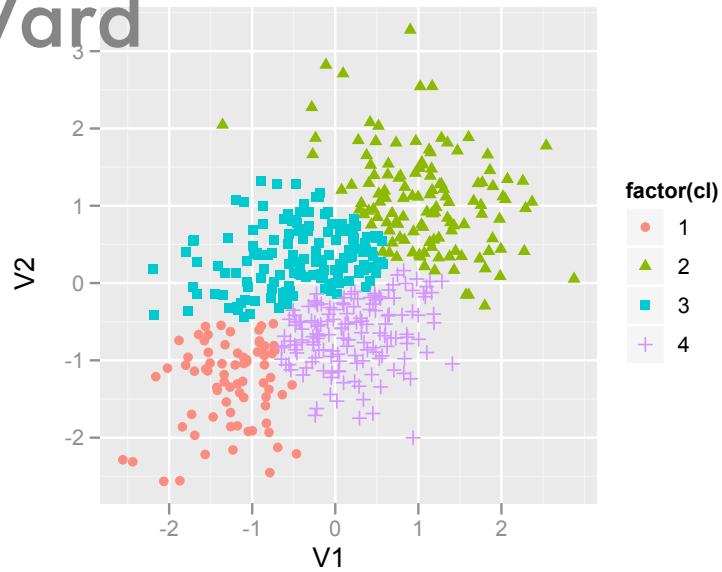


- Many clustering tasks involve partitioning data into chunks.
- There may not be natural clusters.

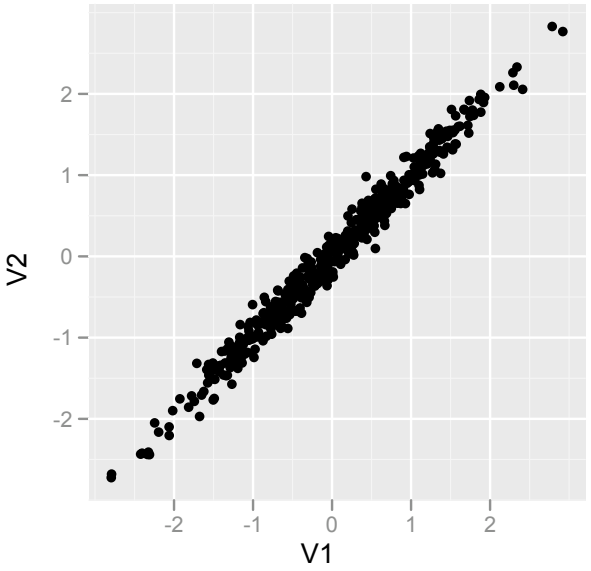
k-means



Ward

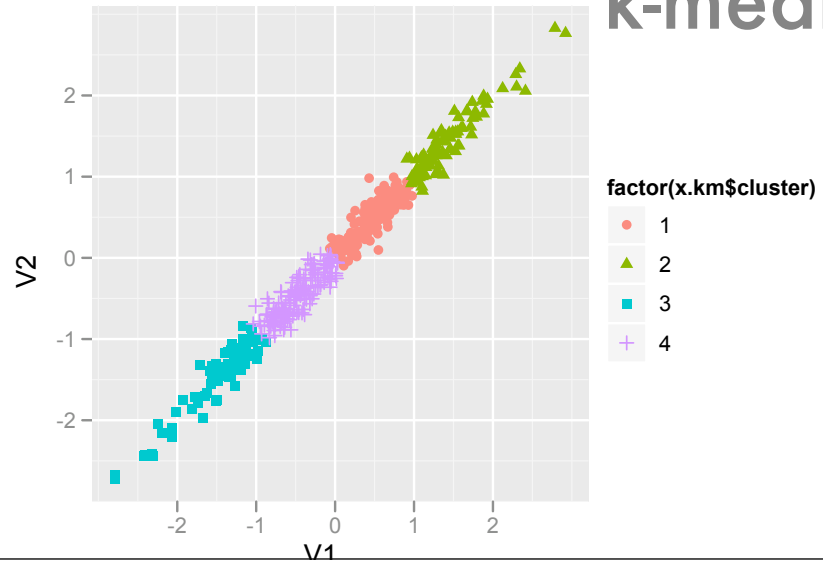


Example - partitioning

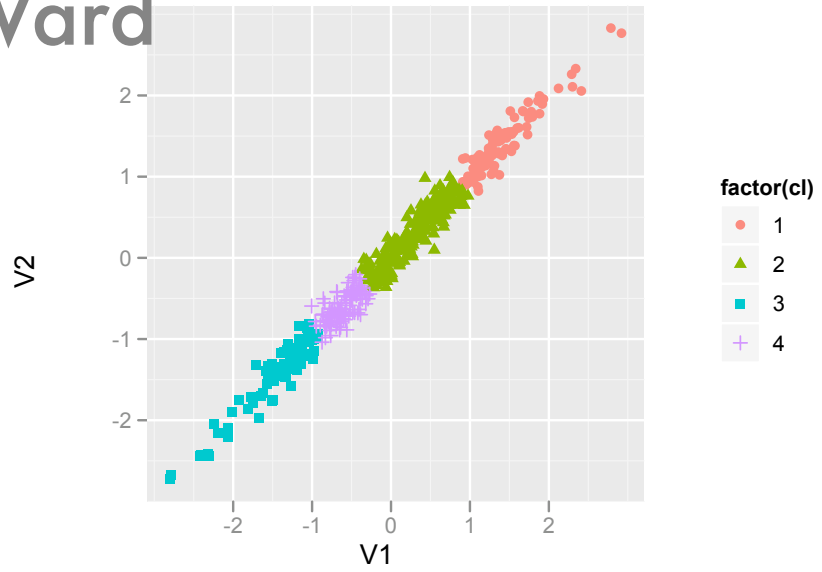


- Covariance matters in the way the data gets partitioned.

k-means



Ward



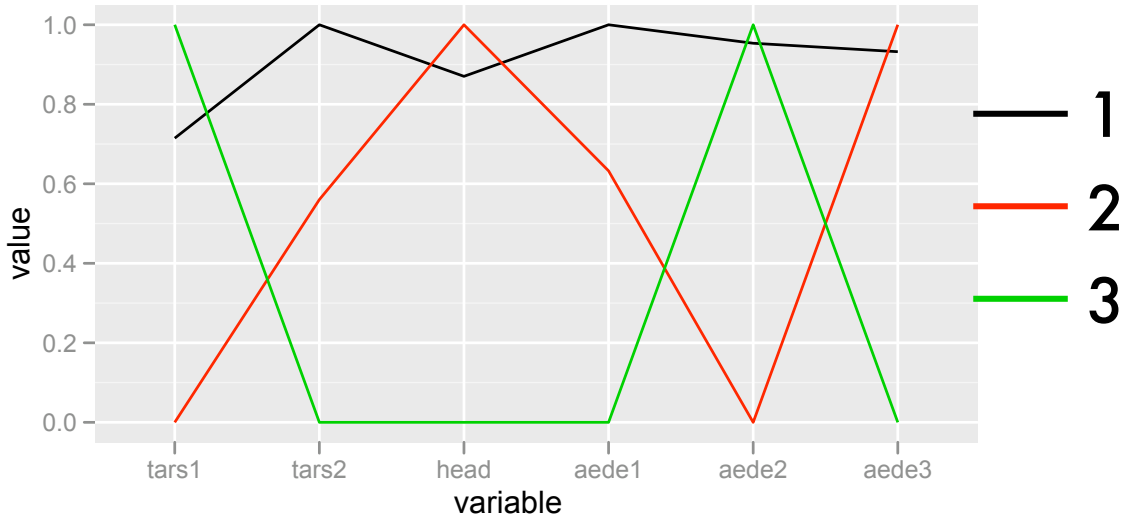
Summarizing results

Need to show how the clusters differ from each other:

- Tabulate the summary statistics for each cluster.
- Make separate plots for each cluster, using same scale
- Plot the means on one plot

Example

cluster	tars1	tars2	head	aede1	aede2	aede3
mean 1	183.10	129.62	51.24	146.19	14.10	104.86
sd 1	12.14	7.16	2.23	5.63	0.89	6.18
mean 2	138.23	125.09	51.59	138.27	10.09	106.59
sd 2	9.34	8.55	2.84	4.14	0.97	5.85
mean 3	201.00	119.32	48.87	124.65	14.29	81.00
sd 3	14.90	6.65	2.35	4.62	1.10	8.93

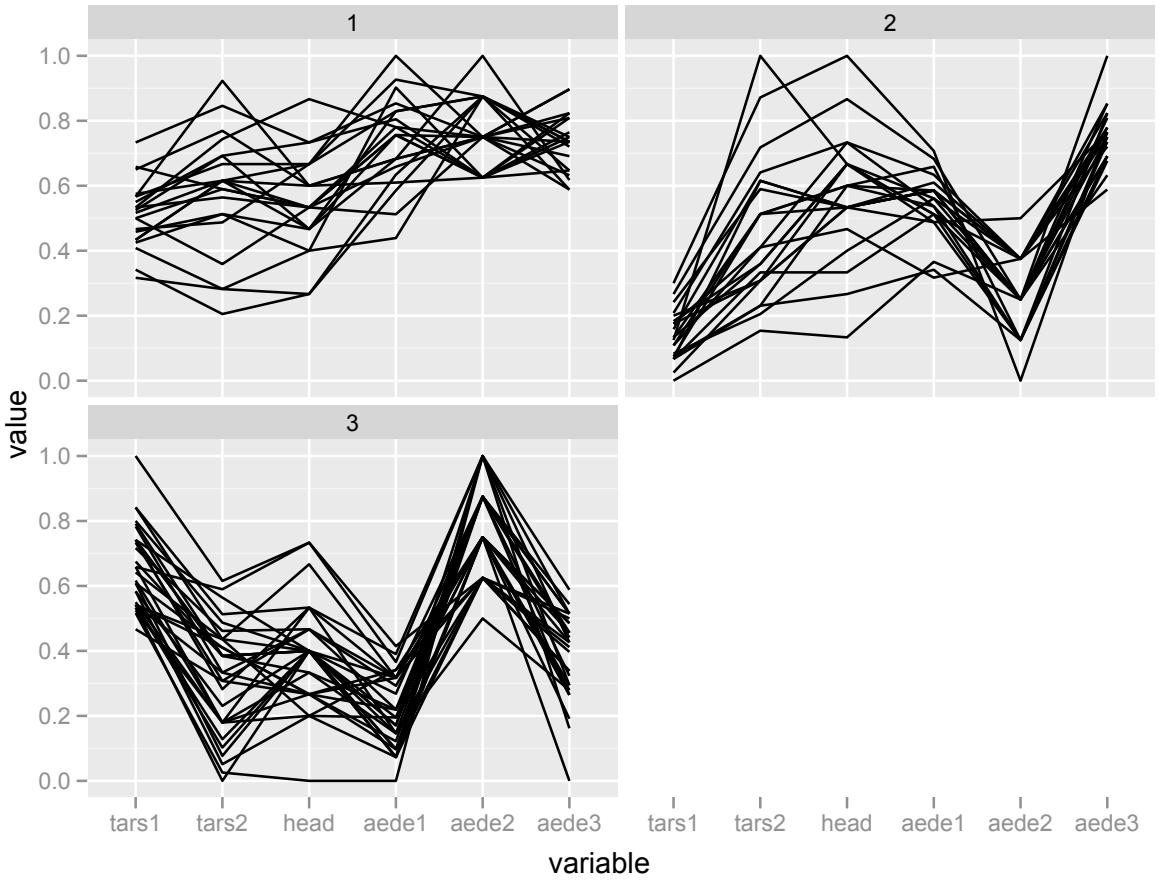


Cluster 1 has high values on all variables.

Cluster 2 has low values for tars1 and aede2, high values of head and aede3.

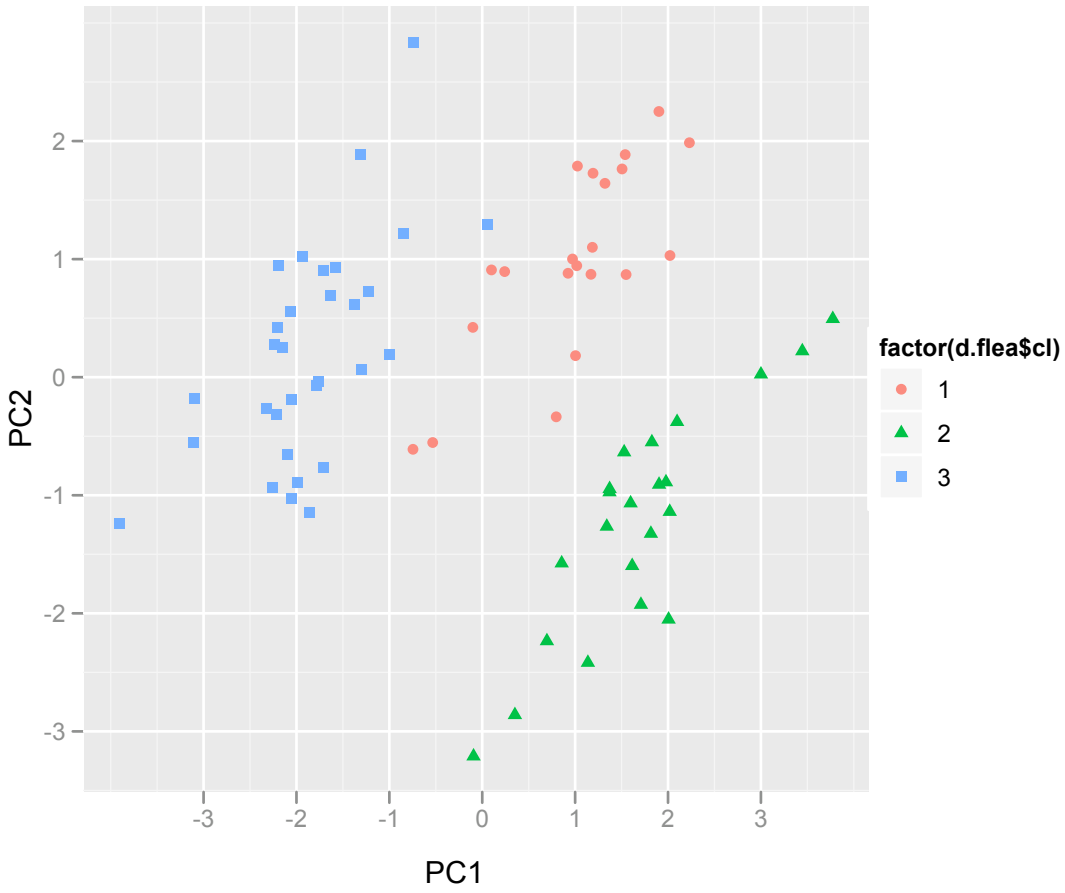
Cluster 3 has high values of tars 1 and aede2, but low values of all other variables.

Example



Plotting all of the data shows the variability in each cluster.

Example



Plotting the clusters in a low-dimensional representation like the first two principal components can also help evaluate the clusters.

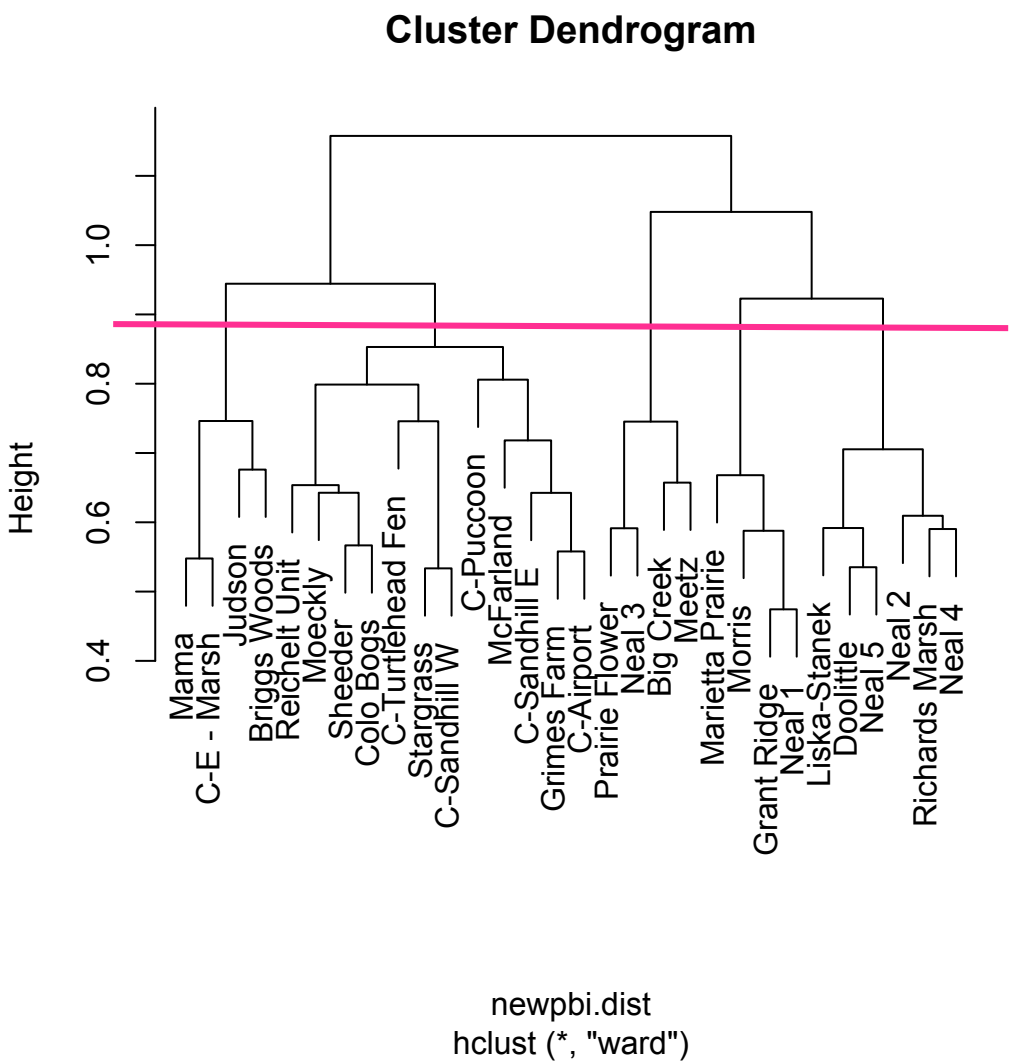
Example

- Iowa prairies
- Which sites are similar?
- Use Canberra distance:

$$d_{A,B} = \frac{1}{\#\text{non-zero entries}} \sum_{j=1}^{p} \frac{|A_j - B_j|}{|A_j + B_j|}$$

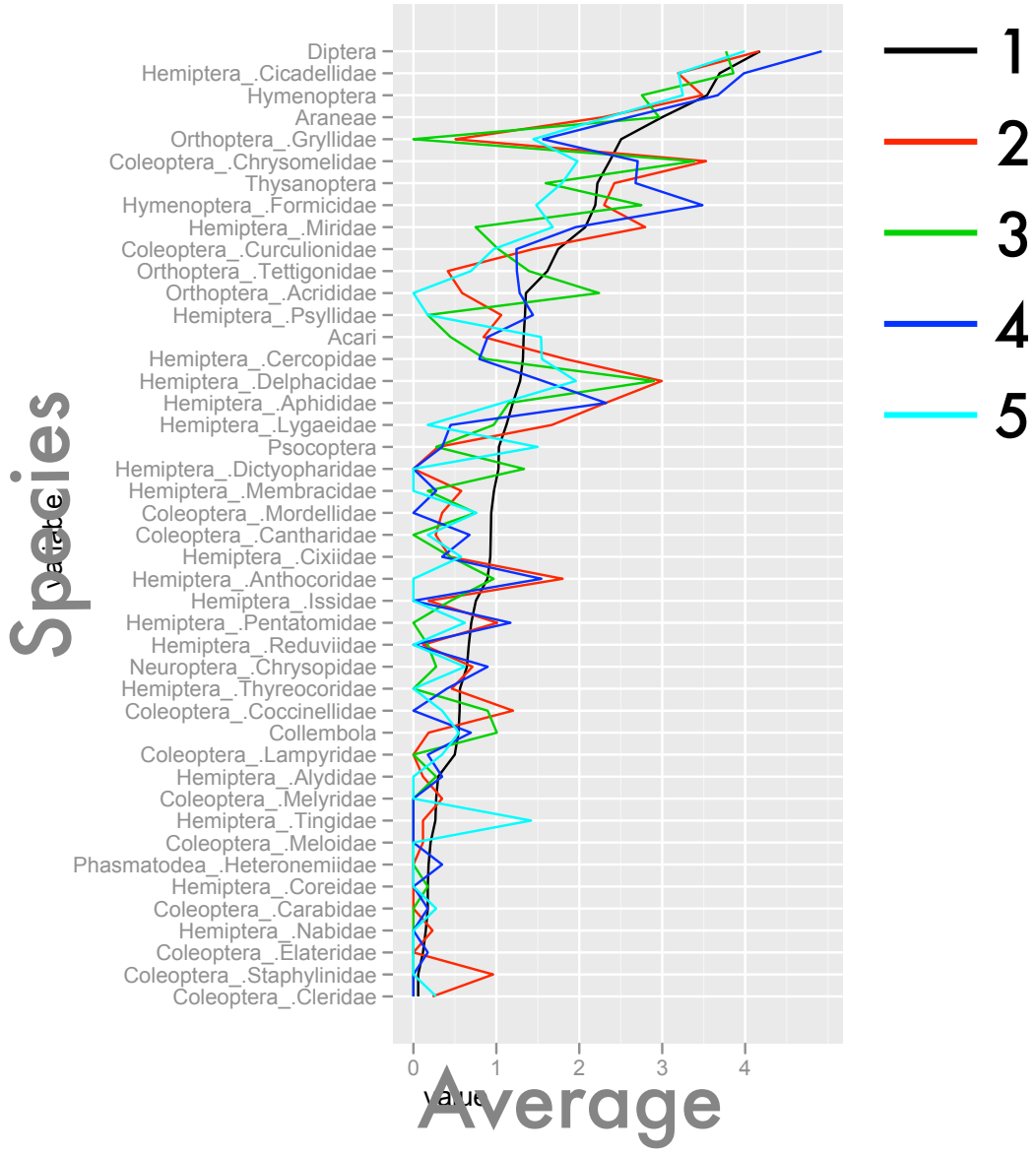
$$\text{where } \frac{0}{0} = 0$$

Example



- Dendrogram suggests 5 clusters?
-

Example



Differences between clusters on:

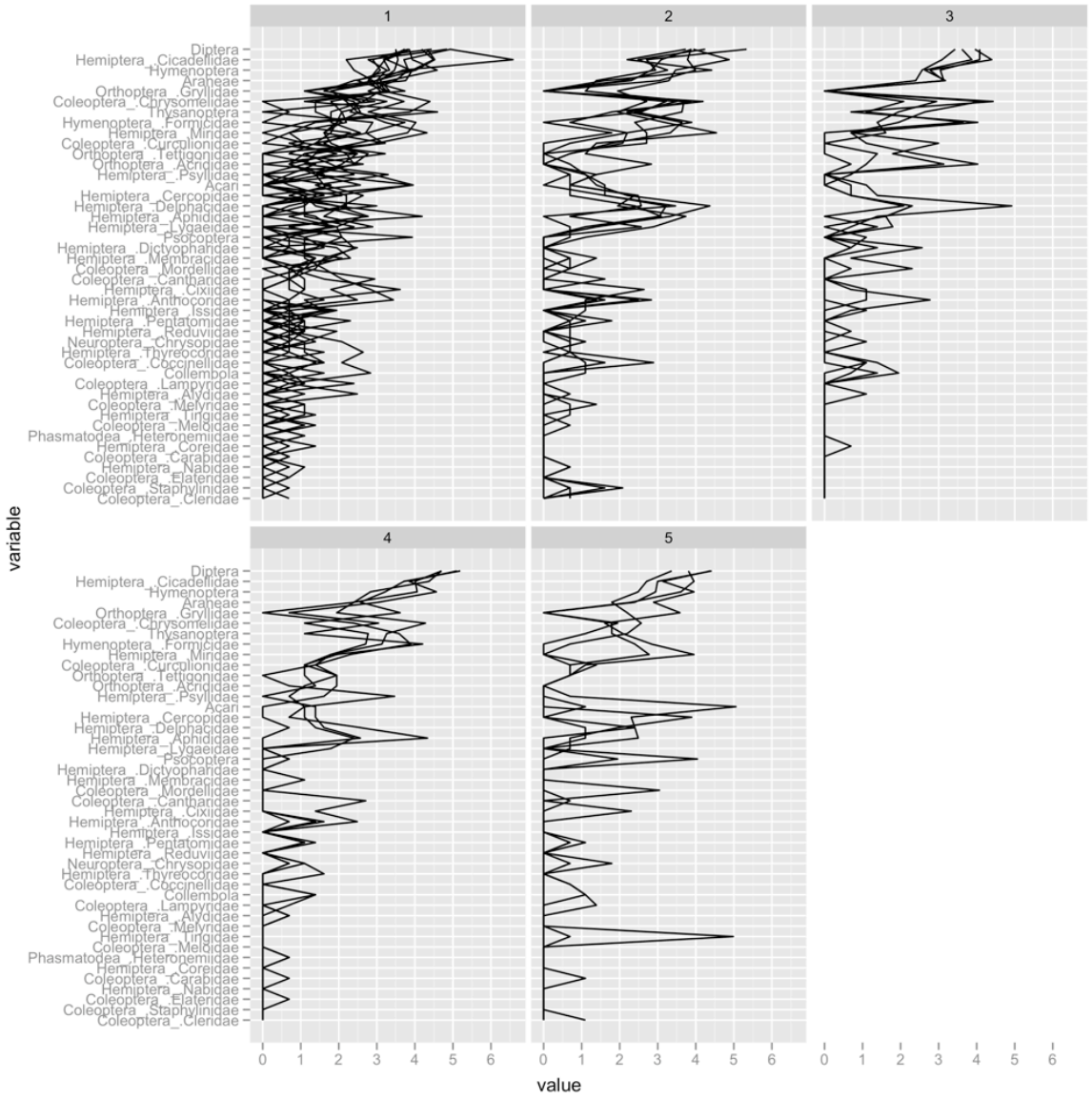
- Orthoptera_Gryllidae (2,3 low; 4,5 medium; 1 high)
- Orthoptera_Acridae (5 low; 2 low'ish; 1,4 med)
- .....
- Cluster 5 has Hemiptera_Tingidae, absent elsewhere.
- Cluster 2 has Coleoptera_Staphylinidae, absent elsewhere.

Example

- List clusters

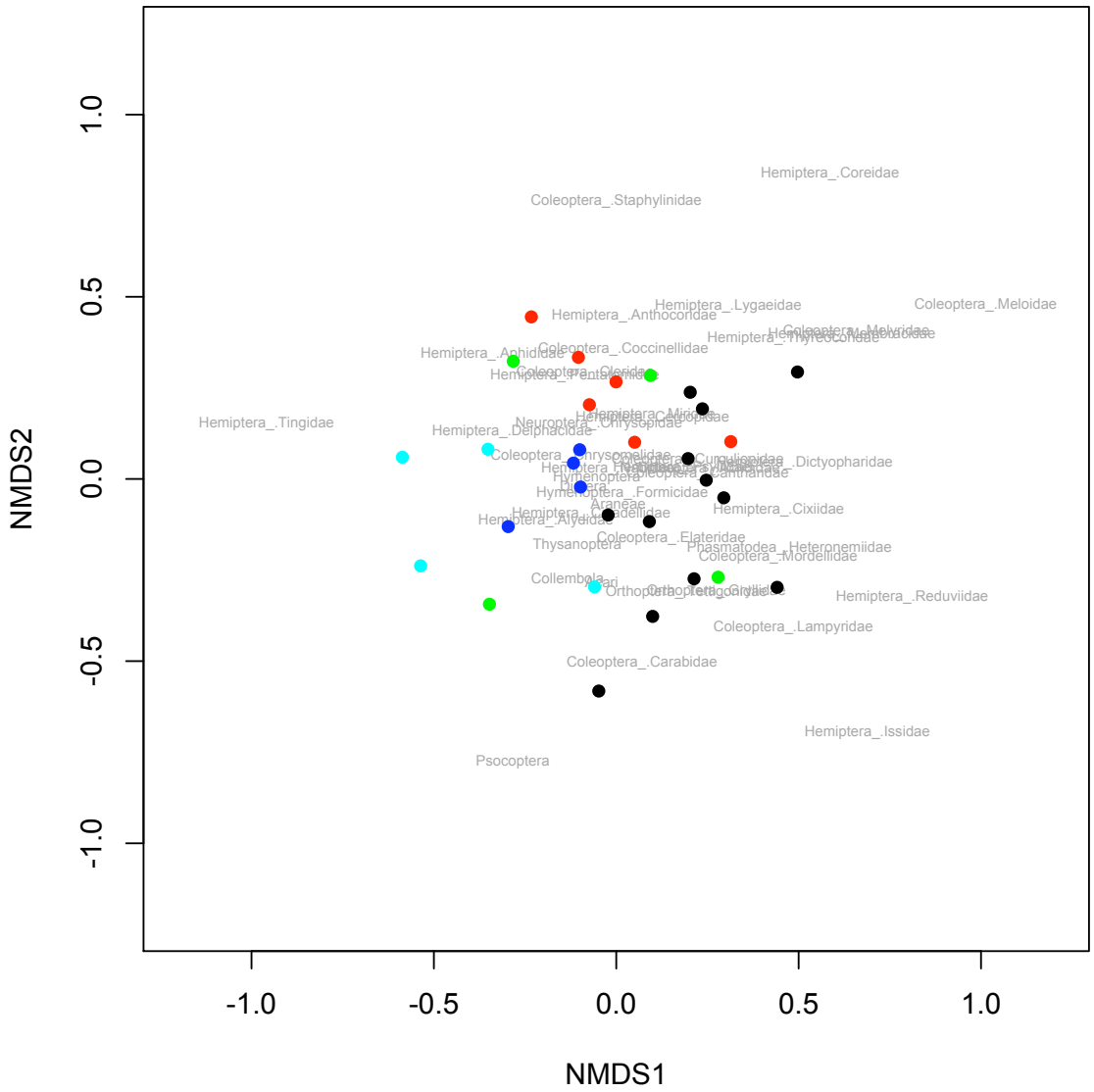
1	2	3	4	5
C-Turtlehead Fen, Moeckly, Reichelt Unit, Sheeder, Colo Bogs, Grimes Farm, McFarland Stargrass, C-Airport, C-Puccoon, C-Sandhill E, C-Sandhill W	Doolittle, Liska-Stanek, Richards Marsh, Neal 2, Neal 4, Neal 5	Judson, Mama, Briggs Woods, C-E - Marsh	Marietta Prairie, Morris, Grant Ridge, Neal 1	Big Creek, Meetz, Prairie Flower, Neal 3

Example



- Really need to take out the nuisance variables to digest the differences.
- Try to write down a few species that have differences between clusters.

Example

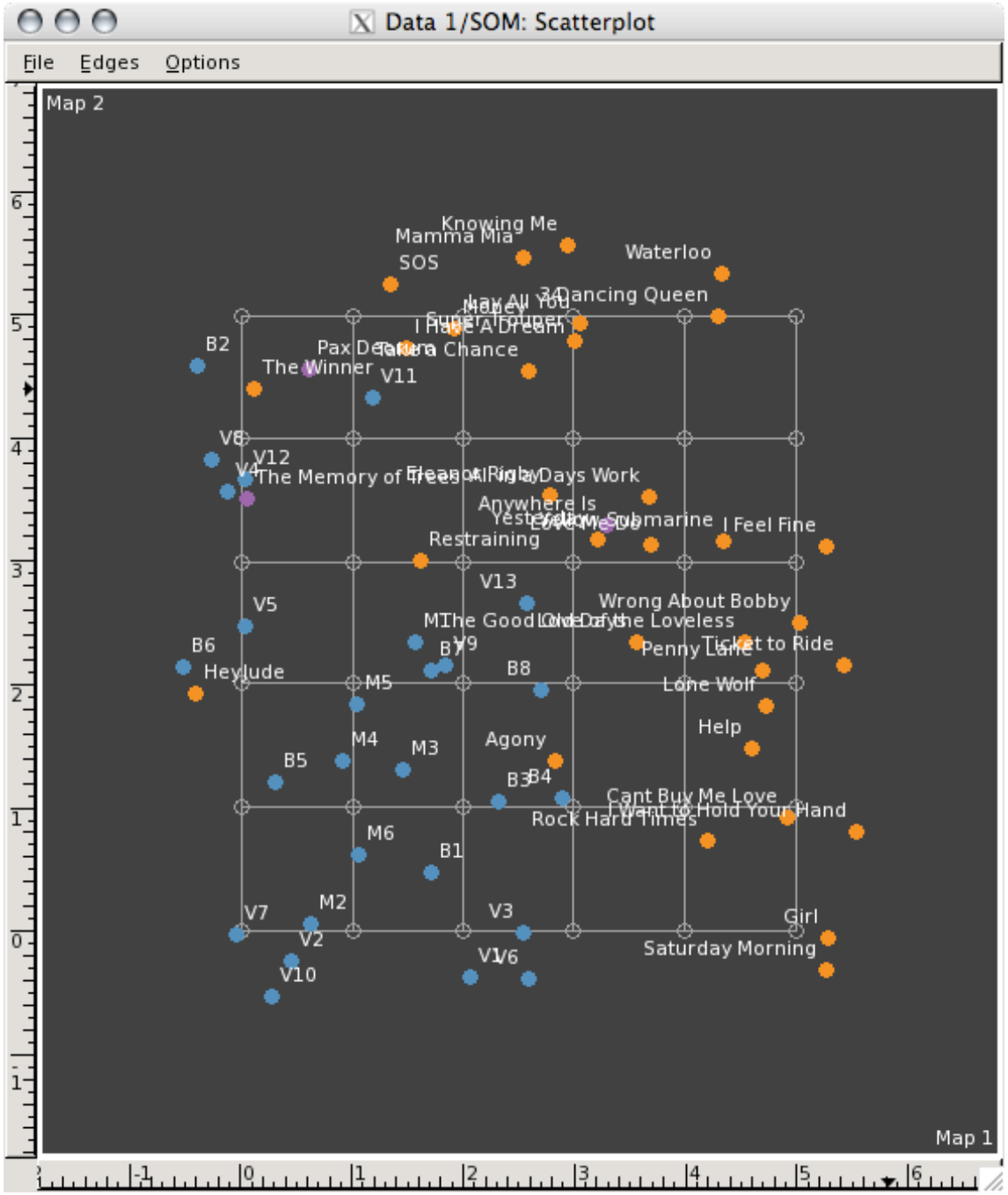


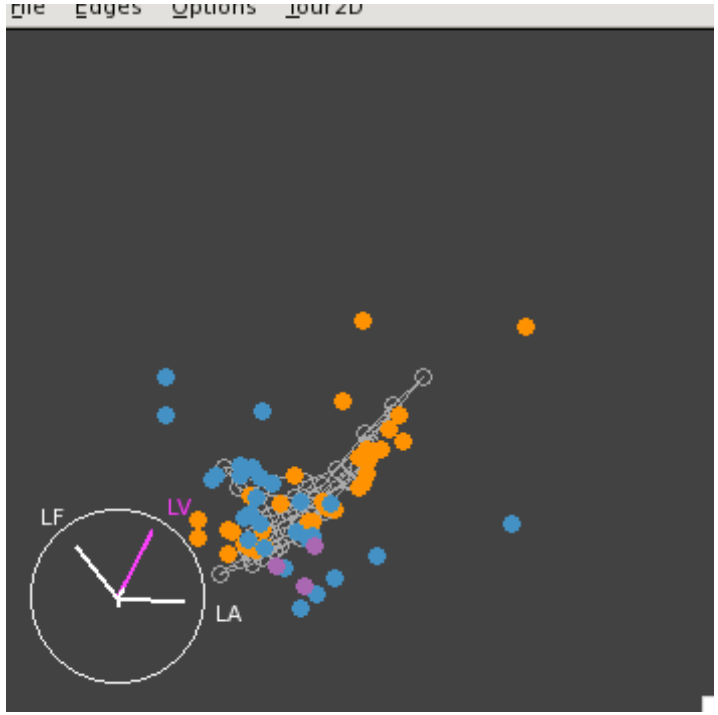
- MDS
- Circle the clusters.
- Match the species to the clusters.

1
2
3
4
5

Self-organizing maps

A self-organizing map is a constrained k-means algorithm. A 1D or 2D net is stretched through the data. The knots in the net form the cluster means, and points closest to the knot are considered to belong to that cluster. The "map" provides a low-dimensional view of the clusters, alternate to MDS or PCA.





The net fits neatly into the data. Some of the extreme points are far from the model, which is not obvious from the map view.

Model-based clustering

Model-based clustering (Fraley and Raftery, 2002) fits a multivariate mixture model to the data. For example, if it is believed that clusters in the data are approximately elliptical in shape, then a multivariate normal distribution might be used. The shape of the clusters is defined by the variance-covariance matrix for each group.

Variance covariance is parametrized as:

$$\Sigma_k = \lambda_k D_k A_k D_k', \quad k = 1, \ldots, g \quad \text{(number of clusters)}$$

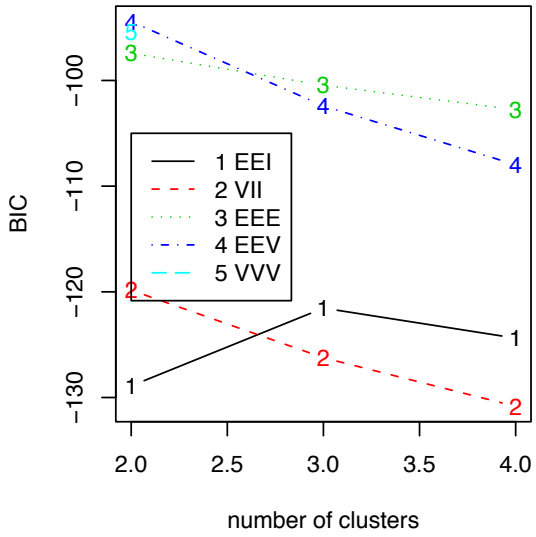
Name	Σ_k	Distribution	Volume	Shape	Orientation
EII	λI	Spherical	equal	equal	NA
VII	$\lambda_k I$	Spherical	variable	equal	NA
EEI	$\lambda D D'$	Diagonal	equal	equal	NA
VEI	$\lambda_k D D'$	Diagonal	variable	equal	NA
VVI	$\lambda_k D_k D_k'$	Diagonal	variable	variable	NA
EEE	$\lambda D A D'$	Ellipsoidal	equal	equal	equal
EEV	$\lambda D A_k D'$	Ellipsoidal	equal	equal	variable
VEV	$\lambda D_k A D_k'$	Ellipsoidal	variable	equal	variable
VVV	$\lambda_k D_k A_k D_k'$	Ellipsoidal	variable	variable	variable

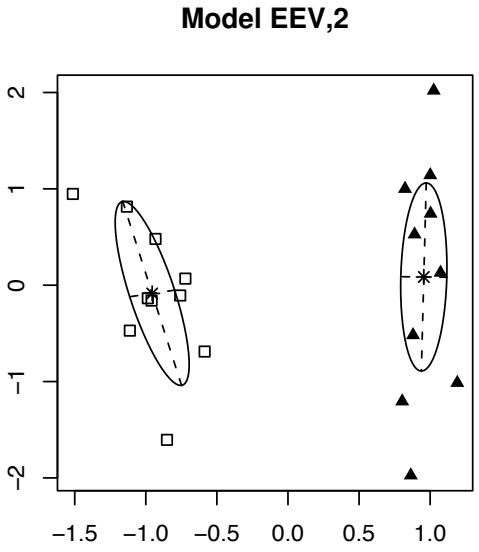
EII VII EEE EEV VVV

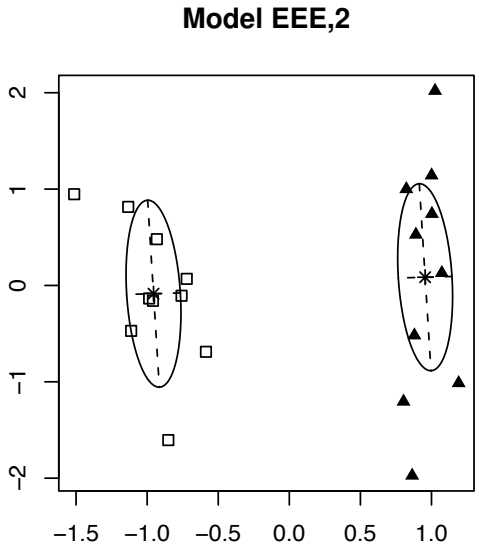
Model fitting

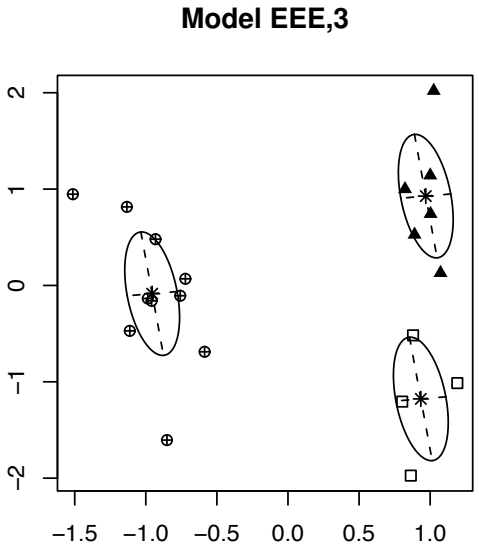
The cluster model is fit by estimating mean, variance-covariance of each population and the mixing proportion, and optimizing these for the sample.

The fit is evaluated by examining the sample variation in relation to the parameter estimates, using Bayes Information Criterion (BIC). The higher the BIC value the better the model.



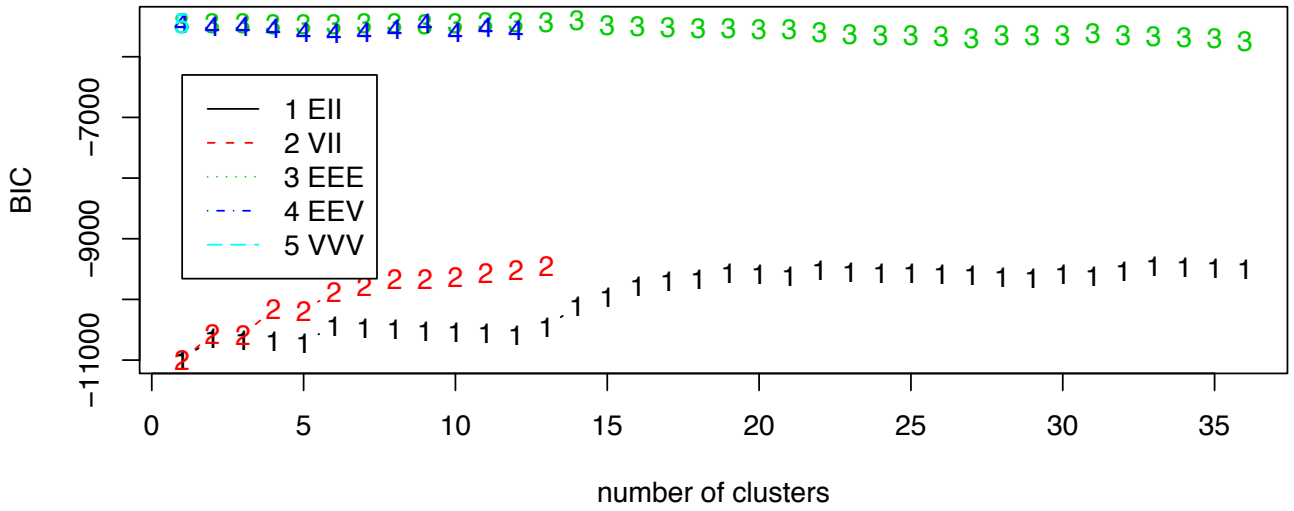




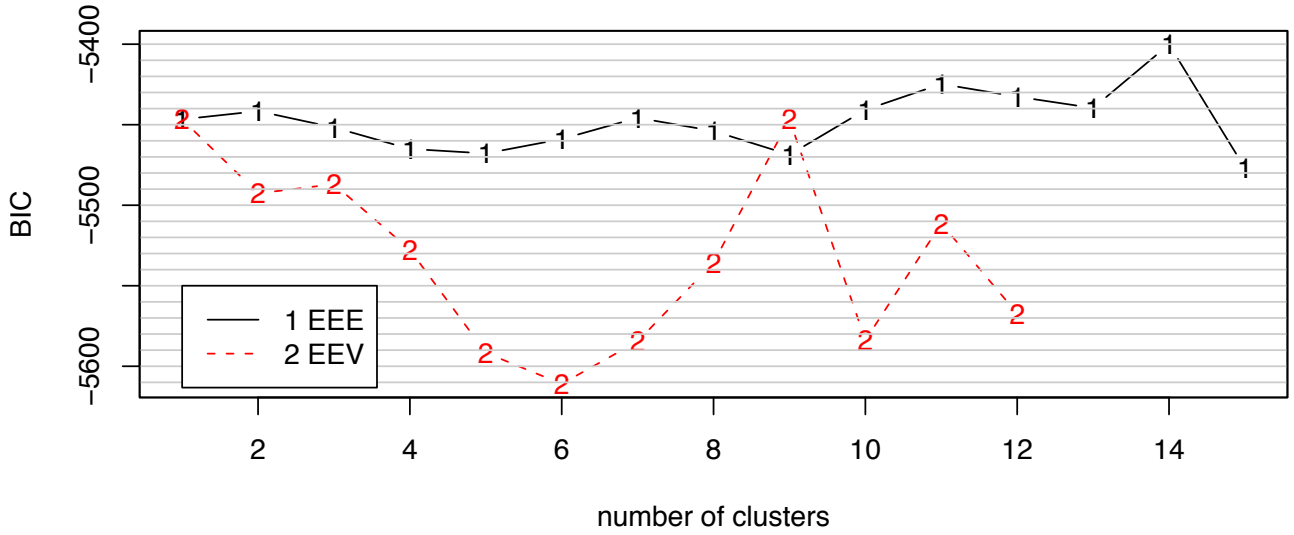


Real model is EEE-2.
Best fit as measured by BIC is EEV-2.
How many parameters need to be estimated?

Example: Music



Elliptical models are much better than spherical models.



best model:
EEE 14
but other models worth exploring too.

Cluster	Cluster Means					Names of tracks
	LVar	LAve	LMax	LFEner	LFreq	
11	1.3×10^8	50.2	3.3×10^4	114	41	Saturday Morning
12	8.3×10^7	-2.8	3.1×10^4	112	246	Girl, Cant Buy Me Love
10	4.2×10^7	-4.3	3.2×10^4	108	108	All in a Days Work, Love of the Loveless, Wrong About Bobby, Yellow Submarine
13	5.9×10^7	-3.5	3.1×10^4	111	160	Rock Hard Times, Lone Wolf, I Want to Hold Your Hand, I Feel Fine, Ticket to Ride, Help, Penny Lane
14	2.4×10^7	-20.1	2.9×10^4	107	232	The Good Old Days, Love Me Do, Yesterday, B4, Waterloo
1	1.6×10^7	-24.2	2.7×10^4	106	169	Dancing Queen, Agony, Eleanor Rigby, B8, Anywhere Is
2	1.4×10^7	216.2	3.0×10^4	105	198	V6
4	7.2×10^6	-83.0	2.7×10^4	102	92	Knowing Me, Take a Chance, Mamma Mia, Lay All You, Super Trouper, Money
3	6.5×10^6	17.5	2.3×10^4	102	274	V1, V3, V9, M3, M6, Restraining, B1, B3, B7, V13
5	4.7×10^6	14.2	1.8×10^4	86	209	V5, HeyJude
6	3.1×10^6	1.6	2.1×10^4	98	552	V7, M4, B5
9	2.4×10^6	-21.3	1.3×10^4	103	233	I Have A Dream, SOS, M1, M2, M5, The Memory of Trees, Pax Deorum, V11
7	5.3×10^5	7.7	8.1×10^3	99	566	V2, V10, B6
8	5.8×10^5	-9.2	5.7×10^3	105	222	The Winner, V4, V8, B2, V12

EEE model 14 clusters

Comparing results

- One approach is to use a confusion table. For example, to compare the results for hierarchical using average linkage and k-means clustering on the flea beatles data we could summarize the clusters like:

	k-means		
hc(ave)	1	2	3
1	0	0	19
2	24	0	0
3	0	30	0
4	0	1	0

Mapping is:

$1 \rightarrow 3$

$2 \rightarrow 1$

$3 \rightarrow 2$

Rearrange confusion table accordingly:

	k-means		
hc(ave)	1	2	3
2	24	0	0
3	0	30	0
1	0	0	19
4	0	1	0