4.3. Cluster analysis

***K*-means method by MacQueen:**

1. Partition the items into K initial clusters.

2. Assign an item to the cluster whose centroid (mean) is nearest using some measure such as Euclidean distance. Recalculate the centroid for the cluster receiving the new item and for the cluster losing the item.

3. Repeat Step 2 until no more reassignments take place.

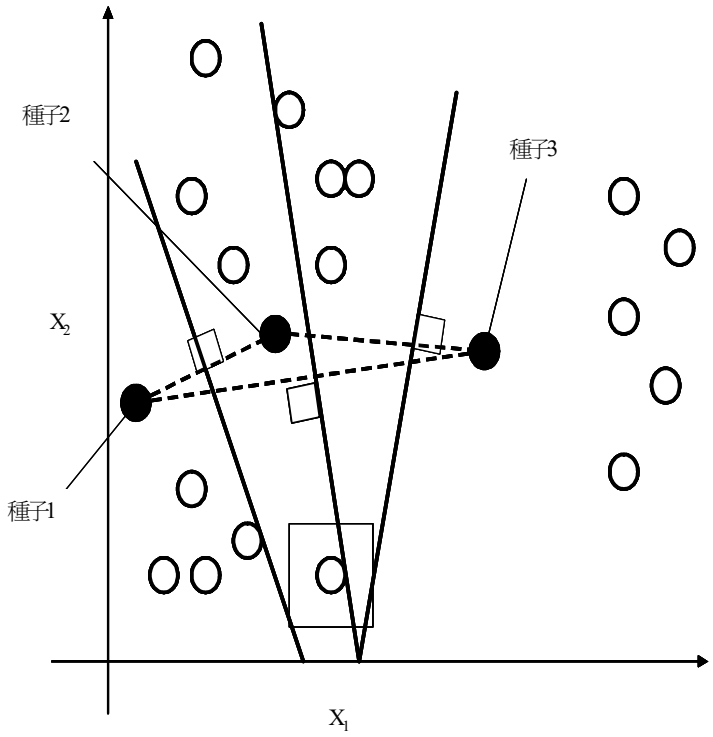
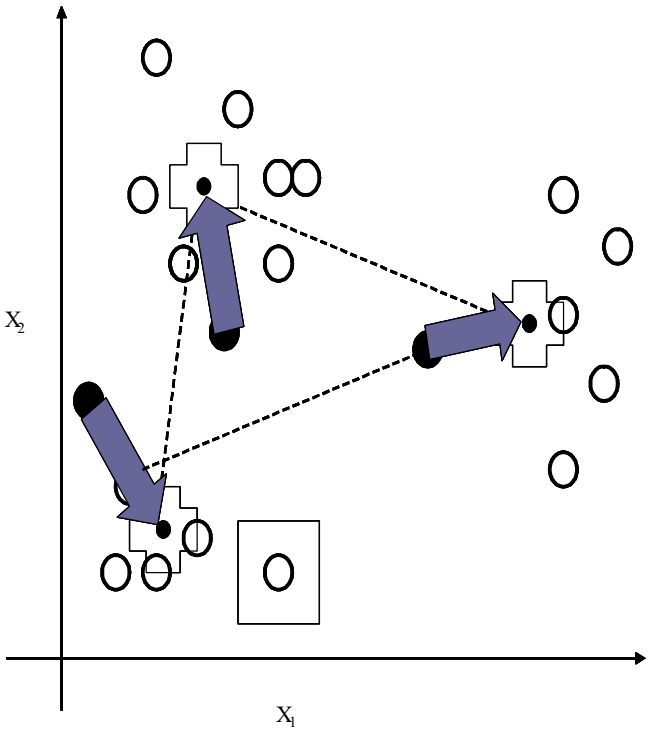


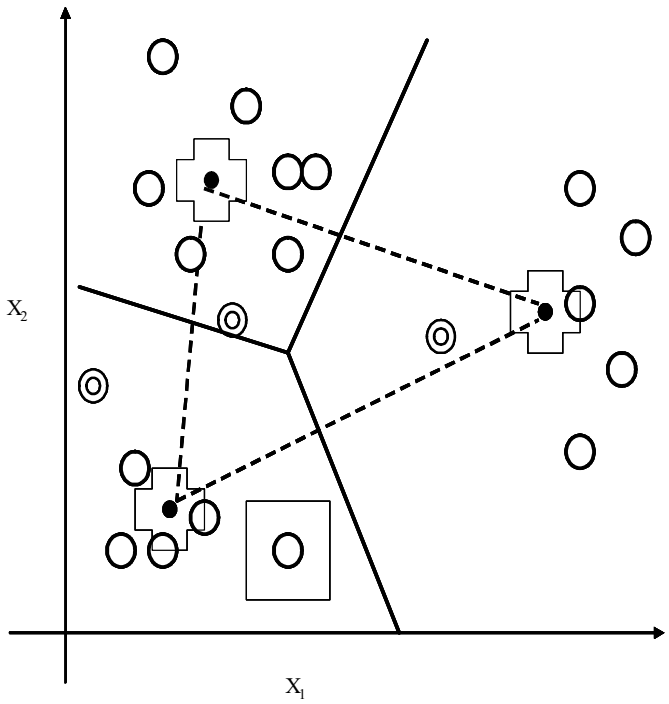
圖10.3 初始種子決定了初始的群集邊界

起初的三個種子由虛線連起來，而實線代表這三個種子所購成群集的邊界。注意到一個被方形圍起來的那一個點。



新的質心由一個十字形圖樣來標示，箭頭代表原本的種子從原來的位置移動到新位置的情況。

圖10.4 計算新群集的質心



顯示出新群集的邊界，這個界線是由與兩個質心距離相等的點所構成。注意到一個被方形圍起來的那一個點，它原本屬於第二個群集，現在被重新分配到第一個群集。

圖10.5 每一次重複的過程中，所有群集分配都必須重新計算一次

Example:

Suppose we measures two variables X_1 and X_2 for four items $\mathrm{A}, \mathrm{B}, \mathrm{C}$, and D. The data are as follows:

Observations		
Item	X_1	X_2
A	5	4
B	1	-2
C	-1	1
D	3	1

Use the K-means clustering technique to divide the items into $K = 2$ clusters. Start with the initial groups (AB) and (CD).

[solution:]

The initial centroids are

$$\bar{x}_1 = \begin{bmatrix} \frac{5+1}{2} \\ \frac{4-2}{2} \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

and

$$\bar{x}_2 = \begin{bmatrix} \frac{-1+3}{2} \\ \frac{1+1}{2} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Since

$$d^2(A, \bar{x}_1) = 13, d^2(A, \bar{x}_2) = 25 \Longrightarrow not\ reassign\ A$$
$$d^2(B, \bar{x}_1) = 13, d^2(B, \bar{x}_2) = 9 \Longrightarrow reassign\ B\ to\ (CD)$$
$$d^2(C, \bar{x}_1) = 16, d^2(C, \bar{x}_2) = 4 \Longrightarrow not\ reassign\ C$$
$$d^2(D, \bar{x}_1) = 0, d^2(D, \bar{x}_2) = 4 \Longrightarrow reassign\ D\ to\ (AB)$$

new clusters are (AD) and (BC). The new centroids are

$$\bar{x}_1 = \begin{bmatrix} \frac{5+3}{2} \\ \frac{4+1}{2} \end{bmatrix} = \begin{bmatrix} 4 \\ 2.5 \end{bmatrix}$$

and

$$\bar{x}_2 = \begin{bmatrix} \frac{-1+1}{2} \\ \frac{-2+1}{2} \end{bmatrix} = \begin{bmatrix} 0 \\ -0.5 \end{bmatrix}.$$

Since

$$d^2(A,\bar{x}_1) = 3.25, d^2(A,\bar{x}_2) = 45.25 \Rightarrow not\ \ reassign\ \ A$$
$$d^2(D,\bar{x}_1) = 3.25, d^2(D,\bar{x}_2) = 11.25 \Rightarrow not\ \ reassign\ \ D$$
$$d^2(B,\bar{x}_1) = 29.25, d^2(B,\bar{x}_2) = 3.25 \Rightarrow not\ \ reassign\ \ B$$
$$d^2(C,\bar{x}_1) = 27.25, d^2(C,\bar{x}_2) = 3.25 \Rightarrow not\ \ reassign\ \ C,$$

the clusters remain the same and the centroids have not changed. Therefore, the final clusters are (AD) and (BC).

Example (Splus):

```
ruspini
plot(ruspini$x,ruspini$y)
kmeans(ruspini,4)
```