A Handbook of Statistical Analyses
Using R

Brian S. Everitt and Torsten Hothorn
Chapter 15

Cluster Analysis: Classifying the Exoplanets

15.1 Introduction

15.2 Cluster Analysis

15.3 Analysis Using R

Sadly Figure 15.2 gives no completely convincing verdict on the number of groups we should consider, but using a little imagination ‘little elbows’ can be spotted at the three and five group solutions. We can find the number of planets in each group using

```R
R> planet_kmeans3 <- kmeans(planet.dat, centers = 3)
R> table(planet_kmeans3$cluster)
1 2 3
28 10 63
```

The centers of the clusters for the untransformed data can be computed using a small convenience function

```R
R> ccent <- function(cl) {
+ f <- function(i) colMeans(planets[cl == i, ])
+ x <- sapply(sort(unique(cl)), f)
+ colnames(x) <- sort(unique(cl))
+ return(x)
+ }
```

which, applied to the three cluster solution obtained by k-means gets

```R
R> ccent(planet_kmeans3$cluster)

  mass     period    eccen
1  7.0532143 839.1644356 0.5184643
2  3.4360000 2420.5500000 0.2718000
3  1.6540635  311.3897180 0.1777984
```

for the three cluster solution and, for the five cluster solution using

```R
R> planet_kmeans5 <- kmeans(planet.dat, centers = 5)
R> table(planet_kmeans5$cluster)
1 2 3 4 5
28 5 7 49 12
R> ccent(planet_kmeans5$cluster)
```

```R

  mass     period    eccen
1  7.0532143 839.1644356 0.5184643
2  3.4360000 2420.5500000 0.2718000
3  1.6540635  311.3897180 0.1777984
4  0.5184643  311.3897180 0.1777984
5  0.2718000  311.3897180 0.1777984
```
R> data("planets", package = "HSAUR")
R> library("scatterplot3d")
R> scatterplot3d(log(planets$mass), log(planets$period),
+    log(planets$eccen), type = "h", angle = 55,
+    scale.y = 0.7, pch = 16, y.ticklabs = seq(0,
+    10, by = 2), y.margin.add = 0.1)

**Figure 15.1** 3D scatterplot of the logarithms of the three variables available for each of the exoplanets.

<table>
<thead>
<tr>
<th></th>
<th>mass</th>
<th>period</th>
<th>eccen</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.2617857</td>
<td>580.6829</td>
<td>0.4910714</td>
</tr>
<tr>
<td>2</td>
<td>14.3480</td>
<td>659.3976</td>
<td>0.3268</td>
</tr>
<tr>
<td>3</td>
<td>2.185714</td>
<td>2557.643</td>
<td>0.19900</td>
</tr>
<tr>
<td>4</td>
<td>1.6846122</td>
<td>282.2686</td>
<td>0.1221082</td>
</tr>
<tr>
<td>5</td>
<td>8.595</td>
<td>1335.74</td>
<td>0.473</td>
</tr>
</tbody>
</table>
R> rge <- apply(planets, 2, max) - apply(planets, 2, min)
R> planet.dat <- sweep(planets, 2, rge, FUN = "/")
R> n <- nrow(planet.dat)
R> wss <- rep(0, 10)
R> wss[1] <- (n - 1) * sum(apply(planet.dat, 2, var))
R> for (i in 2:10) wss[i] <- sum(kmeans(planet.dat, + centers = i)$withinss)
R> plot(1:10, wss, type = "b", xlab = "Number of groups", + ylab = "Within groups sum of squares")

Figure 15.2  Within-cluster sum of squares for different numbers of clusters for the exoplanet data.
We now proceed to apply model-based clustering to the planets data. R functions for model-based clustering are available in package mclust (Fraley et al., 2005; Fraley and Raftery, 2002). Here we use the Mclust function since this selects both the most appropriate model for the data and the optimal number of groups based on the values of the BIC computed over several models and a range of values for number of groups. The necessary code is:

```r
R> library("mclust")
R> planet_mclust <- Mclust(planet.dat)
```

and we first examine a plot of BIC values using

```r
R> plot(planet_mclust, planet.dat)
```

and selecting the BIC option (option number 1 to be selected interactively). The resulting diagram is shown in Figure 15.3. In this diagram the numbers refer to different model assumptions about the shape of clusters:

1. Spherical, equal volume,
2. Spherical, unequal volume,
3. Diagonal equal volume, equal shape,
4. Diagonal varying volume, varying shape,
5. Ellipsoidal, equal volume, shape and orientation,
6. Ellipsoidal, varying volume, shape and orientation.

The BIC selects model 4 (diagonal varying volume and varying shape) with three clusters as the best solution as can be seen from the print output:

```r
R> print(planet_mclust)
best model: diagonal, varying volume and shape with 3 groups
average/median classification uncertainty: 0.043 / 0.012
```

This solution can be shown graphically as a scatterplot matrix. The plot is shown in Figure 15.4. Figure 15.5 depicts the clustering solution in the three-dimensional space. The number of planets in each cluster and the mean vectors of the three clusters for the untransformed data can now be inspected by using

```r
R> table(planet_mclust$classification)
1 2 3
19 41 41
R> ccent(planet_mclust$classification)
        1         2         3
mass 1.1665263 1.5797561 6.0761463
period 6.4718016 313.4127073 1325.5310048
eccen 0.0365263 0.3061463 0.3704951
```
Cluster 1 consists of planets about the same size as Jupiter with very short periods and eccentricities (similar to the first cluster of the $k$-means solution). Cluster 2 consists of slightly larger planets with moderate periods and large eccentricities, and cluster 3 contains the very large planets with very large periods. These two clusters do not match those found by the $k$-means approach.
R> x <- clPairs(planet.dat, classification = planet_mclust$classification, + symbols = 1:3, col = "black")

Figure 15.4 Scatterplot matrix of planets data showing a three cluster solution from Mclust.
ANALYSIS USING R

R> scatterplot3d(log(planets$mass), log(planets$period),
+     log(planets$eccen), type = "h", angle = 55,
+     scale.y = 0.7, pch = planet_mclust$classification,
+     y.ticklabs = seq(0, 10, by = 2), y.margin.add = 0.1)

Figure 15.5 3D scatterplot of planets data showing a three cluster solution from Mclust.
Bibliography
