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> planet_kmeans3 <- kmeans(planet.dat, centers = 3)
R> table(planet_kmeans3$cluster)
```

```
 1  2  3
34 14 53
```

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

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

```
                 1          2           3
mass     2.9276471   10.56786   1.6710566
period 616.0760882 1693.17201 427.7105892
eccen    0.4953529    0.36650   0.1219491
```

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

```
R> planet_kmeans5 <- kmeans(planet.dat, centers = 5)
R> table(planet_kmeans5$cluster)
```

```
 1  2  3  4  5
30 35 18  4 14
```

```
R> ccent(planet_kmeans5$cluster)
```

```
                1           2           3        4
mass     1.743533   1.7448571   3.4916667    2.115
```

```
R> data("planets", package = "HSAUR")
R> library("scatterplot3d")
R> scatterplot3d(log(planets$mass), log(planets$period),
+      log(planets$eccen), type = "h", angle = 55,
+      pch = 16, y.ticklabs = seq(0, 10, by = 2),
+      y.margin.add = 0.1, scale.y = 0.7)
```

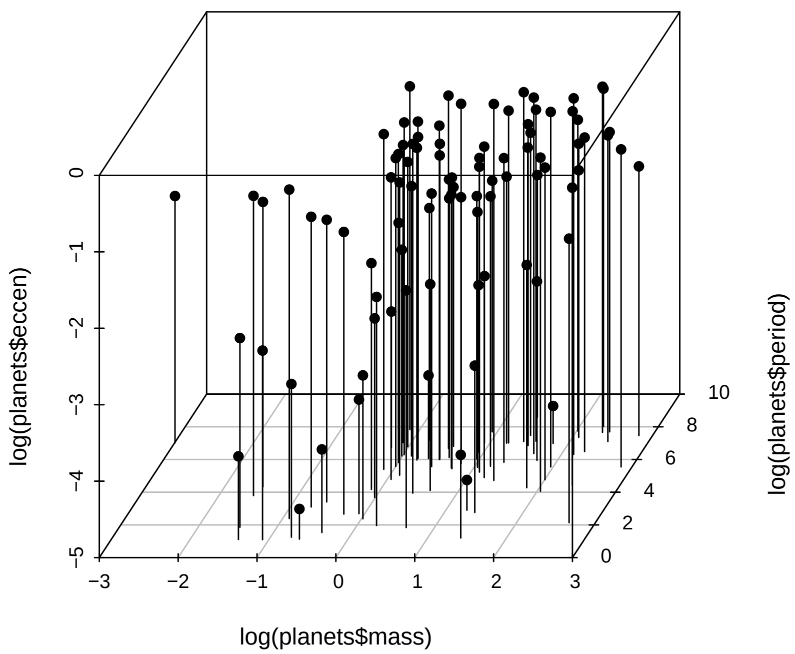

Figure 15.1 3D scatterplot of the logarithms of the three variables available for each of the exoplanets.

```
period 176.297374 552.3494286 638.0220556 3188.250
eccen    0.049310   0.2939143   0.6032778    0.110
                5
mass     10.8121429
period 1318.6505856
eccen     0.3836429
```

```
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")
```

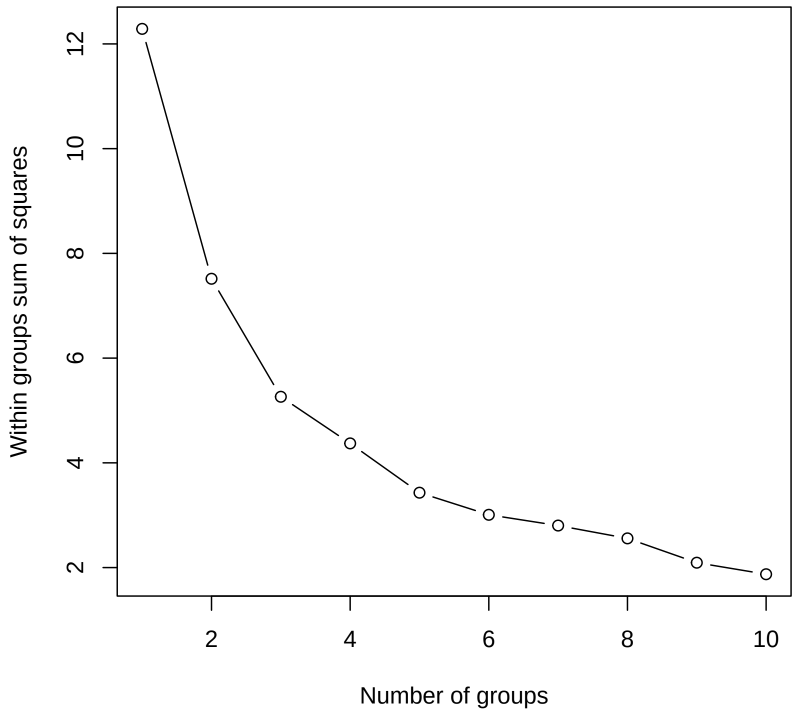

Figure 15.2 Within-cluster sum of squares for different numbers of clusters for the exoplanet data.

```
R> plot(planet_mclust, planet.dat, what = "BIC", col = "black",
+       ylab = "-BIC", ylim = c(0, 350))
```

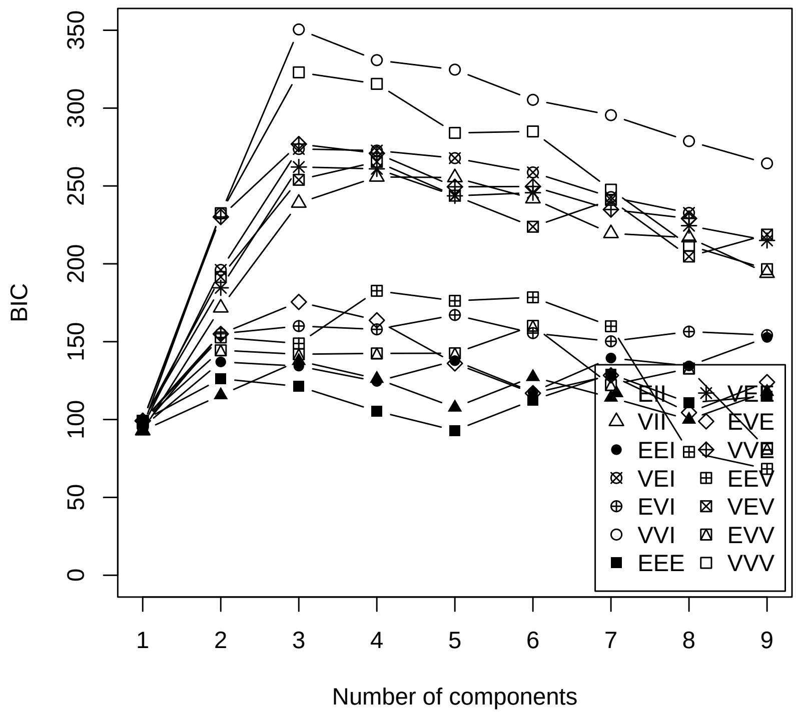

Figure 15.3 Plot of BIC values for a variety of models and a range of number of clusters.

15.3.1 Model-based Clustering in R

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., 2006, 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> library("mclust")
R> planet_mclust <- Mclust(planet.dat)
```

and we first examine a plot of BIC values using 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> print(planet_mclust)
'Mclust' model object: (VVI,3)

Available components:
 [1] "call"           "data"           "modelName"
 [4] "n"              "d"              "G"
 [7] "BIC"            "loglik"         "df"
[10] "bic"            "icl"            "hypvol"
[13] "parameters"     "z"              "classification"
[16] "uncertainty"
```

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> table(planet_mclust$classification)
 1  2  3
14 44 43
```

```
R> ccent(planet_mclust$classification)
                1           2            3
mass   0.52085714   1.6759545    5.9307442
period 5.16207357 272.5978000 1284.9554465
eccen  0.02385714   0.2884545    0.3583791
```

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> clPairs(planet.dat,
+       classification = planet_mclust$classification,
+       symbols = 1:3, col = "black")
```

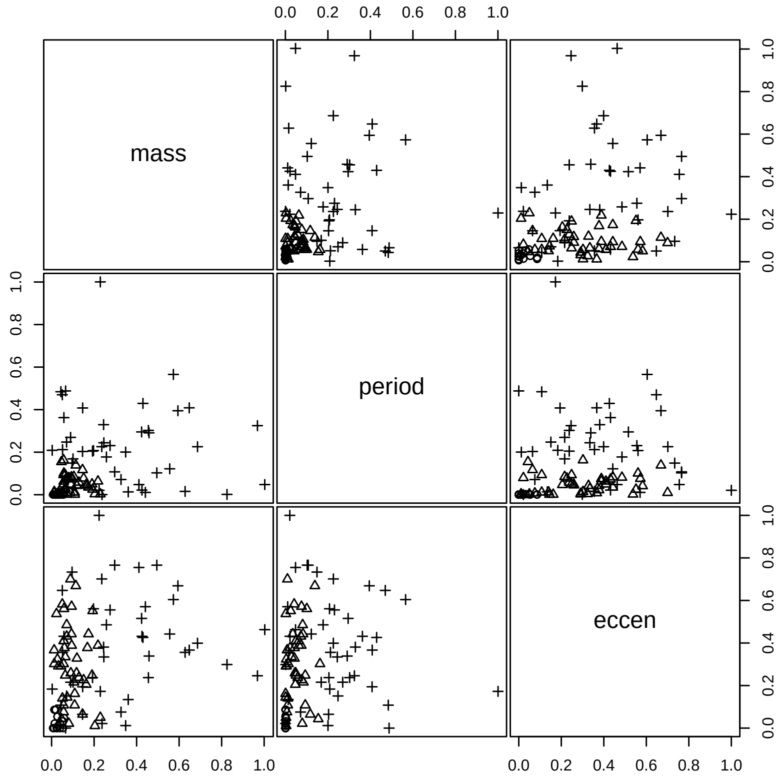

Figure 15.4 Scatterplot matrix of planets data showing a three cluster solution from Mclust.

```
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)
```

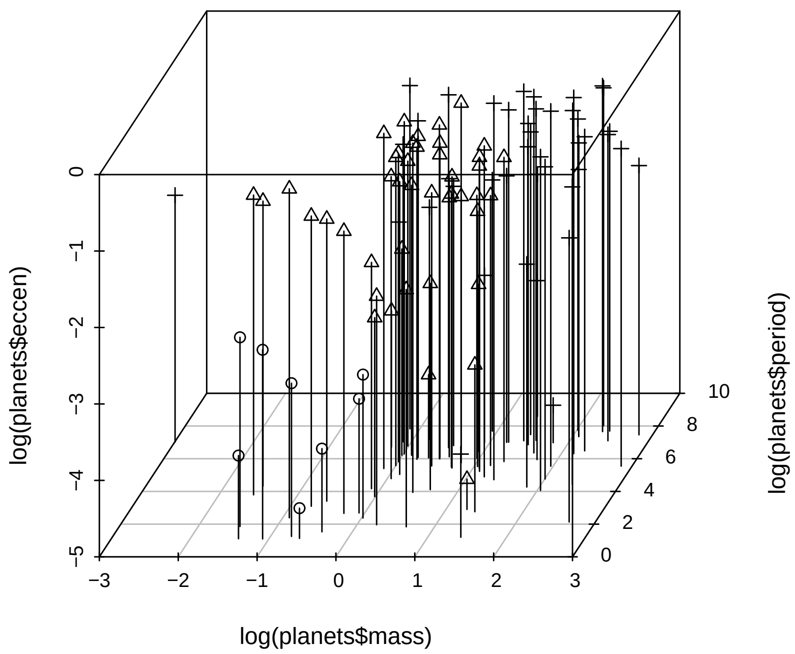

Figure 15.5 3D scatterplot of planets data showing a three cluster solution from `Mclust`.

Bibliography

Fraley, C. and Raftery, A. E. (2002), "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, 97, 611–631.

Fraley, C., Raftery, A. E., and Wehrens, R. (2006), *mclust: Model-based Cluster Analysis*, URL http://www.stat.washington.edu/mclust, R package version 3.1-1.