

Hierarchical Modal Clustering

Surajit Ray and Yansong Cheng

October 8, 2010

Model-based clustering has been widely used in many applications. A approach, named Mode Association Clustering(MAC), which is based on mode identification, is developed by applying new optimization techniques to a non-parametric estimator, kernel density estimator. The kernel density estimate becomes smoother with the increase of smoothing parameters and more points tend to converge to the same mode. This suggests a hierachical clustering approach by choosing a serial of increasing smoothing parameters, namely Hierarchical MAC(HMAC).

For example, consider the data in *disc2d.dat*.

```
> library(Modalclust)
> data(disc2d)
> disc2d.hmac = phmac(disc2d, npart = 1, parallel = F)
```

Parallel computing by using multiple processors is considered to increases the computing speed. not the optimal option,

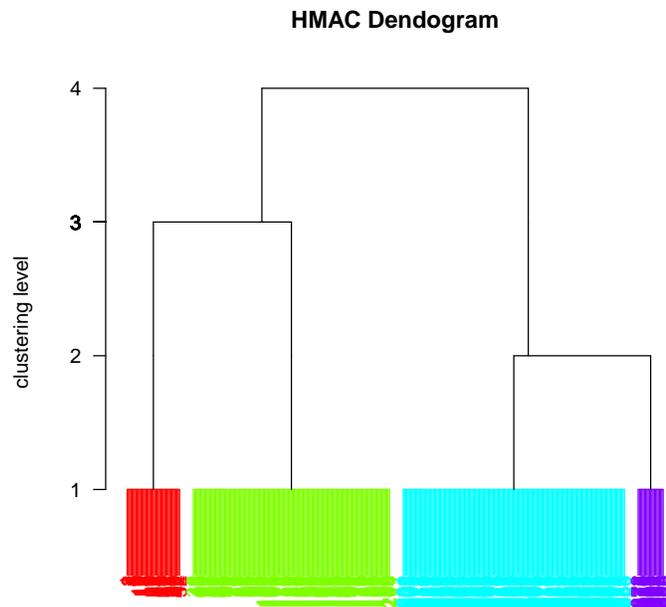
```
> disc2d.phmac = phmac(disc2d, npart = 4, parallel = T)
```

The output may be slightly different from the one without parallel computing. Although it is sub-optimal way, it increases the computing speed especially for large data set.

Consider the output from no parallel computing, which is *disc2d.hmac*, for the rest plotting examples. To get the cluster dendrogram

```
> plot.hmac(disc2d.hmac)
```

```
level 1 2 3
```

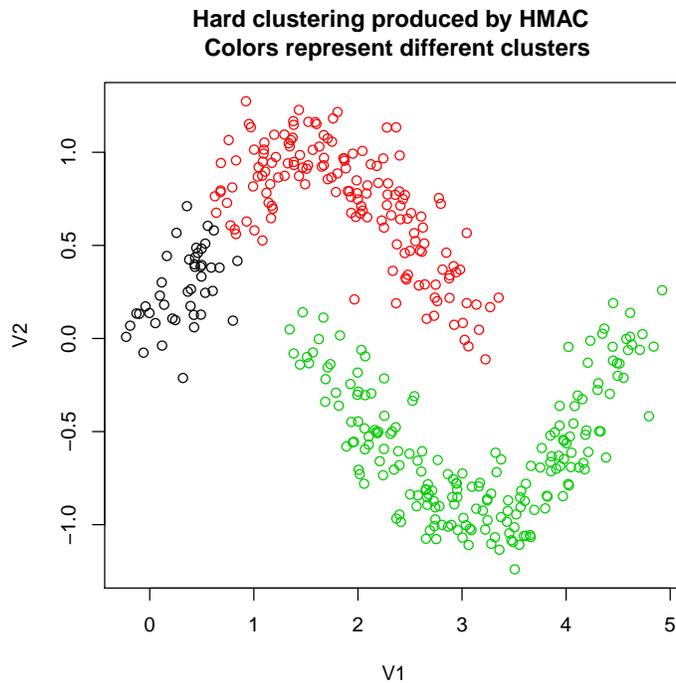


To get the hard and soft cluster plot. Either the number of clusters or the clustering level needs to be specified.

```
> hard.hmac(disc2d.hmac, n.cluster = 3)
```

The level at which there are 3 clusters is 2

```
> soft.hmac(disc2d.hmac, level = 2)
```



With no plot, *hard.hmac* will return the membership of each observation while *soft.hmac* will return the posterior probability of each point and the boundary point as well.

```
> member = hard.hmac(disc2d.hmac, n.cluster = 3, plot = F)
```

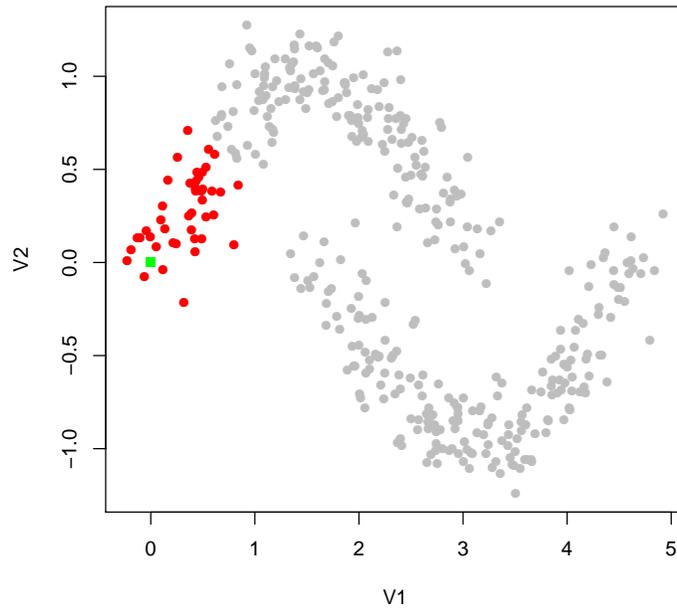
The level at which there are 3 clusters is 2

```
> post.prob = soft.hmac(disc2d.hmac, level = 2, plot = F)
```

For the convenience of application, users can choose clusters by clicking the point on the graph. The clicked point will be lighted up along with other points within the same cluster.

```
> choose.cluster(disc2d.hmac, n.cluster = 3, x = c(0, 0))
```

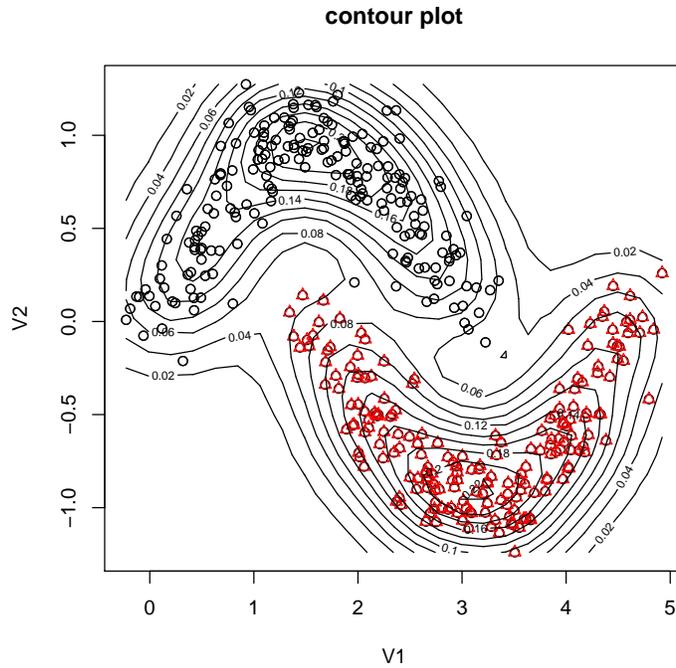
The level at which there are 3 clusters is 2



In addition, contour plot is also available in this package.

```
> contour.hmac(disc2d.hmac, n.cluster = 2)
```

The level at which there are 2 clusters is 3



References

1. Li, J., Ray, S., Lindsay, B. G., "A nonparametric statistical approach to clustering via mode identification," *Journal of Machine Learning Research*, 8(8):1687-1723, 2007.
2. Lindsay, B.G., Markatou M., Ray, S., Yang, K., Chen, S.C. "Quadratic distances on probabilities: the foundations," *The Annals of Statistics* Vol. 36, No. 2, page 983–1006, 2008.