A toolbox for fuzzy clustering using the R programming language

Maria Brigida Ferraro and Paolo Giordani

Department of Statistical Sciences

SAPIENZA
UNIVERSITÀ DI ROMA

Workshop on **Clustering methods and their applications**

November 28, 2014 - Free University of Bozen-Bolzano, Italy

Motivations

The most popular algorithm is the fuzzy *k*-means (f*k*m) (Bezdek, 1974):

objects assigned to clusters
according to membership degrees in [0,1]

Starting from f*k*m, fuzzy clustering has received an increasing attention by researchers from several fields

Nonetheless, popular commercial software solutions (`SAS`, `SPSS`, ...) do not contain routines for fuzzy clustering. Just a few exceptions (limited to f*k*m): `MATLAB` and `R`

R package fclust, version 1.0.1

Suit of functions for fuzzy clustering analysis (algorithms and cluster validity indices)

▷ http://cran.r-project.org/web/packages/fclust/index.html

Fuzzy k-Means (FkM) (Bezdek, 1974)

$$\min_{\mathbf{U},\mathbf{H}} J_{\mathrm{F}k\mathrm{M}} = \sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}^{m} d^2\left(\mathbf{x}_i, \mathbf{h}_g\right) = \sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}^{m} \left\|\mathbf{x}_i - \mathbf{h}_g\right\|^2$$

$$\text{s.t.} \quad u_{ig} \in [0,1], \sum_{g=1}^{k} u_{ig} = 1$$

where

- $\mathbf{X} = [x_{ij}]$: data matrix of order $(n \times t)$
- $\mathbf{U} = [u_{ig}]$: membership degree matrix of order $(n \times k)$
- $\mathbf{H} = [h_{gj}]$: prototype matrix of order $(k \times t)$
- $m(>1)$: parameter of fuzziness (usually $m = 2$)

with

n: number of objects

t: number of variables

k: number of clusters

FkM with covariance matrices (Gustafson & Kessel, 1979)

$$\min_{\mathbf{U},\mathbf{H},\mathbf{F}_1 \cdots \mathbf{F}_k} J_{GK} = \sum_{i=1}^{n} \sum_{g=1}^{k} u_{ig}^m d_M^2 \left(\mathbf{x}_i, \mathbf{h}_g\right)$$

$$\text{s.t.} \quad u_{ig} \in [0,1], \sum_{g=1}^{k} u_{ig} = 1, \quad |\mathbf{F}_g| = \rho_g > 0$$

where

- $d_M^2 \left(\mathbf{x}_i, \mathbf{h}_g\right) = \left(\mathbf{x}_i - \mathbf{h}_g\right)' \mathbf{F}_g \left(\mathbf{x}_i - \mathbf{h}_g\right)$ is the Mahalanobis distance
- $\mathbf{F}_g$: symmetric and definite positive
- ρ_g: volume parameter (usually equal to 1)

Entropic F*k*M (Li & Mukaidono, 1995)

$$\min_{\mathbf{U},\mathbf{H}} J_{ent} = \sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig} d^2\left(\mathbf{x}_i, \mathbf{h}_g\right) + p \sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig} \log u_{ig}$$

$$\text{s.t.} \quad u_{ig} \in [0,1], \sum_{g=1}^{k} u_{ig} = 1$$

where

- p: is the degree of fuzzy entropy,
- p is called the "temperature" in statistical physics.

- GK variant of Entropic F*k*M (Ferraro & Giordani, 2013)

Fuzzy clustering with polynomial fuzzifier (Klawon and Höppner, 2003)

$$\min_{\mathbf{U},\mathbf{H},\mathbf{F}_1 \cdots \mathbf{F}_k} J_{FkM.pf} = \sum_{i=1}^{n} \sum_{g=1}^{k} h(u_{ig}) d^2 \left(\mathbf{x}_i, \mathbf{h}_g\right)$$

$$\text{s.t.} \quad u_{ig} \in [0,1], \sum_{g=1}^{k} u_{ig} = 1$$

where

- $h(u_{ig}) = \left(\frac{1-\beta}{1+\beta} u_{ig}^2 + \frac{2\beta}{1+\beta} u_{ig}\right)$ is the polynomial fuzzifier fuction
- $\beta \in [0,1]$
  - for $\beta = 0$ we obtain the fkm with parameter m equal to 2
  - for $\beta = 1$ the hard k-means

Fuzzy k-Medoids (Krishnapuram *et al.*, 2001)

$$\min_{\mathbf{U},\mathbf{M}} J_{\mathrm{F}k\mathrm{Med}} = \sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}^{m} d^2\left(\mathbf{x}_i, \mathbf{m}_g\right)$$

$$\text{s.t.} \quad u_{ig} \in [0,1], \sum_{g=1}^{k} u_{ig} = 1, \left\{\mathbf{m}_g, g = 1,\ldots,k\right\} \subset \left\{\mathbf{x}_i, i = 1,\ldots,n\right\}.$$

where

- $\{\mathbf{m}_g, g = 1,\ldots,k\} \subseteq \{\mathbf{x}_i, i = 1,\ldots,n\}$
  the medoids are a subset of the observed objects

The fuzzy k-medoids algorithm is usually more robust than the standard fkm algorithm

FkM with noise cluster (Davé, 1991)

$$\min_{\mathbf{U},\mathbf{H}} J_{\text{Noise}} = \sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}^{m} d^2\left(\mathbf{x}_i, \mathbf{h}_g\right) + \sum_{i=1}^{n} \delta^2 \left(1 - \sum_{g=1}^{k} u_{ig}\right)^m$$

$$\text{s.t.} \quad u_{ig} \in [0,1], \sum_{g=1}^{k+1} u_{ig} = 1.$$

where

- δ^2: squared distance of each point to the noise cluster

A partition with $k+1$ clusters is obtained when minimizing J_{Noise}. The first k standard clusters are homogeneous, whereas the noise cluster contains all the outliers and is usually not formed by objects with homogeneous features.

Fuzzy cluster validity indices (i)

- Partition coefficient

$$PC(k) = \sum_{i=1}^{n} \sum_{g=1}^{k} \frac{(u_{ig})^2}{n}$$

- Partition entropy

$$PE(k) = -\sum_{i=1}^{n} \sum_{g=1}^{k} \frac{u_{ig} \log(u_{ig})}{n}$$

- Xie & Beni index

$$XB(k) = \frac{\sum_{i=1}^{n} \sum_{g=1}^{k} u_{ig}^{m} d^2 \left(\mathbf{x}_i, \mathbf{h}_g\right)}{n \min_{g,g'(g \neq g')} d^2(\mathbf{h}_g, \mathbf{h}_{g'})}$$

Fuzzy cluster validity indices (ii)

✎ Fuzzy Silhouette

$$FS(k) = \frac{\sum_{i=1}^{n} (u_{ig} - u_{ig'})^{\alpha} s_i(k)}{\sum_{i=1}^{n} (u_{ig} - u_{ig'})^{\alpha}}$$

where

- $s_i(k) = \frac{b_i - a_i}{\max(b_i, a_i)}$: silhouette index for object i
  - a_i average dissimilarity between the object involved and all the objects belonging to the same cluster
  - b_i lowest average dissimilarity of i to any other cluster which i is not a member
- u_{ig}, $u_{ig'}$: first and second largest elements of the i-th row of $\mathbf{U}$
- α: weighting coefficient (usually $\alpha = 1$)

Visualization techniques (i)

Remark

The recalled cluster validity indices are used to evaluate the clustering results. Nevertheless, they reduce the information of a large dataset to a single value. For this reason, it is necessary to consider visualization techniques for fuzzy clustering, involving different information about the results.

VIFCR (Klawon *et al.*, 2003)

- A chart diagram of the scaled frequency related to the membership degrees

$$\frac{1}{n} \sum_{(i,g):a \leq u_{ig} < b} \left(\frac{k(k-2)}{k-1} u_{ig} + \frac{k}{k-1} \right),$$

with $a, b \in [0,1]$ and $a < b$.

Visualization techniques (ii)

- A diagram whose coordinates, for each object (point) x_i, are
  - u_{ig_1} : the highest membership degree of x_i
  - u_{ig_2} : the second highest membership degree of x_i

  All the points are included in the triangle of vertices $(0,0)$ (noise data), $(0.5,0.5)$ (ambiguous data) and $(1,0)$ (crisp assignments).
- A diagram whose coordinates, for each object (point) x_i, are

$$(d_{ig}, u_{ig})$$

  The ideal situation is to obtain high membership degrees for small distances and low membership degrees for large distances.

Visualization techniques (iii)

VAT (Bezdek & Hataway, 2002)

- The matrix of dissimilarities between the objects, $R = [r_{ij}]$, is considered.
- The matrix is reordered obtaining R^*
- Its image $I(R^*)$ is displayed.

The number of **dark blocks** along its main diagonal represents the number of **clusters** and the size of each block the approximate size of the cluster.

Visualization techniques (iv)

✎ VCV (Hathaway & Bezdek, 2003)

- First of all the clusters are ordered and the objects in each cluster are ordered by taking into account the membership degrees.
- Then, the dissimilarities r_{ij} between object x_i and x_j are taken into account.
- The following dissimilarities are used:

$$r_{ij}^* = \min_{1 \leq g \leq k} \{d_{ig} + d_{jg}\},$$

where $d_{ig} = d(x_i, h_g)$.

- Finally, the information is displayed as and intensity image $I(R^*)$.

Visualization techniques (v)

VCV2 (Huband & Bezdek, 2008)

In this case the membership degrees matrix U is reordered using the index array of R^* obtained by means of the VAT. The resulting matrix $\widehat{U}$ is transformed to the square matrix

$$U^* = \mathbf{1}_n - \left(\widehat{U}^T \widehat{U} / \max\{ \left(\widehat{U}^T \widehat{U} \right)_{ij} \} \right).$$

The display image $I(U^*)$ is compared with $I(R^*)$ to check the adequacy of the number of clusters.

The package

Package 'fclust'

July 2, 2014

Type Package

Title Fuzzy clustering

Version 1.0.1

Date 2014-03-26

Author Paolo Giordani, Maria Brigida Ferraro

Maintainer Paolo Giordani <paolo.giordani@uniroma1.it>

Description Algorithms for fuzzy clustering and cluster validity indices

Depends R(>= 2.8.1), base, graphics, stats

License GPL (>= 2)

LazyLoad yes

Repository CRAN

NeedsCompilation no

Date/Publication 2014-03-26 15:01:18

Main features of the package

- 20 functions + 4 datasets
- Most relevant functions for algorithms:
  - `FKM`: standard f*k*m algorithm
  - `FKM.gk`: Gustafson and Kessel extension of f*k*m
  - `FKM.med`: fuzzy *k*-medoids algorithm
  - `FKM.noise`: f*k*m with noise cluster
- Most relevant functions for cluster validity indices:
  - `PC`: partition coefficient
  - `PE`: partition entropy (PE);
  - `XB`: Xie and Beni index (XB)
  - `SIL.F`: fuzzy silhouette (FS)
- Interactive fuzzy clustering analysis by means of the function `Fclust`

Input arguments (for the algorithms)

- `X`: object of class `matrix` or `data.frame`
- `k`: number of clusters (default: 2)
- `m`: parameter of fuzziness (default: 2)
- `stand`: if `stand=1`, the clustering algorithm is run using standardized data (default: no standardization)
- `RS`: number of (random) starts (default: 1)
- `startU`: rational starting point for the membership degree matrix **U** (default: no rational start)
- `conv`: convergence criterion (default: 1e-9)
- `maxit`: maximum number of iterations (default: 1e+6)
- ...

Output values (for the algorithms)

Object of class `fclust`. List with the following components:

- `U`: membership degree matrix
- `H`: prototype matrix
- `clus`: matrix containing the indices of the clusters where the objects are assigned (column 1) and the associated membership degrees (column 2)
- `medoid`: vector containing the indices of the medoid objects
- `value`: vector containing the loss function values for the `RS` starts
- `cput`: vector containing the computational times (user times) for the `RS` starts
- `Xca`: data used in the clustering algorithm (standardized data if `stand=1`)
- `X`: raw data
- `call`: matched call
- ...

McDonald's data

McDonald's USA Nutrition Facts (81 menu items, no beverages)

```
> library("fclust")
> data(Mc)
```

variables:

- `numeric`:
  `Serving Size`, `Calories`, `Total Fat (g)`, `Saturated Fat (g)`, `Trans Fat (g)`, `Cholesterol (mg)`, `Sodium (mg)`, `Carbohydrates (g)`, `Dietary Fiber (g)`, `Sugars (g)`, `Protein (g)`, `Vitamin A (%DV)`, `Vitamin C (%DV)`, `Calcium (%DV)`, `Iron (%DV)`
- `factor`:
  `Type` (levels: `Burgers & Sandwiches`, `Chicken`, `Breakfast`, `Salads`, `Snacks & Sides`, `Desserts/Shakes`)

Aim of the analysis

Aim

Clustering of the menu items (scores normalized w.r.t. `Serving Size`) to discover whether a cluster structure exists (i.e. similar menu items in terms of their nutrition facts) and, in particular, whether a six-cluster structure is visible emerging a link between the variable `type` and the typology of nutrition facts.

Standard f*k*m algorithm (function `FKM`):

```
> fkm   <-   FKM(X = Mc[,1:(ncol(Mc)-1)], k = c,
                 m = 1.5, stand = 1, RS = 10)
```

Number of clusters

FS index for values of $k = 2, \ldots, 10$:
`FS` vector containing the FS values (script omitted)

```
> round(FS, 2)
  k = 2  k = 3  k = 4   k = 5  k = 6
   0.52   0.49   0.48    0.55   0.62
  k = 7  k = 8  k = 9  k = 10
   0.64   0.57   0.62    0.61
```

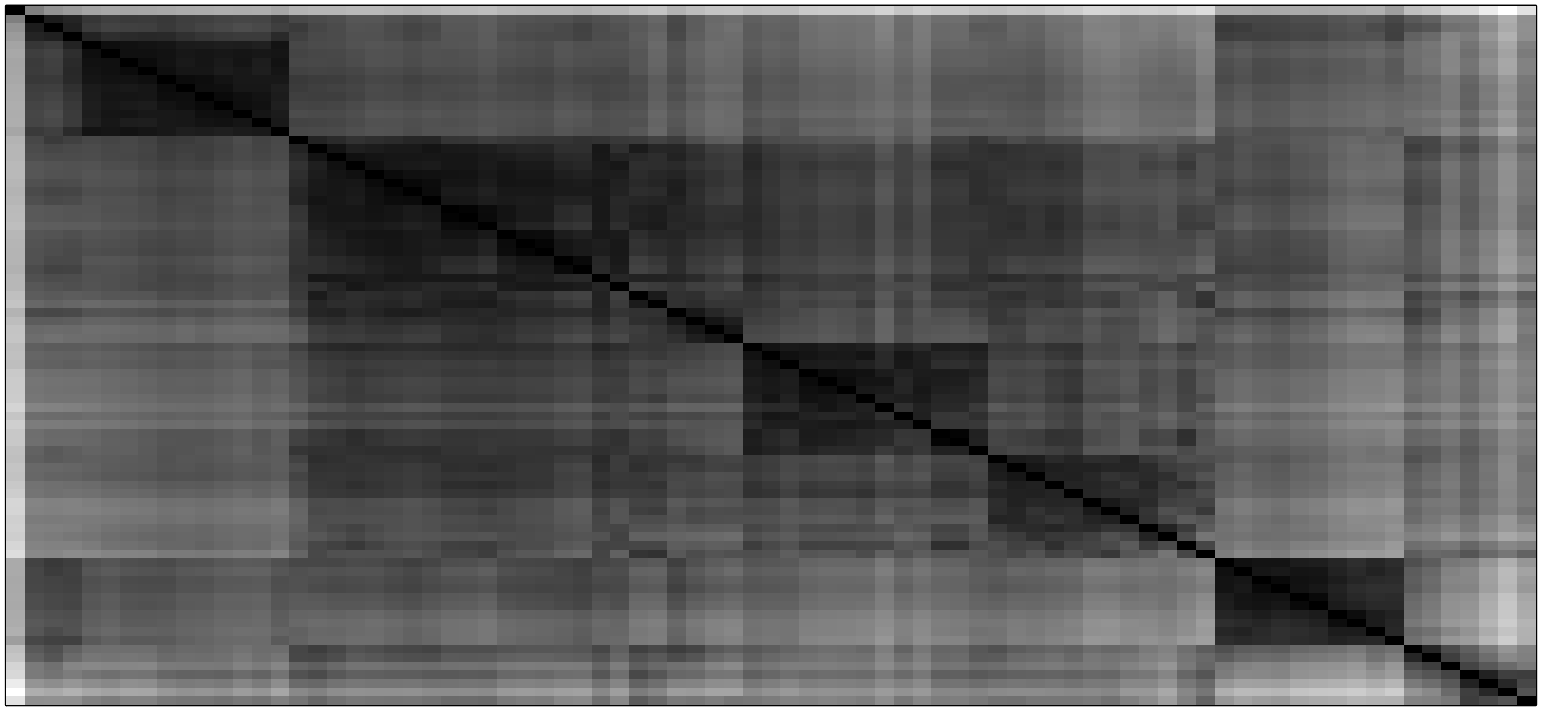
Solution with $k = 7$ clusters (two low-size clusters)

```
> fkm7   <-   FKM(X = Mc[,1:(ncol(Mc)-1)], k = 7,
              m = 1.5, stand = 1, RS = 10)
> cl.size(fkm7$U)
  Cl 1  Cl 2  Cl 3  Cl 4  Cl 5  Cl 6  Cl 7
    24    12     4    13    15    10     3
```

Data Visualization: VAT

Function `VAT(Xca)`

```
> VAT(fkm7$Xca)
```



FkM with $k = 6$ clusters

Trying to avoid low-size clusters, we move to $k = 6$ solution ($FS = 0.62$)

```
> fkm6   <-   FKM(X = Mc[,1:(ncol(Mc)-1)], k = 6,
                 m = 1.5, stand = 1, RS = 10)

> cl.size(fkm6$U)
  Cl 1  Cl 2  Cl 3  Cl 4  Cl 5  Cl 6
    12    26    10    15     5    13
```

Comparison between the solutions with $k = 6$ and $k = 7$ (Adjusted Rand Index = 0.95)

```
> table(fkm6$clus[,1], fkm7$clus[,1])
       Cl 1  Cl 2  Cl 3  Cl 4  Cl 5  Cl 6  Cl 7
Cl 1      0    12     0     0     0     0     0
Cl 2     24     0     0     0     0     0     2
Cl 3      0     0     0     0     0    10     0
Cl 4      0     0     0     0    15     0     0
Cl 5      0     0     4     0     0     0     1
Cl 6      0     0     0    13     0     0     0
```

Interpretation of the clusters (i)

```
> table(Mc$Type, fkm6$clus[,1])
```

	Cl 1	Cl 2	Cl 3	Cl 4	Cl 5	Cl 6
Breakfast	12	5	0	1	1	0
Burgers & Sandwiches	0	10	0	0	0	12
Chicken	0	4	0	0	0	0
Desserts/Shakes	0	0	0	12	4	0
Salads	0	0	10	0	0	0
Snacks & Sides	0	7	0	2	0	1

Clusters

Cluster 1 Breakfast

(Bacon, Egg & Cheese Biscuit)

Cluster 3 Salads

(Premium Southwest Salad with Grilled Chicken)

Interpretation of the clusters (ii)

Cluster 4	`Desserts/Shakes` (ice-creams and fruits)		(McFlurry with OREO Cookies)
Cluster 5	`Desserts/Shakes` (cookies and pies)		(Oatmeal Raisin Cookie)

More complex interpretation for Clusters 2 and 6

- `Burgers & Sandwiches` assigned to Cluster 6 (although no one-to-one relationship)
- Cluster 2 contains food items of different types

Nonetheless, by further inspecting the food items of type `Burgers & Sandwiches` assigned to Cluster 2 (the code is omitted), a clear interpretation of Clusters 2 and 6 can be found

Interpretation of the clusters (iii)

Findings

- Chicken-made food items belong to Cluster 2 along with two other food items with fish and pork
- All the food items assigned to Cluster 6 contain beef
- 6 (out of 7) food items of type `Snacks & Sides` assigned to Cluster 2 are chicken-based

Hence

Cluster 2	"chicken-made food items"		(Premium Crispy Chicken Ranch)
Cluster 6	"beef-made burgers and sandwiches"		(McDouble)

Centroids (i)

```
> fkm6$Hraw <- Hraw(fkm6$X, fkm6$H)
```

`Breakfast` items have highest values of `Cholesterol (mg)` and `Sodium (mg)` (a lot of items with eggs)

"chicken-made food items" presents average values for the nutrition facts except for high values of `Sodium (mg)` and lowest values of `Vitamin A (%DV)`

`Salads` are the most healthy items (lowest values of `Calories`, `Total Fat (g)`, `Saturated Fat (g)` and `Trans Fat (g)` and highest values of `Vitamin A (%DV)` and `Vitamin C (%DV)`)

Centroids (ii)

Ice-creams and fruits (`Desserts/Shakes`) present lowest values of `Cholesterol (mg)`, `Sodium (mg)`, `Dietary Fiber (g)`, `Protein (g)` and `Iron (%DV)` and highest values of `Calcium (%DV)`

Cookies and pies (`Desserts/Shakes`) are the less dietetic ones: highest amounts of `Calories`, `Total Fat (g)`, `Saturated Fat (g)`, `Carbohydrates (g)`, `Sugars (g)`. Also highest values of `Iron (%DV)` and lowest values of `Calcium (%DV)`

"beef-made burgers and sandwiches" present highest values of `Trans Fat (g)` and `Protein (g)`

Membership degrees (examples)

Oatmeal Raisin Cookie (Cluster 5 with membership degree = 0.99)

Baked Hot Apple Pie (Cluster 5 with membership degree = 0.53)

Mean values (more relevant variables)

```
> round(apply(fkm6$X[,c(1,2,3,7,9,13,14)],2,mean),2)
  Calories    Total Fat  Saturated Fat  Carbohydrates     Sugars   Iron (%DV)  Calcium (%DV)
      2.33         0.11           0.04           0.25       0.08         0.09           0.08
```

Centroid of Cluster 5 (more relevant variables)

```
> round(fkm6$Hraw[5,c(1,2,3,7,9,13,14)],2)
  Calories    Total Fat  Saturated Fat  Carbohydrates     Sugars   Iron (%DV)  Calcium (%DV)
      4.35         0.19           0.09           0.59       0.33         0.04           0.16
```

Oatmeal Raisin Cookie (more relevant variables)

```
> round(fkm6$X[''Oatmeal Raisin Cookie'',c(1,2,3,7,9,13,14)],2)
  Calories    Total Fat  Saturated Fat  Carbohydrates     Sugars   Iron (%DV)  Calcium (%DV)
      4.55         0.18           0.08           0.67       0.39         0.06           0.18
```

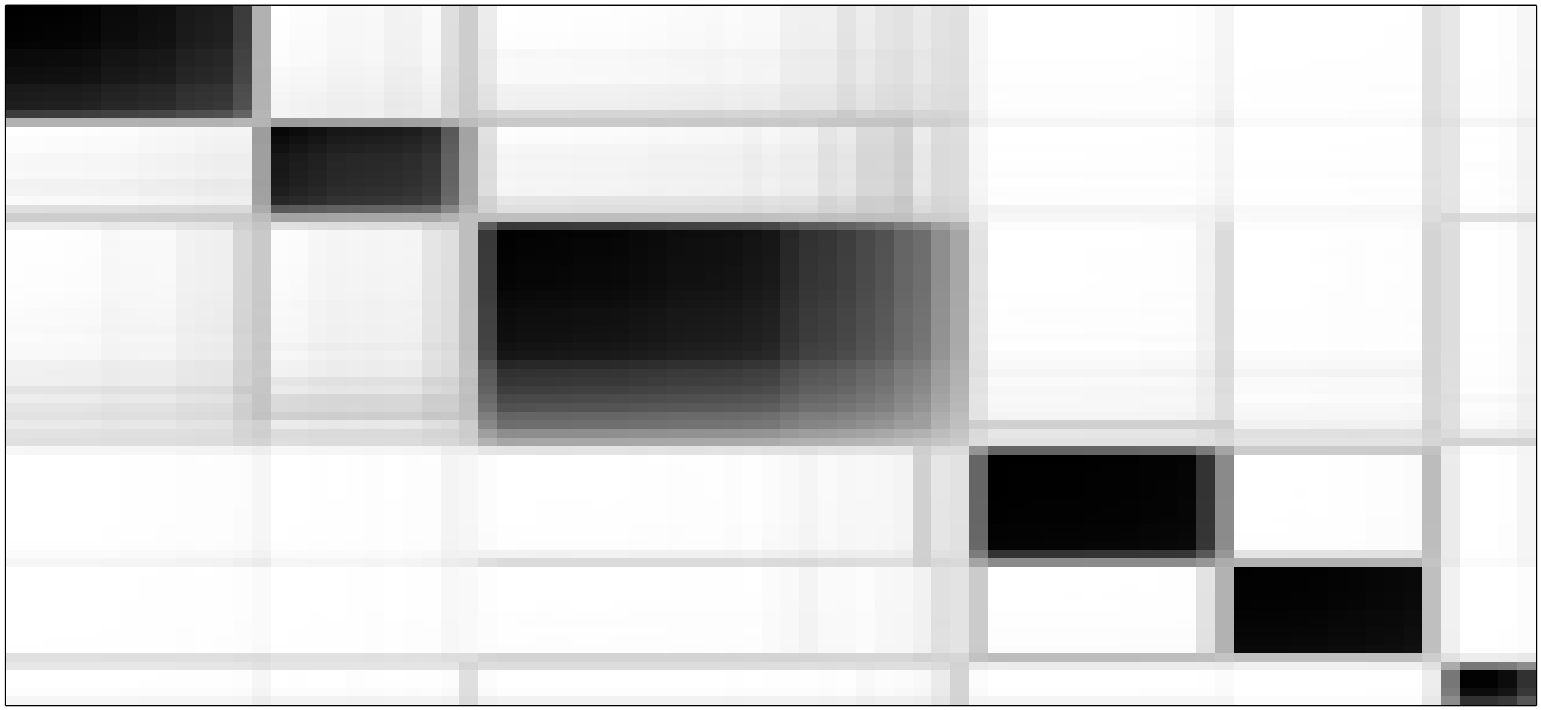
Baked Hot Apple Pie (more relevant variables)

```
> round(fkm6$X[''Baked Hot Apple Pie'',c(1,2,3,7,9,13,14)],2)
  Calories    Total Fat  Saturated Fat  Carbohydrates     Sugars   Iron (%DV)  Calcium (%DV)
      3.25         0.17           0.09           0.42       0.17         0.03           0.08
```

Results Visualization: VCV2

Function `VCV2(Xca, U, which)`

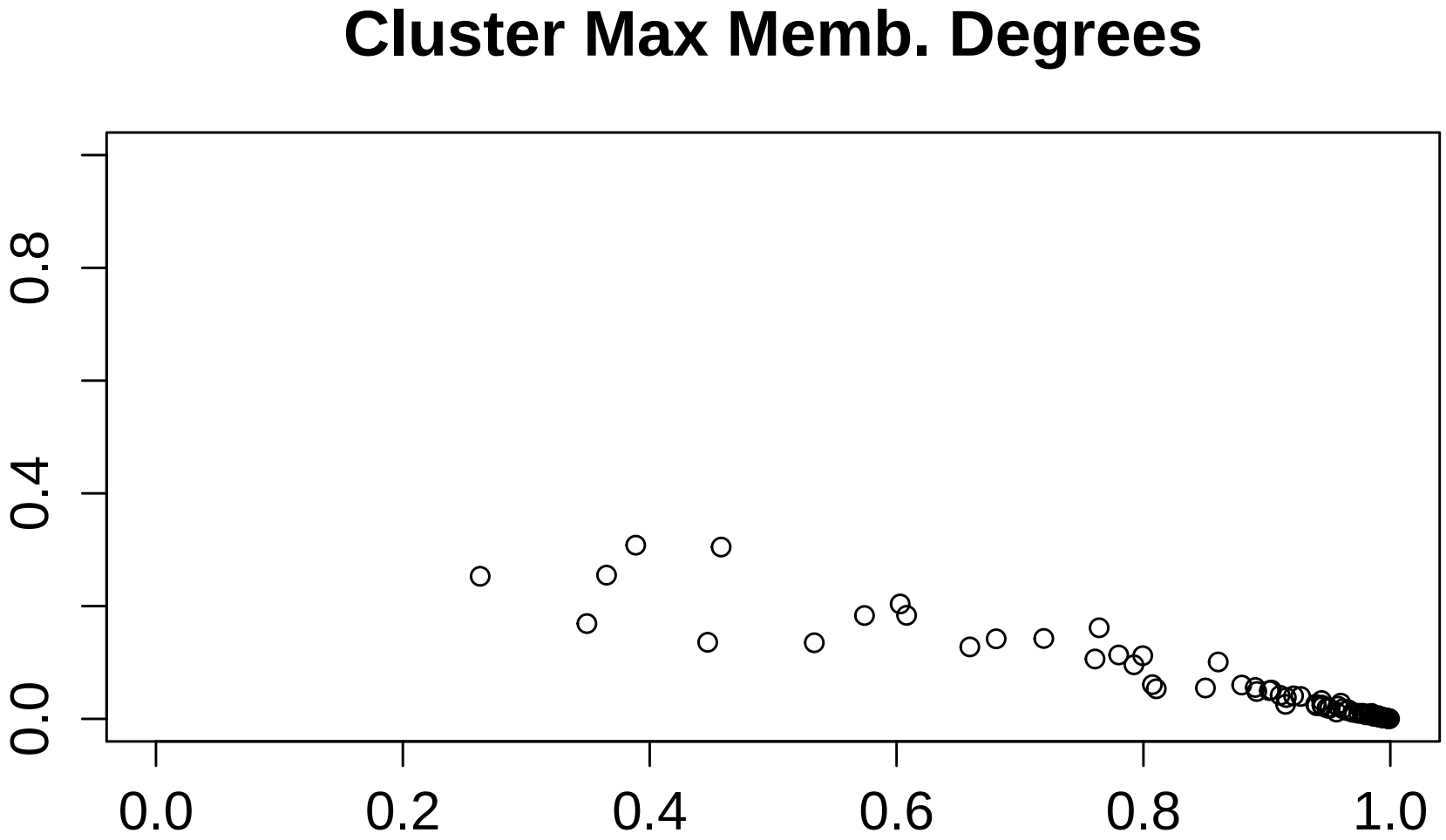
```
> VCV2(fkm6$Xca, fkm6$U, 2)
```



Results Visualization: VIFCR

Function `VIFCR(fclust.obj, which)`

```
> VIFCR(fkm6,2)
```



Unemployment data

The data set contains the unemployment rates and shares of 32 European countries in 2011 (source: Eurostat).

```
> library("fclust")
> data(unemployment)
```

variables:

- `numeric`:
  - `Total.Rate`: the percentage of unemployed persons aged 15-74 in the economically active population
  - `Youth.Rate`: the youth unemployment rate, defined as the unemployment rate for young people aged between 15 and 24
  - `LongTerm.Share`: the long-term unemployment share, defined as the Percentage of unemployed persons who have been unemployed for 12 months or more

Aim of the analysis

Aim

We are interested in finding homogeneous groups of countries characterized by similar unemployment structures.

Correlation structure

$$Corr = \begin{bmatrix} 1 & 0.92 & 0.58 \\ 0.92 & 1 & 0.54 \\ 0.58 & 0.54 & 1 \end{bmatrix}$$

We decide to apply the Gustafson and Kessel extension of f*k*m (function `FKM.gk`) in order to explore the existence of clusters having non-spherical shapes.

FkM.gk with $k = 3$ clusters

Prior analyses on the data set suggest to run the algorithm using standardized data (`stand = 1`), and to choose $k = 3$ (`k = 3`) clusters (the default value $m = 2$ is set). The here-considered algorithm has a high risk of hitting local optima and, thus, 50 random starts are used (`RS = 50`).

```
> clust <- FKM.gk(unemployment, k = 3, RS = 50,
          stand = 1)
> cl.size(clust$U)
 Clus 1  Clus 2  Clus 3
     15       6      11
```

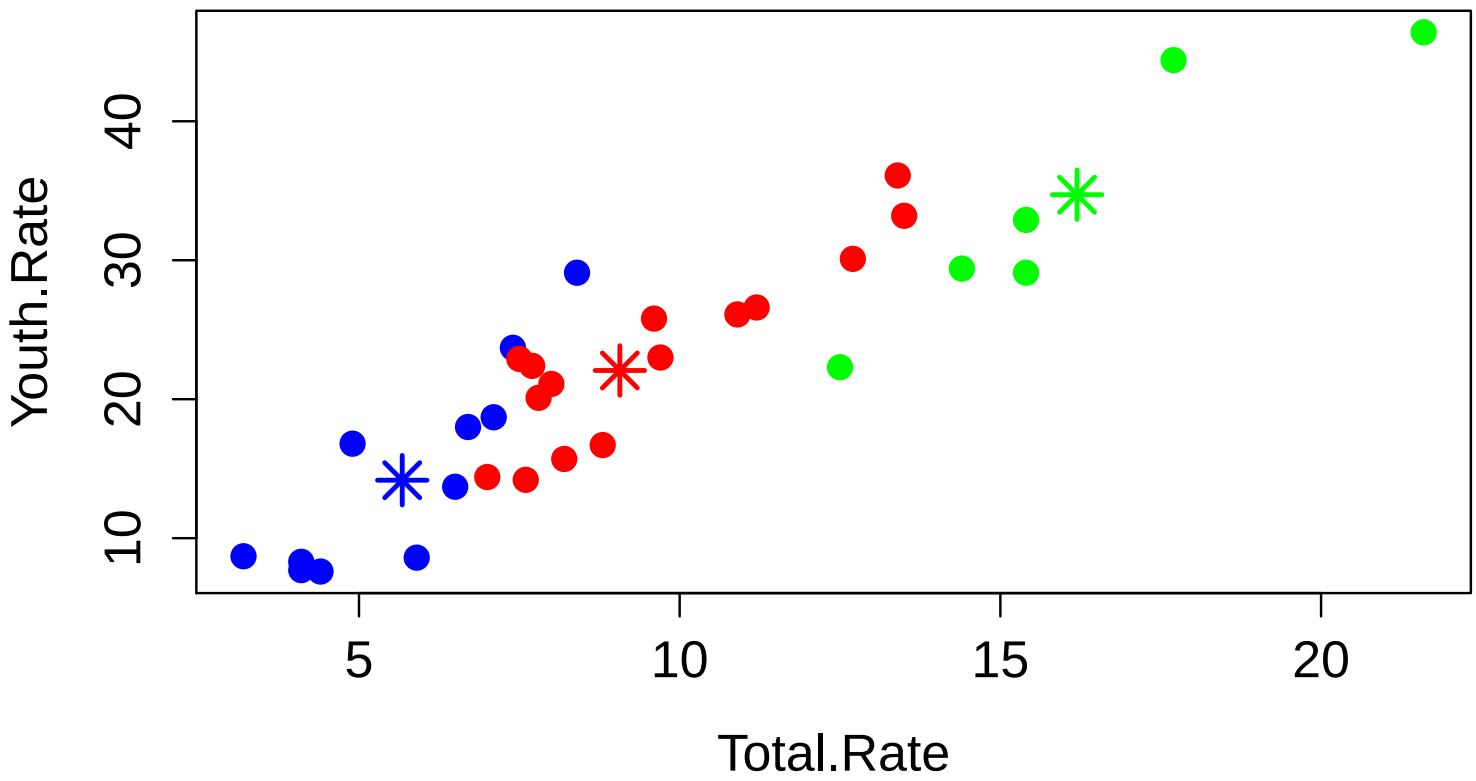
Clusters: covariance matrices

```
> clust$F
, , Clus 1
                Total.Rate  Youth.Rate  LongTerm.Share
 Total.Rate       1.299352    1.386309        2.770606
 Youth.Rate       1.386309    2.088642        2.875459
 LongTerm.Share   2.770606    2.875459        7.180983
, , Clus 2
                Total.Rate  Youth.Rate  LongTerm.Share
 Total.Rate       3.214435    3.511246       -1.801111
 Youth.Rate       3.511246    4.683005       -1.961230
 LongTerm.Share  -1.801111   -1.961230        1.376300
, , Clus 3
                Total.Rate  Youth.Rate  LongTerm.Share
 Total.Rate       1.268973    1.859881        1.906008
 Youth.Rate       1.859881    3.822880        2.140836
 LongTerm.Share   1.906008    2.140836        3.969645
```

Results Visualization: plot.fclust (i)

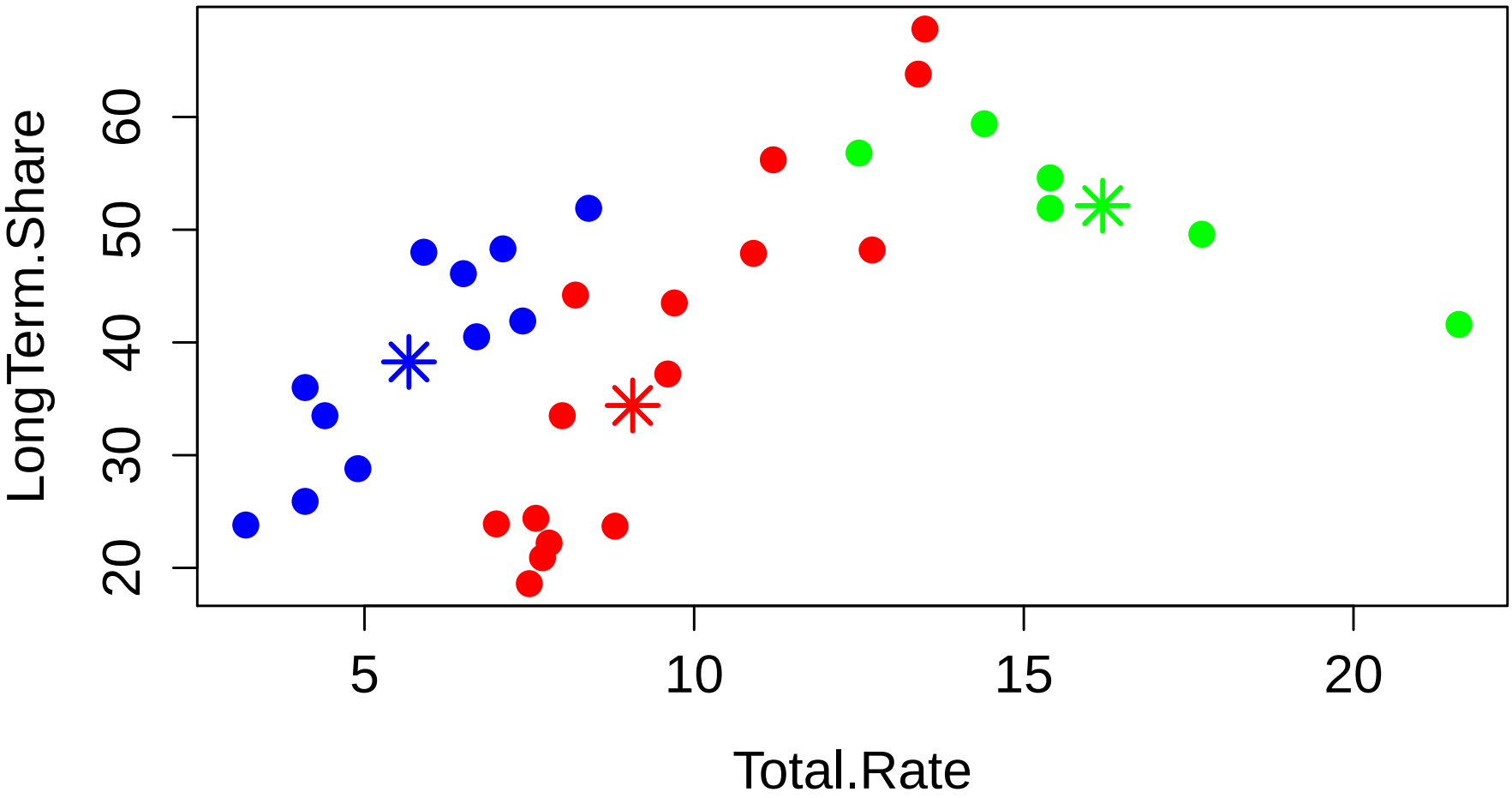
Function `plot.fclust(fclust.obj, v1v2, colclus, umin, ucex, pca)`

```
> plot.fclust(clust)
```



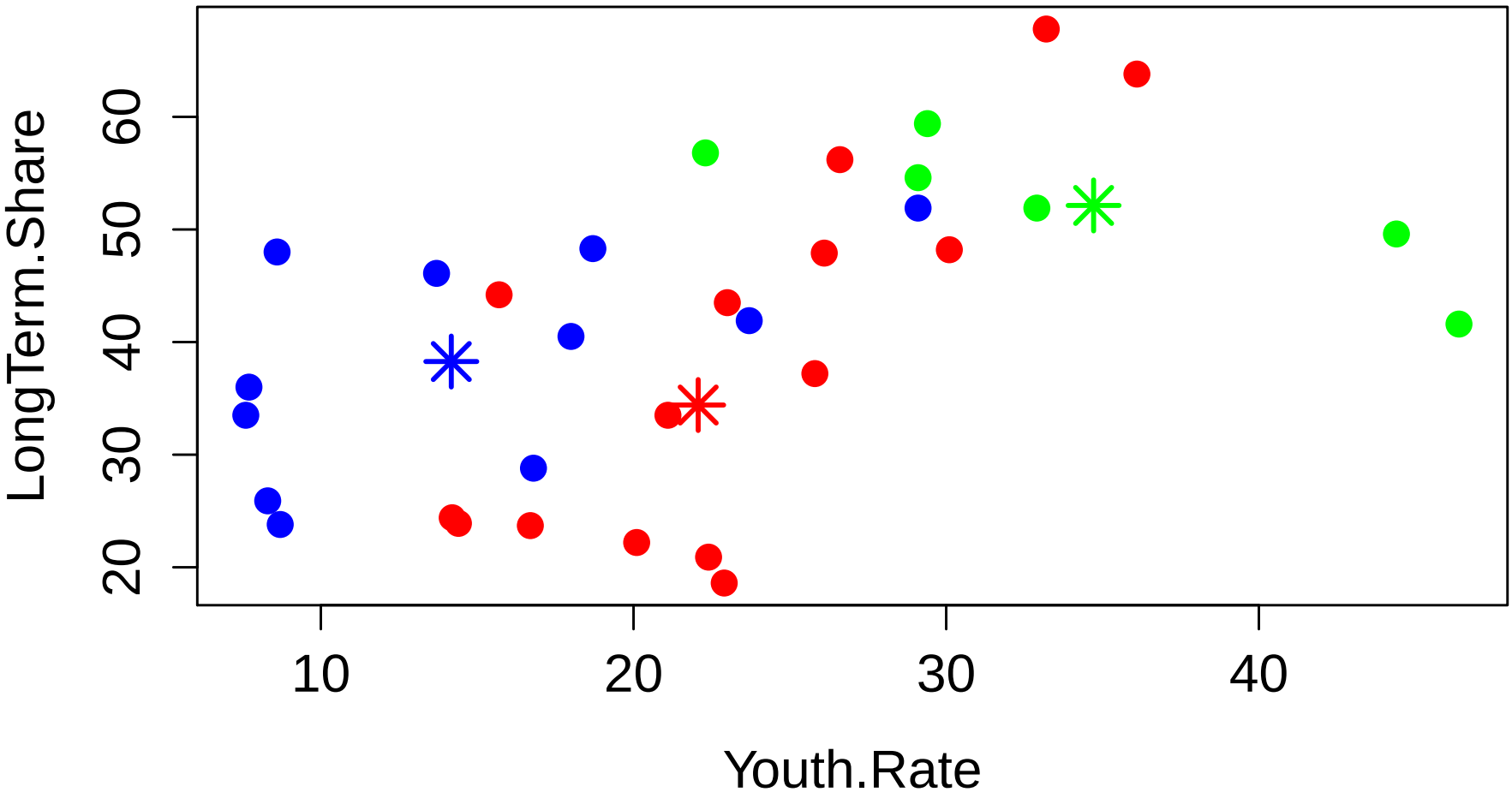
Results Visualization: plot.fclust (ii)

```
> plot.fclust(clust,v1v2=c(1,3))
```



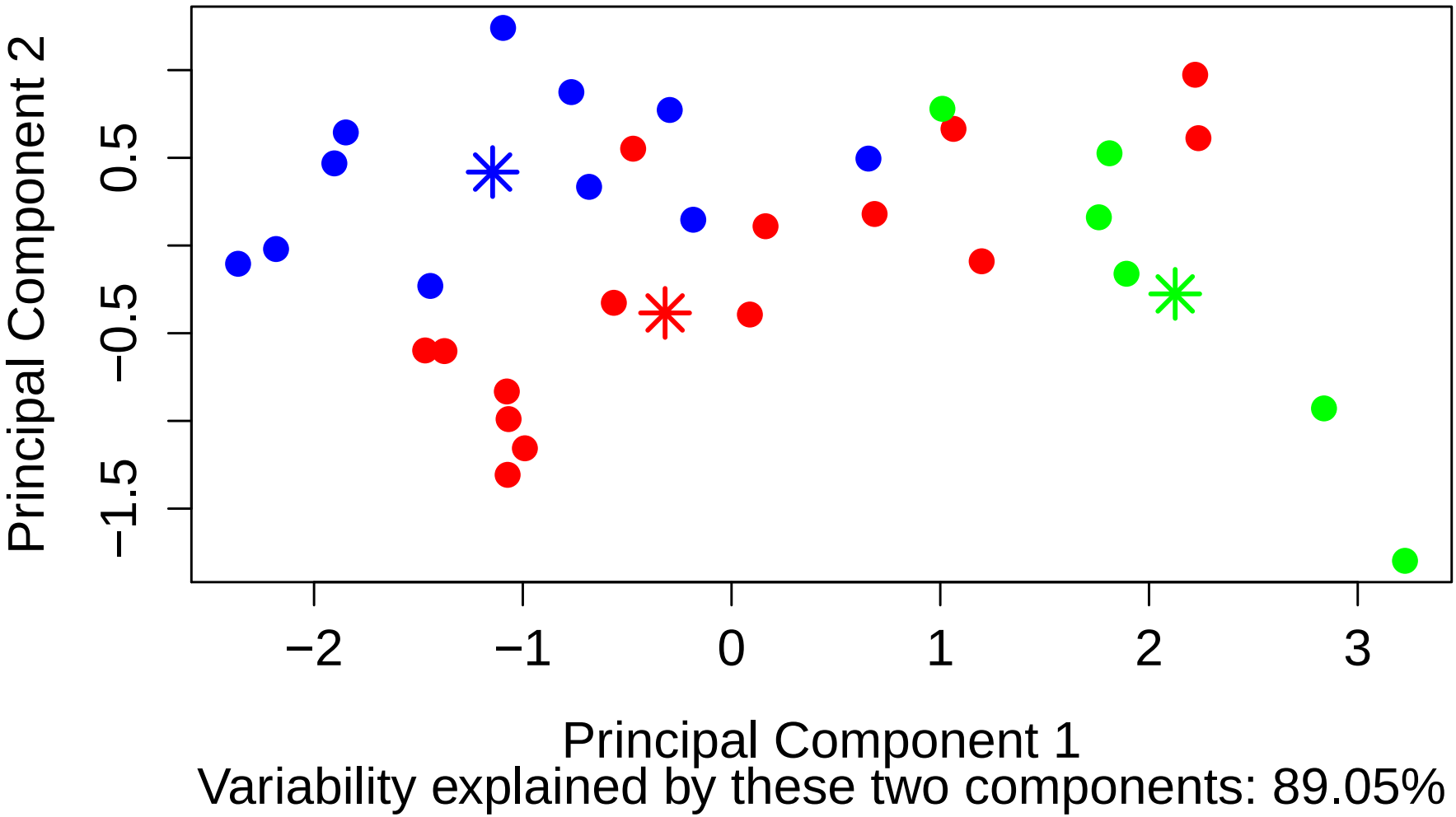
Results Visualization: plot.fclust (iii)

```
> plot.fclust(clust,v1v2=c(2,3))
```



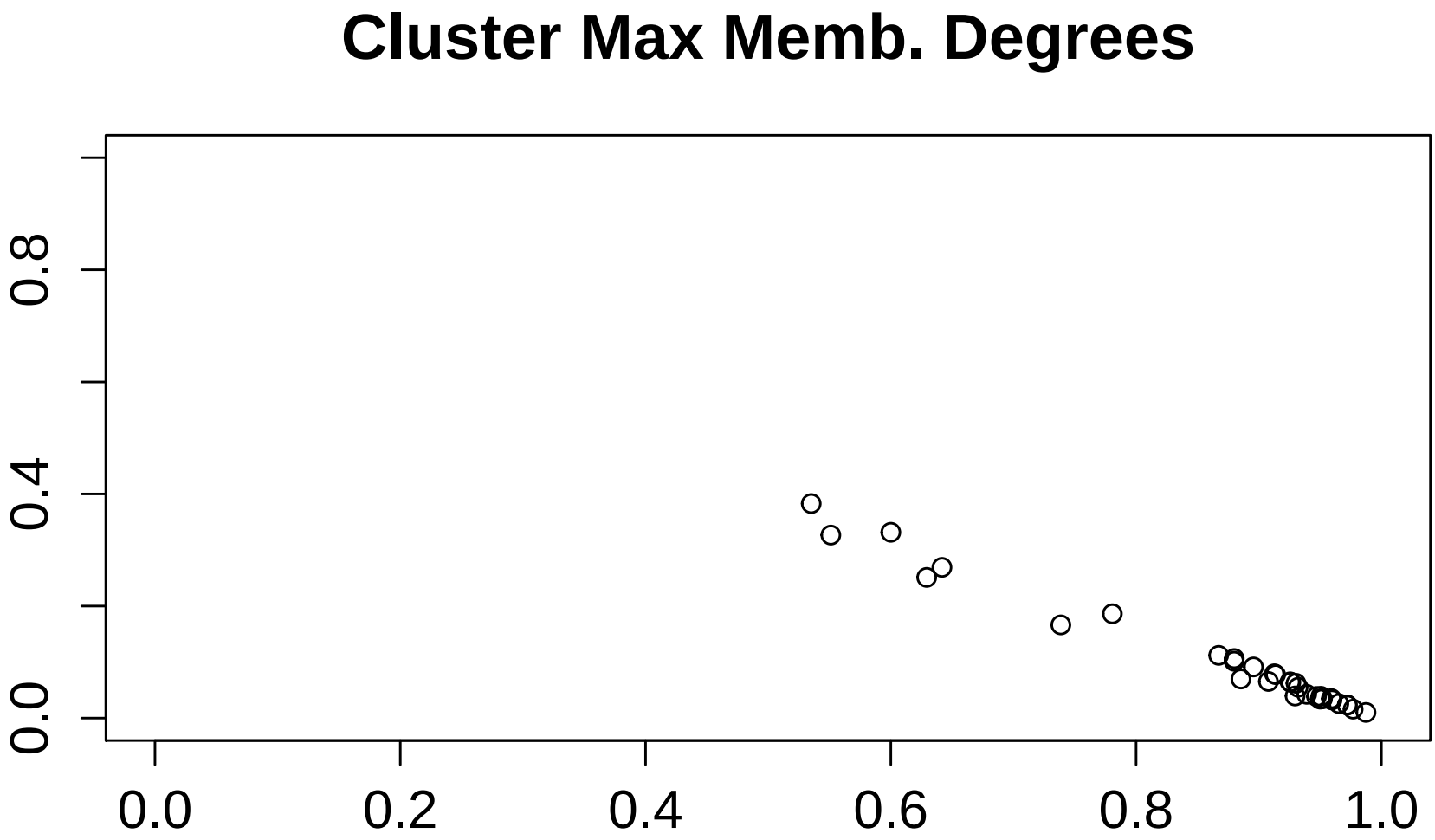
Results Visualization: plot.fclust (iv)

```
> plot.fclust(clust, pca=TRUE)
```



Results Visualization: VIFCR

```
> VIFCR(clust,2)
```



Clusters

Cluster 1: {Bulgaria, Croatia, Cyprus, Portugal, Denmark, Finland, France, Hungary, Iceland, Poland, Slovakia, Slovenia, Sweden, Turkey, UK}

Cluster 2: {Estonia, Ireland, Greece, Latvia, Lithuania, Spain}

Cluster 3: {Austria, Belgium, Czech Republic, Germany, Italy, Luxembourg, Malta, Netherlands, Norway, Romania, Switzerland}

Centroids

We now interpret the obtained clusters by studying the centroids (using the function `Hraw`) and the membership degree information.

```
> round(Hraw(clust$X, clust$H), 2)
         Total.Rate  Youth.Rate  LongTerm.Share
Clus 1         9.07       22.07           34.41
Clus 2        16.20       34.71           52.13
Clus 3         5.67       14.17           38.27
```

Interpretation of the clusters

- Cluster 2 is composed by the Baltic states and a subset of the European countries mostly suffering from the economic crisis. Such a cluster is characterized by the highest levels of all the variables, hence highlighting a critical situation.
- By inspecting the centroids we can conclude that Cluster 1 detects countries with medium total and young unemployment rates and low long-term unemployment shares. Therefore, Cluster 1 seems to highlight dynamic labor markets.
- On the contrary, Cluster 3 represents static labor markets. In detail, it is composed by countries with low total and young unemployment rates and medium long-term unemployment share.

Fish data

Food balance sheet of Fish, year 2009 (FAO)

variables:

- `numeric`:
  - `Production` (tonnes in live weight)
  - `Imports` (tonnes in live weight)
  - `Exports` (tonnes in live weight)
  - `Population`: (thousands)
  - `PCSupply`: Supply (kilograms per capita per year)
  - `FishProtPC`: Fish Proteins (grams per capita per day)
  - `AnimalProtPC`: Animal Proteins (grams per capita per day)
  - `TotalProtPC`: Total Proteins (grams per capita per day)

units: 40 countries

Aim of the analysis

Aim

We are interested in finding homogeneous groups of countries characterized by similar behaviour related to production, imports and exports of fish, supply, fish, animal and total proteins.

- We have divided the first three variables by `Population`.
- We don't consider the variable `Population` in the cluster analysis.
- By inspecting the values of Fuzzy Silhouette for different number of clusters, it results that the optimal number is $k = 3$

FkM ($k = 3$ clusters)

Solution with $k = 3$ clusters (one low-size cluster)

```
> fkm   <-   FKM(X = fish, k = 3,
                m=2, stand = 1, RS = 10)
> cl.size(fkm$U)

  Cl 1  Cl 2  Cl 3
     2    20    18
```

✎ Cluster 1 contains Iceland and Faroe Island

Membership degrees

```
> round(fkm$clus[1:15,], 2)
             Cluster  Membership degree
 Albania                2                    0.90
 Austria                3                    0.59
 Belarus                2                    0.91
 Belgium                3                    0.70
 BosniaHerz             2                    0.91
 Bulgaria               2                    0.92
 Croatia                2                    0.95
 CzechRep               2                    0.92
 Denmark                3                    0.64
 Estonia                2                    0.79
 FaroeIs                1                    0.96
 Finland                3                    0.94
 FYRMacedonia           2                    0.88
 France                 3                    0.97
 Germany                3                    0.56
```

FkM with polynomial fuzzifier ($k = 3$ clusters)

Solution with $k = 3$ clusters (one low-size cluster)

```
> fkm.pf    <-    FKM.pf(X = fish, k = 3,
                   beta = 0.5, stand = 1, RS = 10)
> cl.size(fkm.pf$U)

  Cl 1   Cl 2   Cl 3
    19      2     19
```

- Cluster 2 contains Iceland and Faroe Island
- They seem to be noisy data

Membership degrees

```
> round(fkm.pf$clus[1:15,], 2)
             Cluster  Membership degree
Albania            1                1.00
Austria            3                0.92
Belarus            1                1.00
Belgium            3                1.00
BosniaHerz         1                1.00
Bulgaria           1                1.00
Croatia            1                1.00
CzechRep           1                1.00
Denmark            3                1.00
Estonia            1                1.00
FaroeIs            2                1.00
Finland            3                1.00
FYRMacedonia       1                1.00
France             3                1.00
Germany            3                0.82
```

FkM with polynomial fuzzifier and noise clusters ($k = 2$ clusters)

Solution with $k = 2$ clusters

```
> fkm.pf.noise   <-   FKM.pf.noise(X = fish, k = 2,
                     beta = 0.5, stand = 1, RS = 10)
```

Membership degrees (more relevant countries)

```
> fkm.pf.noise$U
              Clus 1       Clus 2
Austria   0.28612643   0.67056640
FaroeIs   0.00000000   0.03307934
Germany   0.41432136   0.55093605
Iceland   0.00000000   0.29940487
Russian   0.66430809   0.30232321
```

Clusters

Cluster 1: {Albania, Belarusm, BosniaHerz, Bulgaria, Croatia, CzechRep, Estonia, FYRMacedonia, Hungary, Latvia, MoldovaRep, Montenegro, Poland, Romania, Russian, Serbia, Slovakia, Slovenia, Switzerland, Ukrainee}

Cluster 2: { Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Lithuania, Luxemburg, Malta, Netherlands, Norway, Portugal, Spain, Sweden, UK }

Noise cluster: {Faroels, Iceland}

Mean values (more relevant variables)

```
> round(apply(fkm.pf.noise$X[,c(1,2,3,4)],2,mean),2)
   Production   Imports   Exports  PCSupply
       456.55     35.34    261.54     23.39
```

Centroids (more relevant variables)

```
> fkm.pf.noise$Hraw= Hraw(fkm.pf.noise$X, fkm.pf.noise$H)
> round(fkm.pf.noise$Hraw[,c(1,2,3,4)],2)
         Production  Imports  Exports  PCSupply
Clus 1        10.39    12.14    10.73     10.93
Clus 2       139.06    48.14    99.53     31.36
```

FaroeIs (more relevant variables)

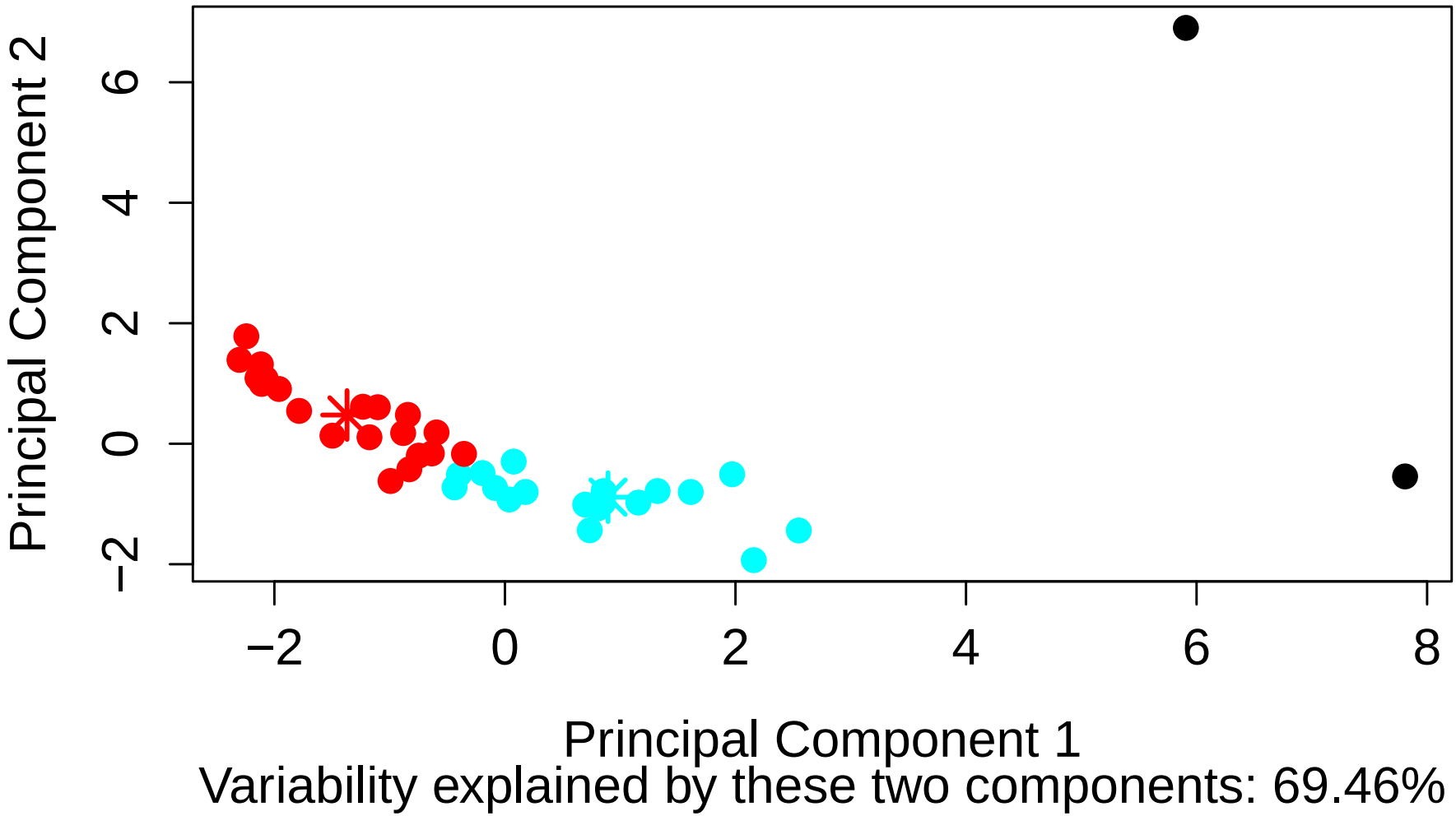
```
> round(fkm.pf.noise$X[``FaroeIs'',c(1,2,3,4)],2)
          Production  Imports  Exports  PCSupply
FaroeIs     12491.59   115.47  6735.61     87.70
```

Iceland (more relevant variables)

```
> round(fkm.pf.noise$X[``Iceland'',c(1,2,3,4)],2)
          Production  Imports  Exports  PCSupply
Iceland      4443.25   240.46  2477.47     88.30
```

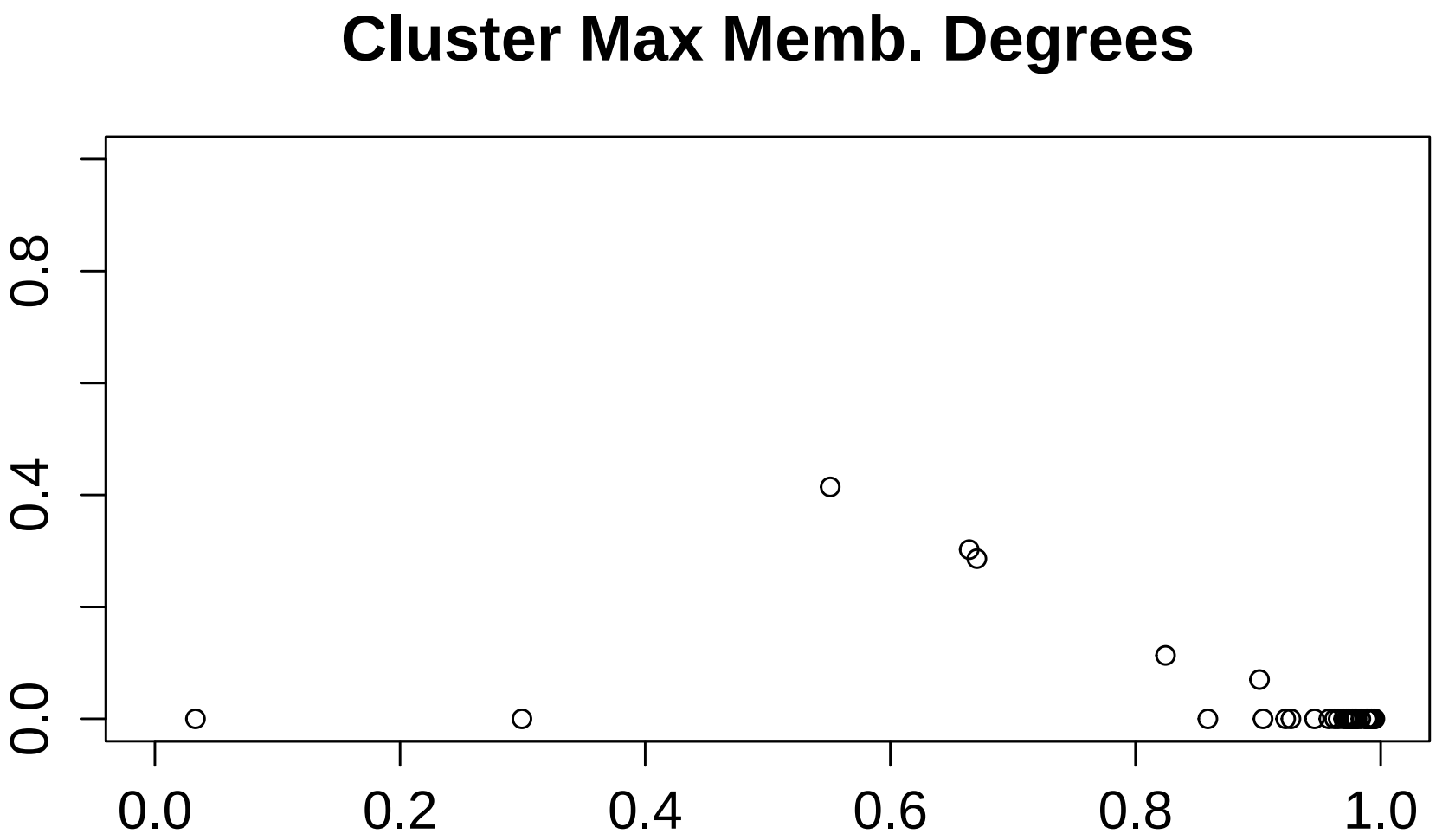
Results Visualization: plot.fclust

```
> plot.fclust(fkm.pf.noise, pca=TRUE)
```



Results Visualization: VIFCR

```
> VIFCR(fkm.pf.noise,2)
```



Conclusions

package fclust (version 1.0.1)

- Functions for fuzzy clustering algorithms
- Functions for fuzzy cluster validity indices

▷ http://cran.r-project.org/web/packages/fclust/index.html

package fclust (version 1.1)

- Noise variants for all fuzzy clustering algorithms
- Fuzzy clustering algorithm with polynomial fuzzifier approach
- Visualization techniques for (fuzzy) clustering
  (displaying clusters, validation, ...)

▷ December 2014 / January 2015

References

Bezdek JC. Journal of Cybernetics **3**, 58–73 (1974)

Bezdek JC. Pattern recognition with fuzzy objective function algorithm (1981)

Bezdek JC, Hathaway RJ. In: Proc. IJCNN 2002, IEEE Press, Piscataway, NJ, pp. 2225–2230 (2002).

Campello RJGB, Hruschka ER. Fuzzy Sets and Systems **157**, 2858–2875 (2006)

Davé RN. Pattern Recognition Letters **12**, 657–664 (1991)

Davé, RN. Pattern Recognition Letters **17**, 613–623 (1996)

Gustafson E, Kessel W. In: Proceedings of IEEE CDC (1978)

Hathaway RJ, Bezdek JC. Pattern Recognition Letters **24**, 1563–1569 (2003).

Huband JM, Bezdek JC. In: J.M. Zurada et al. (eds.), WCCI 2008, LNCS 5050, pp. 293–308 (2008).

Klawonn F, Höppner F. In: Advances in intelligent data analysis (LNCS 2779, pp. 254-264) (2003).

Klawonn F, Chekhtman V, Janz E. In: J. Benitez, O. Cordon, F. Hoffmann, R. Roy (eds.): Advances in Soft Computing: Engineering Design and Manufacturing. Springer, London, 65–76 (2003).

Kaufman L, Rousseeuw PJ. In: Statistical Data Analysis Based on the L1 -norm and Related Methods, 405-416 (1987)

Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis, J. Wiley and Sons (1990)

Li RP, Mukaidono M. In: Proceedings of FUZZ-IEEE/IFES 95, pp. 2227–2232 (1995)

Li RP, Mukaidono M. Fuzzy Sets and Systems **102**, 253–258 (1999)

Xie XL, Beni G. IEEE Transactions on Pattern Analysis and Machine Intelligence **13**, 841–847 (1991)

Thank you!

e-mail: **mariabrigida.ferraro@uniroma1.it**