Fuzzy Clustering

Hard-c-Means, Fuzzy-c-Means, Gustafson-Kessel

Olaf Wolkenhauer

Control Systems Centre

UMIST

o.wolkenhauer@umist.ac.uk

www.csc.umist.ac.uk/people/wolkenhauer.htm

Contents

Learning Objectives

- A more general concept to represent data sampled from a system is that of a data space.
- System properties and behaviour are reflected by clusters of data.
- Clusters may be interpreted as linear submodels of an overall non-linear system.
- Clusters may also be interpreted as if-then rules relating properties of the variables that form the data space.
- Fuzzy clustering provides least-squares solutions to the identify clusters, to partition the data space into clusters or classes.
- Fuzzy boundaries between clusters are differentiable functions and hence are computationally attractive.
- For many real-world problems a fuzzy partitioning of the underlying space is more realistic than ‘hard clustering’.

1. Pattern Recognition

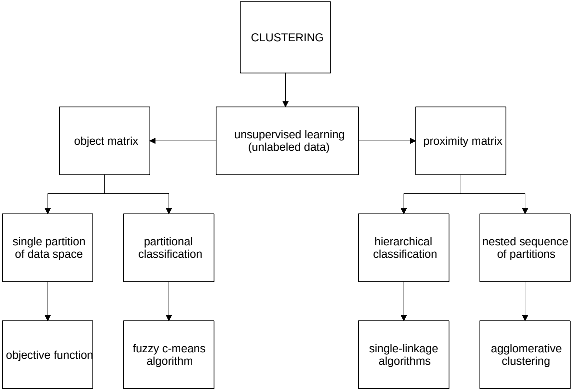
▷ The objective of cluster analysis is the *classification* of objects according to similarities among them.

▷ *Agglomerative Hierarchical Methods*: New clusters are formed by reallocating memberships of one point at a time. Results in a nested sequence of partitions (dendrogram plot). Example: Link algorithms.

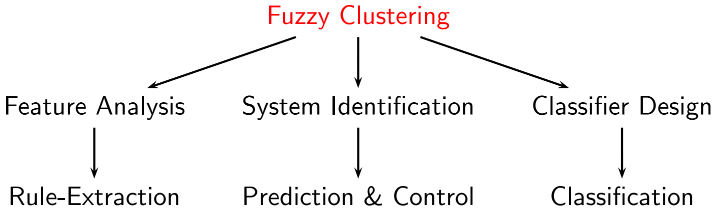
▷ *Partitional Objective Function Clustering*: Group data into clusters such that the objects in one group are more similar to each other than objects in other clusters.

✘ Fuzzy clustering with quadratic objective functions leads to least-squares optimisations.

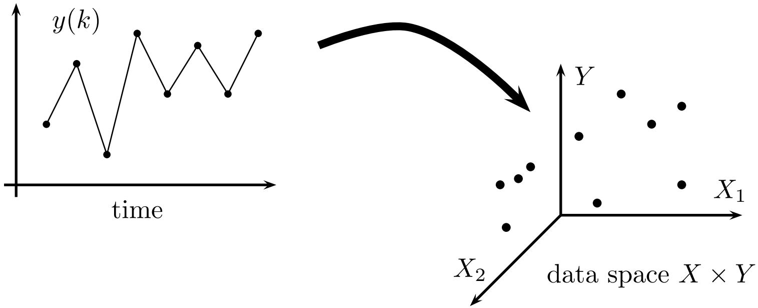
1.1. Hierarchical Clustering



1.2. Overview: Applications



1.3. From Time-Series to Pattern in Data Space



1.4. Data Space, Data Matrix

▷ *Data space*: $\Xi \subset \mathbb{R}^n$

▷ *Objects*: $\mathbf{o} \in \Xi$

▷ $j = 1, \ldots, d$ measurements, observations, *data objects*

$$\mathbf{m}_j = [m_{1j}, \ldots, m_{nj}]^T$$

▷ $n \times d$ *data matrix:*

$$\mathbf{M} = \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1d} \\ m_{21} & m_{22} & \cdots & m_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n1} & m_{n2} & \cdots & m_{nd} \end{bmatrix}$$

1.5. Example: NARX Model:

Second-order model structure :

$$y(k+1) = f\big(y(k), y(k-1), u(k), u(k-1)\big)$$

Regressor vector :

$$\mathbf{x} \doteq \big[y(k), y(k-1), u(k), u(k-1)\big]^T$$

Data vectors

$$\mathbf{m}_j = \big[y(j), y(j-1), u(j), u(j-1), y(j+1)\big]^T$$

with $n = r + 1$, forming the matrix :

$$\mathbf{M} = \begin{bmatrix} y(2) & y(3) & \cdots & y(d-1) \\ y(1) & y(2) & \cdots & y(d-2) \\ u(2) & u(3) & \cdots & u(d-1) \\ u(1) & u(2) & \cdots & y(d-2) \\ y(3) & y(4) & \cdots & y(d) \end{bmatrix}$$

2. Clustering

▷ Objective: Group objects $\mathbf{m}_j$ into c clusters.

▷ Assume the clusters exist, let $\mathbf{C} = [\mathbf{c}^{(1)}, \ldots, \mathbf{c}^{(c)}]$ be a set of *prototypes* or cluster centres :

$$\mathbf{c}^{(i)} = \frac{\sum_{j=1}^{d} u_{ij} \cdot \mathbf{m}_j}{\sum_{j=1}^{d} u_{ij}} \qquad i = 1, 2, \ldots, c$$

▷ $u_{ij} \in \mathbf{U}$ denotes the membership of $\mathbf{m}_j$ in the ith cluster. $\mathbf{U}$ is therefore called *partition matrix*.

✘ A cluster can be seen as describing an *equivalence class*

$$[\mathbf{c}^{(i)}]_E \doteq \left\{ \mathbf{o} : \ \mathbf{o} \in \Xi, \ E(\mathbf{c}^{(i)}, \mathbf{o}) = 1 \right\} .$$

2.1. Equivalence Relations

▷ An equivalence relation is *reflexive*, $E(\mathbf{o}, \mathbf{o}) = 1$, *symmetric*, $E(\mathbf{o}, \mathbf{o}') = 1$ implies $E(\mathbf{o}', \mathbf{o}) = 1$ and *transitive*, $E(\mathbf{o}, \mathbf{o}') = 1$ and $E(\mathbf{o}', \mathbf{o}'') = 1$ implies $E(\mathbf{o}, \mathbf{o}'') = 1$.

▷ If a cluster is described by an *equivalence class*

$$[\mathbf{c}^{(i)}]_E \doteq \left\{ \mathbf{o} : \ \mathbf{o} \in \Xi, \ E(\mathbf{c}^{(i)}, \mathbf{o}) = 1 \right\} .$$

then the set of equivalence classes $\{[\mathbf{c}^{(i)}]_E\}$ forms a *partition*.

▷ The set of equivalence classes is called a *quotient set*

$$\Xi / E \doteq \left\{ [\mathbf{c}^{(i)}]_E \right\} .$$

✘ The map from Ξ onto Ξ/E, called *natural map*, defines a *classifier*.

$$\psi : \ \Xi \ \rightarrow \ \Xi/E \qquad \mathbf{o} \ \mapsto \ [\mathbf{o}']_E$$

3. Hard-c-Means Clustering

▷ Let c be the number of clusters, the *hard partitioning space* :

$$M_{hc} = \left\{ \mathbf{U} \in V_{cd} \;:\; u_{ij} \in \{0,1\}, \forall (i,j);\; \sum_{i=1}^{c} u_{ij} = 1;\; 0 < \sum_{j=1}^{d} u_{ij} < d, \forall i \right\}$$

▷ Clustering criterion (*objective function*, cost function) :

$$J_{hc}(\mathbf{M}; \mathbf{U}, \mathbf{C}) = \sum_{i=1}^{c} \sum_{j=1}^{d} u_{ij}\; d^2_{\mathbf{A}} \left(\mathbf{m}_j, \mathbf{c}^{(i)} \right)$$

▷ *Distance measure* :

$$d^2_{\mathbf{A}} \left(\mathbf{m}_j, \mathbf{c}^{(i)} \right) \doteq \left\| \mathbf{m}_j - \mathbf{c}^{(i)} \right\|^2_{\mathbf{A}} = \left(\mathbf{m}_j - \mathbf{c}^{(i)} \right)^T \mathbf{A} \left(\mathbf{m}_j - \mathbf{c}^{(i)} \right)$$

3.1. Objective Function, Partition Space

▷ Start with an initial partition (randomly chosen).

▷ Minimise the 'within-cluster-overall-variance' :

$$(\mathbf{U}, \mathbf{C}) = \arg \min_{M_{hc} \times \mathbb{R}^{d \times c} \times V_{dd}} J_{hc}(\mathbf{M}; \mathbf{U}, \mathbf{C}, \mathbf{A})$$

✘ Problem: Due to the discrete nature of u_{ij}, the size of the partition space is huge :

$$|M_{hc}| = \frac{1}{c!} \left[\sum_{i=1}^{c} \binom{c}{i} (-1)^{c-i} \cdot i^d \right] .$$

3.2. Hard-c-Means Algorithm

Repeat for $l = 1, 2, \dots$:

Step 1: Calculate centres of clusters; c-mean vectors :

$$\mathbf{c}_l^{(i)} = \left(\sum_{j=1}^{d} u_{ij}^{(l-1)} \cdot \mathbf{m}_j \right) \Big/ \left(\sum_{j=1}^{d} u_{ij}^{(l-1)} \right) , \quad 1 \leq i \leq c \; .$$

Step 2: Update $\mathbf{U}^{(l)}$: Reallocate cluster memberships to minimise squared errors:

$$u_{ij}^{(l)} = \begin{cases} 1 & \text{if } d(\mathbf{m}_j, \mathbf{c}_i^{(l)}) = \min_{1 \leq k \leq c} d(\mathbf{m}_j, \mathbf{c}_k^{(l)}) \\ 0 & \text{otherwise.} \end{cases}$$

Until $\left\| \mathbf{U}^{(l)} - \mathbf{U}^{(l-1)} \right\| < \delta$.

4. Fuzzy Clustering

▷ *Fuzzy partition space* (cf. M_{hc})

$$M_{fc} = \left\{ \mathbf{U} \in V_{cd} \ : \ u_{ij} \in [0,1], \forall (i,j); \ \sum_{i=1}^{c} u_{ij} = 1; \ 0 < \sum_{j=1}^{d} u_{ij} < d, \forall i \right\}$$

▷ *Fuzzy objective function* .. is a least-squares functional :

$$J_{fc}(\mathbf{M}; \mathbf{U}, \mathbf{C}) = \sum_{i=1}^{c} \sum_{j=1}^{d} (u_{ij})^{w} \ d_{\mathbf{A}}^{2} \left(\mathbf{m}_j, \mathbf{c}^{(i)} \right)$$

▷ *Weighting factor* $w \in [1, \infty)$.

- $w \to 1$: hard, crisp clustering.
- $w \to \infty$: $u_{ij} \to 1/c$.
- Typical values: 1.25 and 2.

4.1. Fuzzy-c-Means Algorithm

Preparations:

1. Fix c, $2 \leq c < d$
2. Choose any inner product norm metric for $\mathbb{R}^n$.
3. Choose the termination tolerance $\delta > 0$, e.g between 0.01 and 0.001.
4. Fix w, $1 \leq w < \infty$, e.g 2.
5. Initialise $\mathbf{U}^{(0)} \in M_{fc}$, (e.g randomly).

Repeat for $l = 1, 2, \ldots$:

1. **Step 1:** Compute cluster prototypes:

$$\mathbf{c}_l^{(i)} = \frac{\sum_{j=1}^{d} \left(u_{ij}^{(l-1)} \right)^w \mathbf{m}_j}{\sum_{j=1}^{d} \left(u_{ij}^{(l-1)} \right)^w}, \quad 1 \leq i \leq c .$$

2. **Step 2:** Compute distances:

   For all clusters $1 \leq i \leq c$,

   For all data objects $1 \leq j \leq d$,

$$d_{\mathbf{A}}^2 \left(\mathbf{m}_j, \mathbf{c}_l^{(i)} \right) = \left(\mathbf{c}_l^{(i)} - \mathbf{m}_j \right)^T \mathbf{A} \left(\mathbf{c}_l^{(i)} - \mathbf{m}_j \right) .$$

1. **Step 3:** Update the partition matrix:

If $d_{\mathbf{A}}\left(\mathbf{m}_j, \mathbf{c}_l^{(i)}\right) > 0$ for $1 \le i \le c$, $1 \le j \le d$,

$$u_{ij}^{(l)} = \frac{1}{\sum_{k=1}^{c} \left(d_{\mathbf{A}}^2(\mathbf{m}_j, \mathbf{c}^{(i)})/d_{\mathbf{A}}^2(\mathbf{m}_j, \mathbf{c}^{(k)})\right)^{1/(w-1)}}$$

otherwise

$$u_{ij}^{(l)} = 0 \text{ if } d_{\mathbf{A}}\left(\mathbf{m}_j, \mathbf{c}_l^{(i)}\right) > 0, \text{ and } u_{ij}^{(l)} \in [0,1] \text{ with } \sum_{i=1}^{c} u_{ij}^{(l)} = 1 \; .$$

Until $\left\|\mathbf{U}^{(l)} - \mathbf{U}^{(l-1)}\right\| < \delta$.

5. Example: Butterfly Data Set

The data set **M** consists of 15 points in the plane :

j	1	2	3	4	5	6	7	8
$\mathbf{m}_j$	(0,0)	(0,2)	(0,4)	(1,1)	(1,2)	(1,3)	(2,2)	(3,2)
j	9	10	11	12	13	14	15	
$\mathbf{m}_j$	(4,2)	(5,1)	(5,2)	(5,3)	(6,0)	(6,2)	(6,4)	

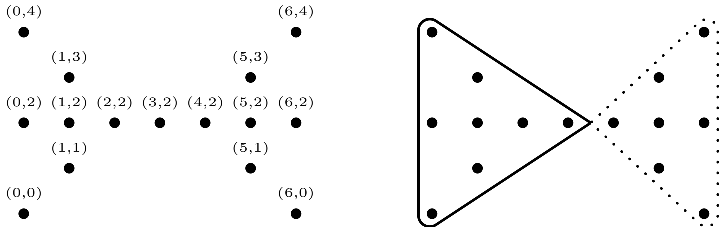


Figure 1: *The butterfly data set and hard-c-means result.*

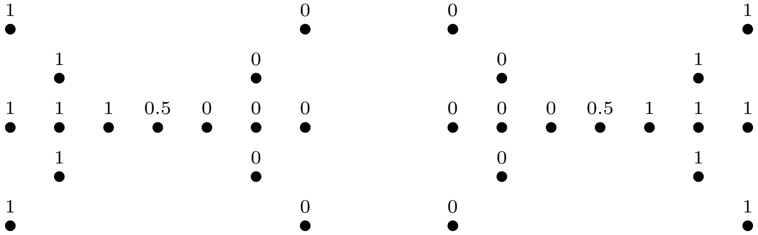


Figure 2: *Fuzzy c-means clustering of the butterfly data set.* $w = 1.25$

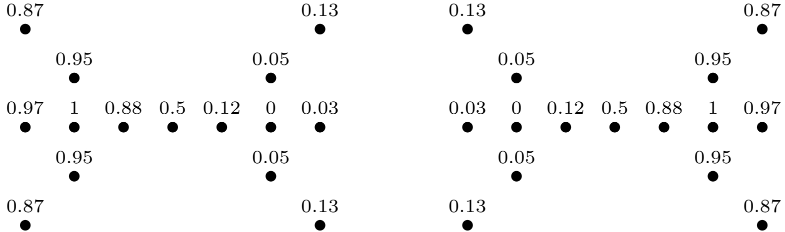


Figure 3: *Fuzzy c-means clustering of the butterfly data set.* $w = 2$

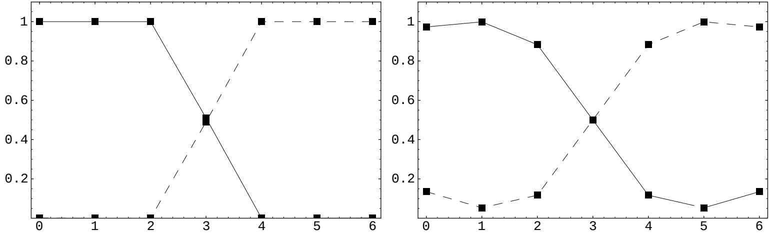
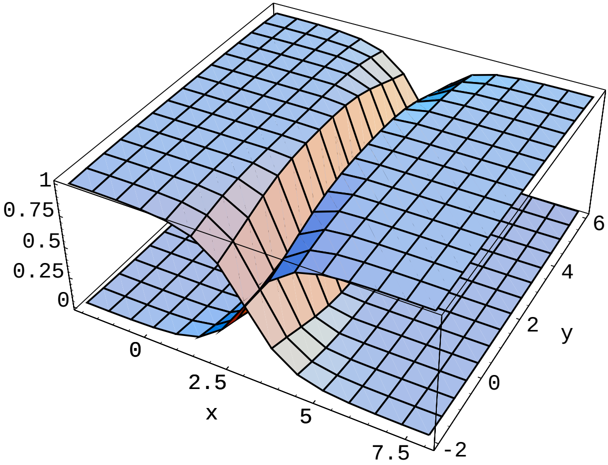


Figure 4: *FCM:* $w = 1.25$ *(left),* $w = 2$ *(right). Result after 7 iterations.*


1
0.75
0.5
0.25
0
0
2.5
x
5
7.5
-2
0
2
y
4
6

6. Gustafson-Kessel Clustering

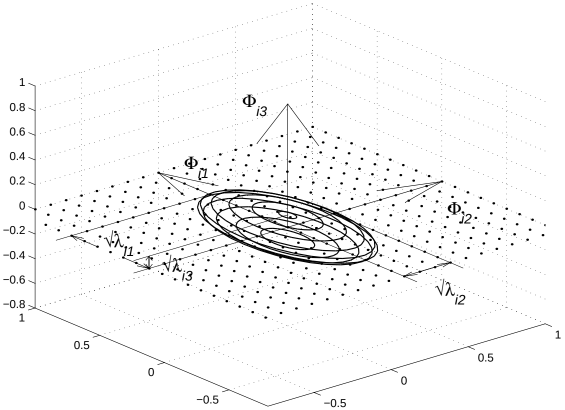
✘ Problem: Fuzzy-c-means searches for spherical clusters.

▷ Each cluster is characterised by its centre and covariance matrix :

$$\mathbf{F}^{(i)} = \frac{\sum_{j=1}^{d} (u_{ij})^w (\mathbf{m}_j - \mathbf{c}^{(i)})(\mathbf{m}_j - \mathbf{c}^{(i)})^T}{\sum_{j=1}^{d} (u_{ij})^w}$$

▷ Let λ_{ik} denote the k^{th} eigenvalue of $\mathbf{F}^{(i)}$ and Φ_{ik} the kth unit eigenvector of $\mathbf{F}^{(i)}$ and have the eigenvalues arranged in decreasing order, $\lambda_{i1} \geq \lambda_{i2} \geq \cdots \geq \lambda_{in}$.

▷ Then the eigenvectors Φ_{i1} to $\Phi_{i(n-1)}$ span the i^{th} cluster's linear subspace and the nth eigenvector Φ_{in} is the normal to this linear subspace.


Φ i3
Φ i1
Φ i2
√λ i1
√λ i3
√λ i2
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
1
0.5
0
−0.5
−0.5
0
0.5
1

6.1. Covariance Matrix

For the Gustafson-Kessel algorithm, each cluster has its own norm-inducing matrix $\mathbf{A}^{(i)}$:

$$d^2_{\mathbf{A}^{(i)}} = \left(\mathbf{c}_l^{(i)} - \mathbf{m}_j\right)^T \mathbf{A}^{(i)} \left(\mathbf{c}_l^{(i)} - \mathbf{m}_j\right) .$$

Where

$$\mathbf{A}^{(i)} \doteq \left(|\mathbf{F}^{(i)}|\right)^{1/(r+1)} \cdot \left(\mathbf{F}^{(i)}\right)^{-1} .$$

and

$$\mathbf{F}^{(i)} = \frac{\sum_{j=1}^{d} (u_{ij})^w (\mathbf{m}_j - \mathbf{c}^{(i)})(\mathbf{m}_j - \mathbf{c}^{(i)})^T}{\sum_{j=1}^{d} (u_{ij})^w}$$

6.2. Example: Nonlinear First-Order AR Process

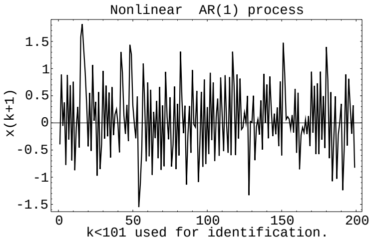
Nonlinear AR(1) process :

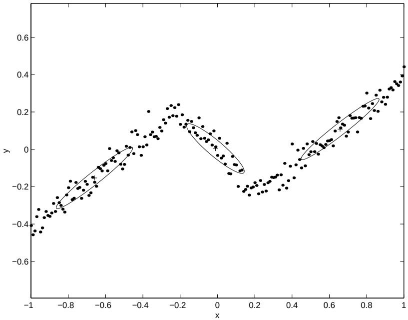
$$x(k+1) = f\big(x(k)\big) + \varepsilon(k), \quad f(x) = \begin{cases} 2x-2, & 0.5 \le x, \\ -2x, & -0.5 < x < 0.5 \\ 2x+2, & x \le -0.5 \end{cases}$$

where $\varepsilon(k) \sim N(0, \sigma^2)$ with $\sigma = 0.3$. $x(0) = 0.1$.

Model structure :

$$x(k+1) = f\big(x(k), x(k-1), \dots, x(k-r+1)\big),$$


Nonlinear AR(1) process
x(k+1)
1.5
1
0.5
0
-0.5
-1
-1.5
0
50
100
150
200
k<101 used for identification.
100 data for identification
x(k+1)
1.5
1
0.5
0
-0.5
-1
-1.5
-1.5
-1
-0.5
0
0.5
1
1.5
x(k)


0.6
0.4
0.2
0
−0.2
−0.4
−0.6
y
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
x

6.3. The Algorithm

Preparations:

- Fix c, $2 \leq c < d$
- Choose termination criteria $\delta > 0$.
- Fix w, $1 \leq w < \infty$, e.g 2.
- Initialise $\mathbf{U}^{(0)} \in M_{fc}$, (e.g randomly).

Repeat for $l = 1, 2, \ldots$:

1. **Step 1:** Compute cluster prototypes (means) :

$$\mathbf{c}_l^{(i)} = \frac{\sum_{j=1}^{d} \left(u_{ij}^{(l-1)}\right)^w \mathbf{m}_j}{\sum_{j=1}^{d} \left(u_{ij}^{(l-1)}\right)^w}, \quad 1 \le i \le c .$$

2. **Step 2:** Compute the cluster covariance matrices :

$$\mathbf{F}^{(i)} = \frac{\sum_{j=1}^{d} \left(u_{ij}^{(l-1)}\right)^w \left(\mathbf{m}_j - \mathbf{c}_l^{(i)}\right)\left(\mathbf{m}_j - \mathbf{c}_l^{(i)}\right)^T}{\sum_{j=1}^{d} \left(u_{ij}^{(l-1)}\right)^w}$$

3. **Step 3:** Compute distances for $1 \leq i \leq c$ and $1 \leq j \leq d$:

$$d^2_{\mathbf{F}^{(i)}}\left(\mathbf{c}_l^{(i)}, \mathbf{m}_j\right) = \left(\mathbf{c}_l^{(i)} - \mathbf{m}_j\right)^T \left[\left|\mathbf{F}^{(i)}\right|^{\frac{1}{(r+1)}} \cdot \left(\mathbf{F}^{(i)}\right)^{-1}\right]\left(\mathbf{c}_l^{(i)} - \mathbf{m}_j\right)$$

4. **Step 4:** Update partition matrix :
   If $d_{\mathbf{F}^{(i)}} > 0$ for $1 \leq i \leq c$, $1 \leq j \leq d$,

$$u_{ij}^{(l)} = \frac{1}{\sum_{k=1}^{c} \left(d_{\mathbf{F}^{(i)}}/d_{\mathbf{F}^{(i)}}\right)^{2/(w-1)}}$$

   otherwise

$$u_{ij}^{(l)} = 0 \text{ if } d_{\mathbf{F}^{(i)}}\left(\mathbf{c}^{(j)}, \mathbf{m}_j\right) > 0, \text{ and } u_{ij}^{(l)} \in [0,1] \text{ with } \sum_{i=1}^{c} u_{ij}^{(l)} = 1\ .$$

Until $\left\|\mathbf{U}^{(l)} - \mathbf{U}^{(l-1)}\right\| < \delta$.

References

[1] Babuska, R. : *Fuzzy Modelling for Control*. Kluwer, 1998. See http://lcewww.et.tudelft.nl/.

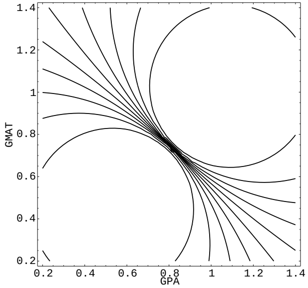
[2] Backer, E. : *Computer-Assisted Reasoning in Cluster Analysis* Prentice Hall, 1995.

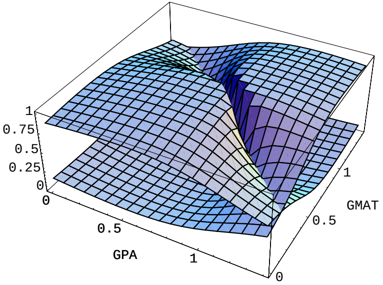
[3] Bezdek, J.C. : *Pattern Recognition with Fuzzy Objective Function Algorithms* Plenum Press, 1981.

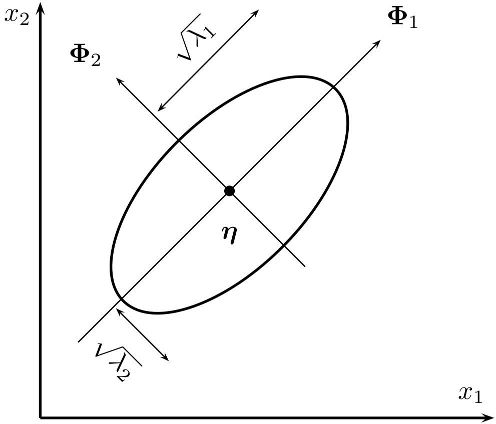
[4] Höppner, F. et.al. : *Fuzzy Cluster Analysis* John Wiley, 1999.

[5] Bezdek, J.C. and Pal, S.C. : *Fuzzy Models for Pattern Recognition* IEEE, 1992.

[6] Wolkenhauer, O. : *Data Engineering*. http://www.csc.umist.ac.uk/people/wolkenhauer.htm.


GMAT
GPA
0.2
0.4
0.6
0.8
1
1.2
1.4


GPA
GMAT
0
0.25
0.5
0.75
1


x_2
$\mathbf{\Phi}_1$
$\mathbf{\Phi}_2$
$\sqrt{\lambda_1}$
$\boldsymbol{\eta}$
$\sqrt{\lambda_2}$
x_1