01_03_Fuzzy Systems and Fuzzy Clustering

1. Fuzzy Systems

Linguistic （Mamdani）fuzzy model

$$\text{if } x \text{ is } A \text{ then } y \text{ is } B$$
$$x \text{ is } A \rightarrow \text{antecedent(fuzzy proposition)}$$
$$y \text{ is } B \rightarrow \text{consequent(fuzzy proposition)}$$

Fuzzy Relational Model

$$\text{if } x \text{ is } A \text{ then } y \text{ is } B_1(0.1), B_2(0.8)$$

Takagi-sugeno Fuzz model

$$\text{if } x \text{ is } A \text{ then } y = f(x)$$

2. Mamdani Fuzzy Systems

Given the if-then **rules** and an **input** fuzzy set, deduce the **corresponding output fuzzy set**

Fuzzy implication and Conjunctions

$$R : [0,1] \times [0,1] \rightarrow [0,1]$$
$$\mu_R(x, y) = \mathrm{I}\left(\mu_A(x), \mu_B(y)\right)$$

"classical"	Kleene–Diene	$\mathrm{I}(a, b) = \max(1 - a, b)$
	Łukasiewicz	$\mathrm{I}(a, b) = \min(1, 1 - a + b)$
T-norms	Mamdani	$\mathrm{I}(a, b) = \min(a, b)$
	Larsen	$\mathrm{I}(a, b) = a \cdot b$

implication

conjunctions

same value as classical

Mamdani implication

$$\mathrm{I}\left(\mu_A(\mathbf{x}), \mu_B(\mathbf{y})\right) = \min\left(\mu_A(\mathbf{x}), \mu_B(\mathbf{y})\right)$$

Larsen Implication

$$\mathrm{I}\left(\mu_A(\mathbf{x}), \mu_B(\mathbf{y})\right) = \mu_A(\mathbf{x}) \cdot \mu_B(\mathbf{y})$$

Normal Inference

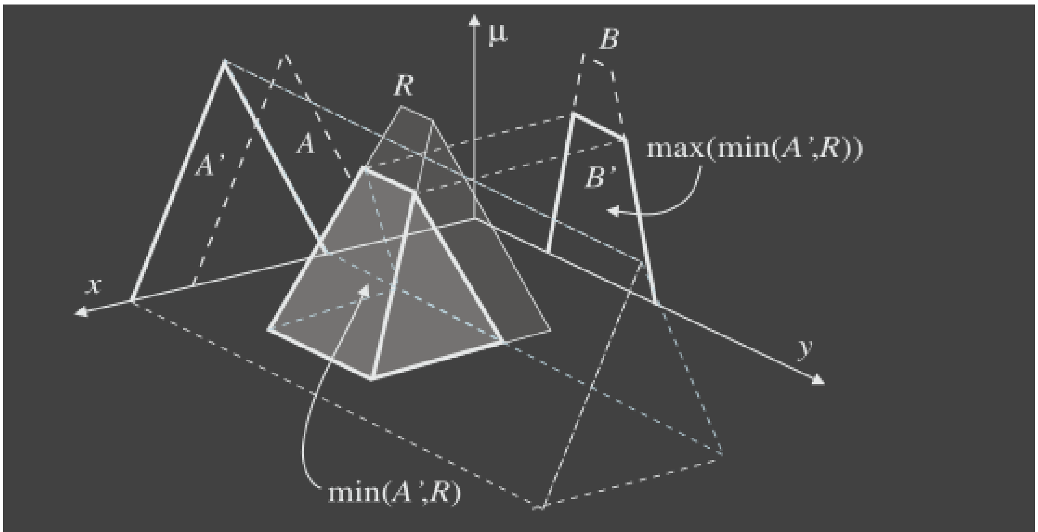
Inference with One Rule

1. construct implication relation

$$\mu_R(x,y) = \mathrm{I}\left(\mu_A(x), \mu_B(y)\right)$$

2. use relational composition to derive B' from A'

$$B' = A' \circ R$$



(b) Fuzzy inference.

For the minimum t-norm

$$\mu_{B'}(\mathbf{y}) = \max_X\left(\min\left(\mu_{A'}(\mathbf{x}), \mu_R(\mathbf{x},\mathbf{y})\right)\right)$$

Inference with Several Rules

1. Construct implication relation for each rule i

$$\mu_{R_i}(x,y) = \mathrm{I}\left(\mu_{A_i}(x), \mu_{B_i}(y)\right)$$

2. Aggregate relations R_i into one

$$\mu_R(x,y) = \operatorname{aggr}\left(\mu_{R_i}(x,y)\right)$$

The aggr operator is the minimum for implications and the maximum for conjuncti

3. Use relational composition to derive B' from A'

$$B' = A' \circ R$$

<u>If R is Fuzzy implication</u>

$$R = \bigcap_{i=1}^{K} R_i, \quad \text{that is,} \quad \mu_R(\mathbf{x}, \mathbf{y}) = \min_{1 \le i \le K} \mu_{R_i}(\mathbf{x}, \mathbf{y})$$

<u>If R is T-norm</u>

$$R = \bigcup_{i=1}^{K} R_i, \text{ that is, } \mu_R(\mathbf{x}, \mathbf{y}) = \max_{1 \le i \le K} \mu_{R_i}(\mathbf{x}, \mathbf{y})$$

Example (3.3 in Lecture Notes)

T-norm VS Fuzzy Implication

Since the input fuzzy set A′ is different from the antecedent set A, the derived conclusion B′ is in both cases "**less certain**" than B.

For **<u>Fuzzy Implication</u>**: uncertainty is reflected in the **increased membership** values for the domain elements that have low or zero membership in B

For **<u>T-norm</u>**: results in **decreasing the membership degree** of the elements that have high membership in B

A simpler way: Mamdani (Max-min) Inference

For the t-norm, we have a simpler way

$$\mu_{B'}(\mathbf{y}) = \max_{1 \le i \le K} [\beta_i \wedge \mu_{B_i}(\mathbf{y})], \quad \mathbf{y} \in Y$$

$$\beta_i = \max_{X} [\mu_{A'}(\mathbf{x}) \wedge \mu_{A_i}(\mathbf{x})]$$

1. Compute the degree of fulfillment for each rule by $\beta_i = \max_X [\mu_{A'}(\mathbf{x}) \wedge \mu_{A_i}(\mathbf{x})]$, Note that for a singleton set ($\mu_{A'}(\mathbf{x}) = 1$ for $\mathbf{x} = \mathbf{x}_0$ and $\mu_{A'}(\mathbf{x}) = 0$ otherwise) the equation β_i simplifies to $\beta_i = \mu_{A_i}(x_0)$

2. Derive the output fuzzy sets
   $B_i' : \mu_{B_i'}(\mathbf{y}) = \beta_i \wedge \mu_{B_i}(\mathbf{y}), \quad \mathbf{y} \in Y, \quad 1 \le i \le K$
3. **Aggregate** output fuzzy sets of all the rules into one fuzzy set.
   $B_i' : \mu_{B'}(\mathbf{y}) = \max_{1 \le i \le K} \mu_{B_i'}(\mathbf{y}), \quad \mathbf{y} \in Y$

It can be seen as :

1. first finding the highest point (maximum fulfilment)
2. Then all inference result should less than the result of maximum fulfilment and at the same time meet the property of initial result.

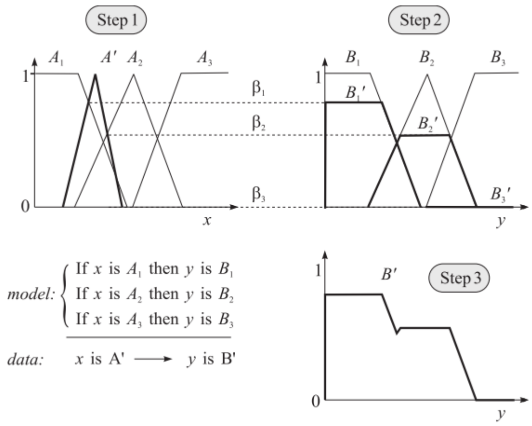


Figure 3.8.: A schematic representation of the Mamdani inference algorithm.

3. Singleton and Takagi-Sugeno Fuzzy System

Singleton Fuzzy model

$$\text{If } x \text{ is } A_i \text{ then } y = b_i$$

- Defuzzication/Infererence:

$$\begin{aligned} y &= \frac{\sum_{i=1}^{K} \mu_{A_i}(x) b_i}{\sum_{i=1}^{K} \mu_{A_i}(x)} \\ &= \Sigma_{i=1}^{k} \phi_i(x) b_i \end{aligned}$$

  - the basis functions $\phi_i(x)$ are given by the (normalized) degrees of fulfillment of the rule antecendents
  - the constants b_i are the consequents

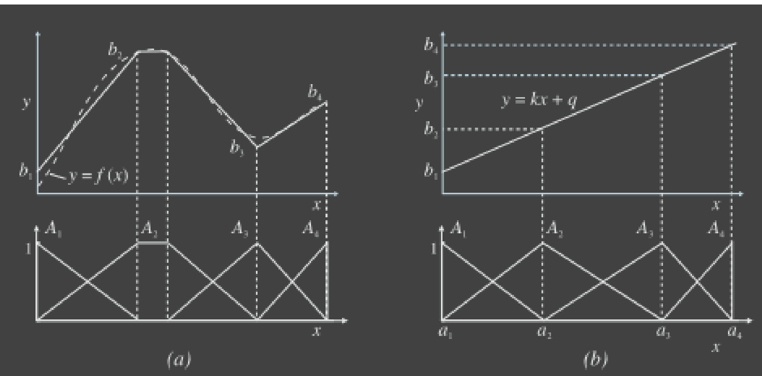


Figure 3.14.: Singleton model with triangular or trapezoidal membership functions results in a piecewise linear input-output mapping (a), of which a linear mapping is a special case (b).

a singleton model can also **represent any given linear mapping of the form**:

$$y = \mathbf{k}^T \mathbf{x} + q = \sum_{i=1}^{p} k_i x_i + q.$$

Takagi-Sugeno (TS) Fuzzy model

If x is A_i then $y_i = a_i x + b_i$

$$y = \frac{\sum_{i=1}^{K} \mu_{A_i}(x) y_i}{\sum_{i=1}^{K} \mu_{A_i}(x)} = \frac{\sum_{i=1}^{K} \mu_{A_i}(x)\,(a_i x + b_i)}{\sum_{i=1}^{K} \mu_{A_i}(x)}$$

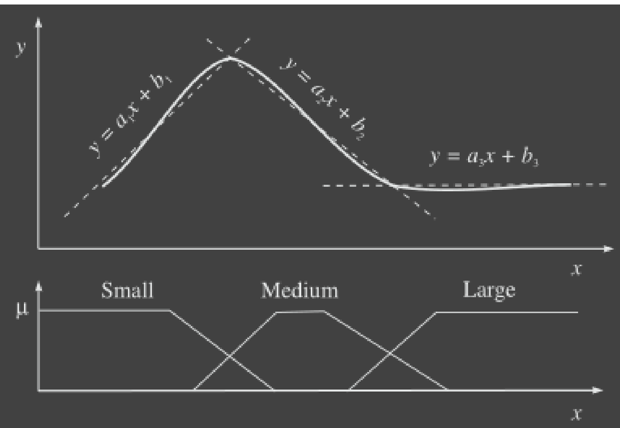


Figure 3.17.: Takagi–Sugeno fuzzy model as a smoothed piece-wise linear approximation of a nonlinear function.

- Quasi-Linear Property

$$y = \left(\sum_{i=1}^{K} \gamma_i(\mathbf{x}) \mathbf{a}_i^T \right) \mathbf{x} + \sum_{i=1}^{K} \gamma_i(\mathbf{x}) b_i = \mathbf{a}^T(\mathbf{x}) \mathbf{x} + b(\mathbf{x})$$

$$\gamma_i(\mathbf{x}) = \frac{\beta_i(\mathbf{x})}{\sum_{j=1}^{K} \beta_i(\mathbf{x})}$$

$$\mathbf{a}(\mathbf{x}) = \sum_{i=1}^{K} \gamma_i(\mathbf{x})\mathbf{a}_i, \quad b(\mathbf{x}) = \sum_{i=1}^{K} \gamma_i(\mathbf{x}) b_i$$

The ‘parameters’ $a(x), b(x)$ are convex linear combinations of the consequent parameters a_i and b_i

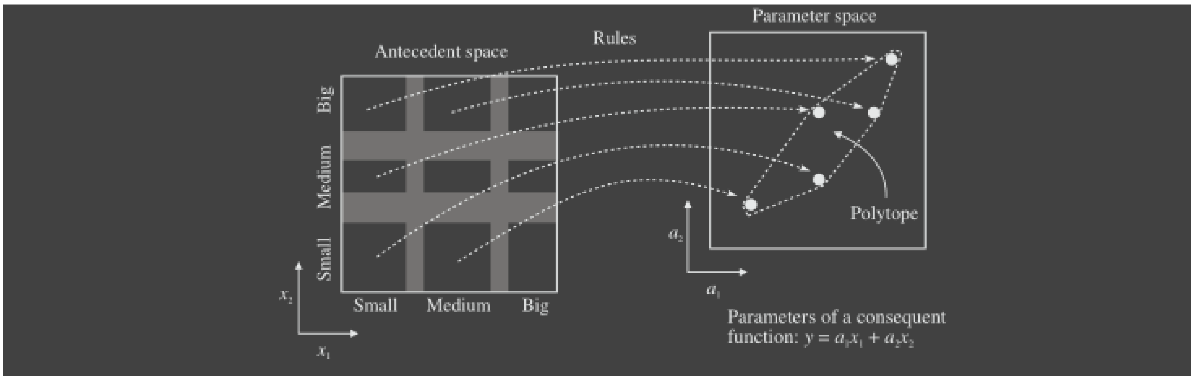


Figure 3.18.: A TS model with affine consequents can be regarded as a mapping from the antecedent space to the space of the consequent parameters.

Inference in the TS Model

$$y = \frac{\sum_{i=1}^{K} \beta_i y_i}{\sum_{i=1}^{K} \beta_i} = \frac{\sum_{i=1}^{K} \beta_i \left(\mathbf{a}_i^T \mathbf{x} + b_i\right)}{\sum_{i=1}^{K} \beta_i}$$

4. Defuzzification

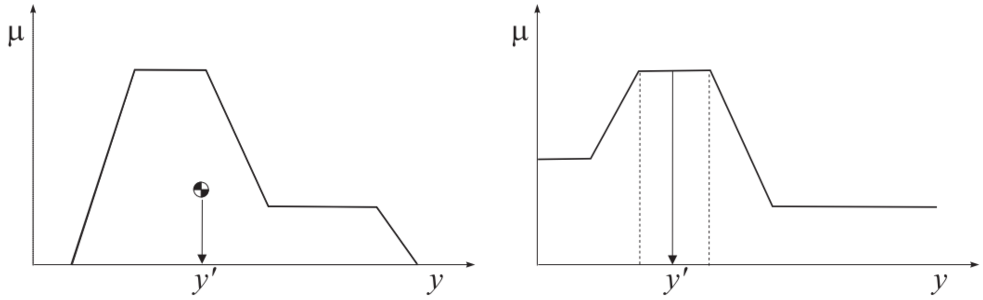
Defuzzification convert a fuzzy set to a crisp value

- Center of Gravity Method (COG):

$$y' = \frac{\sum_{j=1}^{F} \mu_{B'}(y_j) y_j}{\sum_{j=1}^{F} \mu_{B'}(y_j)}$$

- Mean of Maxima (MOM):

$$\operatorname{mom}(B') = \operatorname{cog}\left\{ y \mid \mu_{B'}(y) = \max_{y \in Y} \mu_{B'}(y) \right\}$$



(a) center of gravity (b) mean of maxima

For **Mamdani max-min inference**, we always use COG

- Mamdani: only use min, no implication, so no interpolate itself. By using COG, consider each part weight, not only max

For **Fuzzy Implication**, we always use MOM

- The COG method cannot be directly used in this case, because the **uncertainty in the output** results in an increase of the membership degrees,

<u>**Fuzzy-mean defuzzification**</u>

- consequent fuzzy sets are first defuzzified, for example, use MOM $b_j = \operatorname{mom}(B_j)$
- Then use COG:

$$y' = \frac{\sum_{j=1}^{M} \omega_j b_j}{\sum_{j=1}^{M} \omega_j}$$
$$\omega_j = \mu_{B'}(b_j)$$

5. Clustering

Hard Partitions

Problem Model

$$\begin{aligned} &\bigcup_{i=1}^{c} A_i = \mathbf{Z} \\ &A_i \cap A_j = \emptyset, \quad 1 \leq i \neq j \leq c \\ &\emptyset \subset A_i \subset \mathbf{Z}, \quad 1 \leq i \leq c \end{aligned}$$

Optimization Approach Model

$$
\begin{array}{ll}
\mu_{ik} \in \{0,1\}, & 1 \leq i \leq c, \quad 1 \leq k \leq N, \\
\sum_{i=1}^{c} \mu_{ik} = 1, & 1 \leq k \leq N \\
0 < \sum_{k=1}^{N} \mu_{ik} < N, & 1 \leq i \leq c
\end{array}
$$

Hard Partitioning Space

$$
M_{hc} = \left\{ \mathbf{U} \in \mathbb{R}^{c \times N} \mid \mu_{ik} \in \{0,1\}, \forall i, k; \sum_{i=1}^{c} \mu_{ik} = 1, \forall k; 0 < \sum_{k=1}^{N} \mu_{ik} < N, \forall i \right\}
$$

Shortcoming

Boundary data points may represent patterns with a mixture of properties of data in A_1 and A_2, and therefore cannot be fully assigned to either of these classes, or do they constitute a separate class.

Fuzzy Partitions

More information is preserved in this form of clustering, which is exactly the advantage of fuzzy clustering over hard partitioning.

Problem Model

Given:

$$
z_k = [z_{1k}, z_{2k}, \ldots, z_{nk}]^T \in \mathbb{R}^n, \quad k = 1, \ldots, N
$$

Find:

- the fuzzy partition matrix

$$
\boldsymbol{U} = \begin{bmatrix} \mu_{11} & \ldots & \mu_{1k} & \ldots & \mu_{1N} \\ \vdots & \ldots & \vdots & \ldots & \vdots \\ \mu_{c1} & \ldots & \mu_{ck} & \ldots & \mu_{cN} \end{bmatrix}
$$

- the cluster centers

$$
\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_c\}, \quad \boldsymbol{v}_i \in \mathbb{R}^n
$$

An optimization Model

$$J(Z;V,U,A)=\sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{i,j}^{m}d_{A_i}^{2}\left(z_j,v_i\right)$$

$$\text{s.t.}$$

$$0\leq\mu_{i,j}\leq 1,\quad i=1,\ldots,c,j=1,\ldots,N\quad\text{(membership degree)}$$

$$0<\sum_{j=1}^{N}\mu_{i,j}<N,\quad i=1,\ldots,c\quad\text{(no cluster empty)}$$

$$\sum_{i=1}^{c}\mu_{i,j}=1,\quad j=1,\ldots,N\ \text{(total membership)}$$

Partitioning Space

$$M_{hc}=\left\{\mathbf{U}\in\mathbb{R}^{c\times N}\mid\mu_{ik}\in[0,1],\forall i,k;\sum_{i=1}^{c}\mu_{ik}=1,\forall k;0<\sum_{k=1}^{N}\mu_{ik}<N,\forall i\right\}.$$

Distance Matrics

- **Euclidean norm**

$$d^2\left(z_j,v_i\right)=\left(z_j-v_i\right)^T\left(z_j-v_i\right)$$

- **Inner-product norm**

$$d_{A_i}^2\left(z_j,v_i\right)=\left(z_j-v_i\right)^T A_i\left(z_j-v_i\right)$$

Possibilistic Partition

Optimization Approach Model

$$\begin{aligned}\mu_{ik}\in[0,1],&\qquad 1\leq i\leq c,\quad 1\leq k\leq N,\\ \exists i,\mu_{ik}>0,&\qquad\qquad\qquad\qquad\forall k,\\ 0<\sum_{k=1}^{N}\mu_{ik}<N,&\qquad\qquad\qquad 1\leq i\leq c\end{aligned}$$

The difference between possibilistic partition and fuzzy partition is the second constraint.

possibilistic partitioning space

$$M_{pc}=\left\{\mathrm{U}\in\mathbb{R}^{c\times N}\mid\mu_{ik}\in[0,1],\forall i,k;\forall k,\exists i,\mu_{ik}>0;0<\sum_{k=1}^{N}\mu_{ik}<N,\forall i\right\}.$$

Fuzzy c-Means Clustering

Functional

$$\bar{J}(\mathbf{Z};\mathbf{U},\mathbf{V},\boldsymbol{\lambda})=\sum_{i=1}^{G}\sum_{k=1}^{\mathbb{N}}(\mu_{ik})^{m}D_{ik\mathbf{A}}^{2}+\sum_{k=1}^{N}\lambda_k\left[\sum_{i=1}^{c}\mu_{ik}-1\right],$$

data matrix Z contains all the samples (column)

fuzzy partition matrix U

prototype matrix V

μ_{ik}s the membership degrees of sample k to cluster is

λ_k ensures that the constraint

D^2_{ikA} is the squared inner-product distance norm

Algorithms

Algorithm 4.1 Fuzzy c-means (FCM).

Given the data set $\mathbf{Z}$, choose the number of clusters $1 < c < N$, the weighting exponent $m > 1$, the termination tolerance $\epsilon > 0$ and the norm-inducing matrix $\mathbf{A}$. Initialize the partition matrix randomly, such that $\mathbf{U}^{(0)} \in M_{fc}$.

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

Step 1: Compute the cluster prototypes (means):

$$\mathbf{v}_i^{(l)} = \frac{\sum_{k=1}^{N} \left(\mu_{ik}^{(l-1)}\right)^m \mathbf{z}_k}{\sum_{k=1}^{N} \left(\mu_{ik}^{(l-1)}\right)^m}, \quad 1 \le i \le c\,.$$

Step 2: Compute the distances:

$$D_{ik\mathbf{A}}^2 = (\mathbf{z}_k - \mathbf{v}_i^{(l)})^T \mathbf{A} (\mathbf{z}_k - \mathbf{v}_i^{(l)}), \quad 1 \le i \le c, \quad 1 \le k \le N\,.$$

Step 3: Update the partition matrix:
for $1 \le k \le N$
if $D_{ik\mathbf{A}} > 0$ for all $i = 1, 2, \ldots, c$

$$\mu_{ik}^{(l)} = \frac{1}{\sum_{j=1}^{c} (D_{ik\mathbf{A}}/D_{jk\mathbf{A}})^{2/(m-1)}},$$

otherwise

$$\mu_{ik}^{(l)} \begin{cases} = 0 & \text{if } D_{ik\mathbf{A}} > 0 \\ \in [0, 1] & \text{if } D_{ik\mathbf{A}} = 0 \end{cases} \quad \text{with} \quad \sum_{i=1}^{c} \mu_{ik}^{(l)} = 1\,.$$

until $\|\mathbf{U}^{(l)} - \mathbf{U}^{(l-1)}\| < \epsilon$.

- The FCM algorithm **converges to a local minimum** of the c-means functional. Hence, **different initializations may lead to different results.**
- Alternatively, the algorithm can be initialized with $V^{(0)}$, loop through $V^{(l-1)} \rightarrow U^{(l)} \rightarrow V^{(l)}$, and terminate on $||V^{(l)} - V^{(l-1)}|| < \epsilon$

Parameters Analysis

Number of Clusters c

The number of clusters c is the most important parameter. So, how to choose c?

Validity measures

Validity measures are scalar indices that assess the goodness of the obtained partition. It use measures to **quantify the separation and the compactness of the clusters.**

Xie-Beni index: minimize:

$$\chi(\mathbf{Z};\mathbf{U},\mathbf{V})=\frac{\sum_{i=1}^{c}\sum_{k=1}^{N}\mu_{ik}^{m}\|\mathbf{z}_k-\mathbf{v}_i\|^2}{c\cdot\min_{i\neq j}\left(\|\mathbf{v}_i-\mathbf{v}_j\|^2\right)\mathbf{V}}$$

- numerator: within-group variance
- denominator: separation of the cluster centers

Iterative merging or insertion of clusters

start with a sufficiently large number of clusters

- merging clusters that are similar (compatible)

start with a small number of clusters

- iteratively insert clusters in the regions where the data points have low degree of membership in the existing clusters

Fuzziness Parameters (m and v)

As m approaches one from above, the partition becomes **hard** ($\mu_{ik}\in\{0,1\}$) and v_i are ordinary means of the clusters.

m influnces the fuzziness of the resulting partition. As $m\to\infty$ the partition become completely fuzzy ($\mu_{ik}=\frac{1}{c}$) and the cluster means are all equal to the mean of Z

Termination Criterion （ϵ）

the usual choice is $\epsilon=0.001$, even though $\epsilon=0.01$ works well

Norm-Inducing Matrix (A)

changing the **measure of dissimilarity.**

A common choice is $A=I$, which generates

$$D_{ik}^2=(\mathbf{z}_k-\mathbf{v}_i)^T(\mathbf{z}_k-\mathbf{v}_i)$$

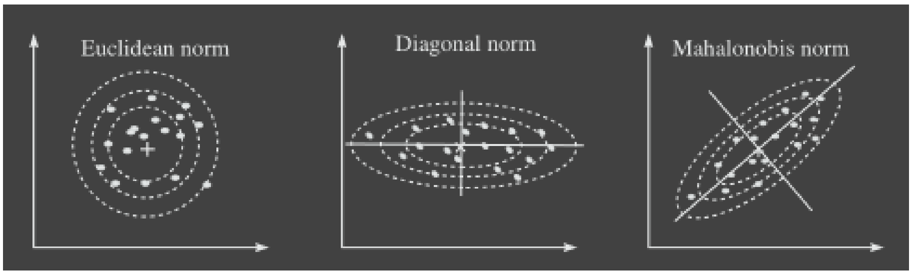
Another choice is A is a a **diagonal matrix** that accounts for **different variances in the directions** of the coordinate axes of Z:

$$\mathbf{A} = \begin{bmatrix} (1/\sigma_1)^2 & 0 & \cdots & 0 \\ 0 & (1/\sigma_2)^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & (1/\sigma_n)^2 \end{bmatrix}$$

Another choice is the inverse of the covariance matrix of Z: $A = R^{-1}$ (Mahalanobis norm)

$$\mathbf{R} = \frac{1}{N}\sum_{k=1}^{N}(\mathbf{z}_k - \bar{\mathbf{z}})(\mathbf{z}_k - \bar{\mathbf{z}})^T$$

The norm influences the clustering criterion by changing the measure of dissimilarity.



ShortComes

- A common limitation of clustering algorithms based on a fixed distance norm is that such a norm **forces the objective function to prefer clusters of a certain shape** even if they are not present in the data.
- inclusion of pre-defined volume per cluster (ρ_i)

Extensions of the Fuzzy c-Means Algorithm

There are several well-known extensions of the basic c-means algorithm:

- Algorithms using an adaptive distance measure, such as the Gustafson–Kessel algorithm (Gustafson and Kessel, 1979) and the fuzzy maximum likelihood estimation algorithm (Gath and Geva, 1989).
- Algorithms based on hyperplanar or functional prototypes, or prototypes defined by functions. They include the fuzzy c-varieties (Bezdek, 1981), fuzzy c-elliptotypes (Bezdek et al., 1981), and fuzzy regression models (Hathaway and Bezdek, 1993).
- Algorithms that search for possibilistic partitions in the data, i.e., partitions where the constraint (4.4b) is relaxed.