Script **generated by TTT**

Title: profile1 (18.06.2013)

Date: Tue Jun 18 11:59:08 CEST 2013

Duration: 90:22 min

Pages: 65

Data Mining:
Metric Clustering

Finding Clusters in Profiles

Examples for profile elements that can be embedded in metric spaces:

- Location & Velocity: Metric space: ($\mathbb{R}^3$, || . ||)
- Text describing Interests: Metric space: ($\mathbb{R}^{|Voc|}$, || . ||) where Voc denotes the Vocabulary of the text.

“I like to dance samba, bake pizza, watch tv and plant trees in the garden. I also like to bake cakes.” ⇒

I	2
like	2
to	2
dance	1
samba	1
bake	2
pizza	1
watch	1
tv	1
and	1
plant	1
trees	1
in	1
the	1
garden	1
also	1
cakes	1

*Often: Instead of term-frequency (tf) alone: use term-frequency * inverse document frequency (idf); idf = log (#of docs where t occurs / #of docs)*

Finding Clusters in Profiles

- How do we compute clusters in metric spaces?
- Group models: How do we compute socially meaningful clusters in metric spaces (and thus avoid quasi-groups)?
- First some notations / basics:
  - In graph clustering we had: A graph clustering **C**={C_1, C_2, ..., C_K} is a partion of V into non-empty subsets C_k
  - Now: clustering $\mathscr{C} : \mathcal{X} \to \mathcal{I}$: mapping of a metric value space $\mathcal{X}$ to a set of cluster indices $\mathcal{I}$
  - Clusterings can be:
    - exclusive or non-exclusive
    - crisp or fuzzy
    - hierarchical or non-hierarchical

Finding Clusters in Profiles

- How do we compute clusters in metric spaces?
- Group models: How do we compute socially meaningful clusters in metric spaces (and thus avoid quasi-groups)?
- First some notations / basics:
  - In graph clustering we had: A graph clustering **C**={C_1, C_2, ..., C_K} is a partion of V into non-empty subsets C_k
  - Now: clustering $\mathscr{C} : \mathcal{X} \rightarrow \mathcal{I}$: mapping of a metric value space X to a set of cluster indices I
  - Clusterings can be:
    - exclusive or non-exclusive
    - crisp or fuzzy
    - hierarchical or non-hierarchical

Finding Clusters in Profiles

- Exclusive → non overlapping clusters; non-exclusive → overlapping clusters
- Hierarchical clustering → imposes a tree structure (Dendrogram) on the C_k where an edge C_i → C'_j implies C_i ⊂ C'_j;
- Crisp clusterings: Conventional characteristic functions α_k for each Cluster C_k

$$\alpha_k : \mathcal{X} \rightarrow \{0,1\} \text{ with } \alpha_k(x \in \mathcal{X}) = \begin{cases} 1 & x \in \mathcal{C}_k \\ 0 & x \notin \mathcal{C}_k \end{cases}$$

- Fuzzy clusterings: fuzzy membership function α _k for each Cluster C_k

$$\alpha_k : \mathcal{X} \rightarrow [0,1]$$

Finding Clusters in Profiles

- Exclusive → non overlapping clusters; non-exclusive → overlapping clusters
- Hierarchical clustering → imposes a tree structure (Dendrogram) on the C_k where an edge C_i → C'_j implies C_i ⊂ C'_j;
- Crisp clusterings: Conventional characteristic functions α_k for each Cluster C_k

$$\alpha_k : \mathcal{X} \rightarrow \{0,1\} \text{ with } \alpha_k(x \in \mathcal{X}) = \begin{cases} 1 & x \in \mathcal{C}_k \\ 0 & x \notin \mathcal{C}_k \end{cases}$$

- Fuzzy clusterings: fuzzy membership function α _k for each Cluster C_k

$$\alpha_k : \mathcal{X} \rightarrow [0,1]$$

Metric variant of Single / Complete link clustering

- Metric variant of Single / Complete link clustering: Hierarchical, crisp, non-overlapping
- Completely analogous to graph clustering case: Start with singletons and on each level of the dendrogram merge two clusters with minimal distance (cost)
  - Single link:

$$d(\mathcal{C}_{k_1}, \mathcal{C}_{k_2}) = \min_{\{n_1,n_2 | x_{n_1} \in \mathcal{C}_{k_1} \wedge x_{n_2} \in \mathcal{C}_{k_2}\}} ||x_{n_1} - x_{n_2}||$$

  - Complete link:

$$d(\mathcal{C}_{k_1}, \mathcal{C}_{k_2}) = \max_{\{n_1,n_2 | x_{n_1} \in \mathcal{C}_{k_1} \wedge x_{n_2} \in \mathcal{C}_{k_2}\}} ||x_{n_1} - x_{n_2}||$$

Metric variant of Single / Complete link clustering

- Metric variant of Single / Complete link clustering: Hierarchical, crisp, non-overlapping
- Completely analogous to graph clustering case: Start with singletons and on each level of the dendrogram merge two clusters with minimal distance (cost)
  - Single link:

$$d(\mathcal{C}_{k_1},\mathcal{C}_{k_2}) = \min_{\{n_1,n_2|x_{n_1}\in\mathcal{C}_{k_1}\wedge x_{n_2}\in\mathcal{C}_{k_2}\}} ||x_{n_1}-x_{n_2}||$$

  - Complete link:

$$d(\mathcal{C}_{k_1},\mathcal{C}_{k_2}) = \max_{\{n_1,n_2|x_{n_1}\in\mathcal{C}_{k_1}\wedge x_{n_2}\in\mathcal{C}_{k_2}\}} ||x_{n_1}-x_{n_2}||$$

K-Means Clustering

- General idea (also valid in graph clustering): Optimize objective function that formalizes clustering paradigm.
- K-Means: Optimize intra cluster coherence:
  - Describe cluster C_k by prototype μ_k; prototype need not be an actual pattern (If so, algorithm works with slight modifications as well)
  - Determine cluster for each pattern x_n by nearest neighbour rule:

$$\mathscr{C}(x_n) = k_a \leftrightarrow ||x_n-\mu_{k_a}|| = \min_i ||x_n-\mu_k||$$

K-Means Clustering

- K-Means: Optimize intra cluster coherence:
  - Find prototypes by optimizing objective function modeling intra cluster coherence as mean square error

$$J_{\mathrm{SQE}} = \sum_{k=1}^{K}\sum_{\{n|x_n\in\mathcal{C}_k\}} ||x_n-\mu_k||^2$$

$$\frac{\mathrm{d}J_{\mathrm{SQE}}}{\mathrm{d}\mu_k} \overset{!}{=} 0 \quad\Longrightarrow\quad \mu^k = \frac{1}{|\mathcal{C}_k|}\sum_{\{n|x_n\in\mathcal{C}_k\}} x_n$$

  - → cluster prototypes are barycenters („centers of gravity“) of their clusters.

K-Means Clustering

- K-Means: Optimize intra cluster coherence:
  - Find prototypes by optimizing objective function modeling intra cluster coherence as mean square error

$$J_{\mathrm{SQE}} = \sum_{k=1}^{K}\sum_{\{n|x_n\in\mathcal{C}_k\}} ||x_n-\mu_k||^2$$

$$\frac{\mathrm{d}J_{\mathrm{SQE}}}{\mathrm{d}\mu_k} \overset{!}{=} 0 \quad\Longrightarrow\quad \mu^k = \frac{1}{|\mathcal{C}_k|}\sum_{\{n|x_n\in\mathcal{C}_k\}} x_n$$

  - → cluster prototypes are barycenters („centers of gravity“) of their clusters.

K-Means Clustering

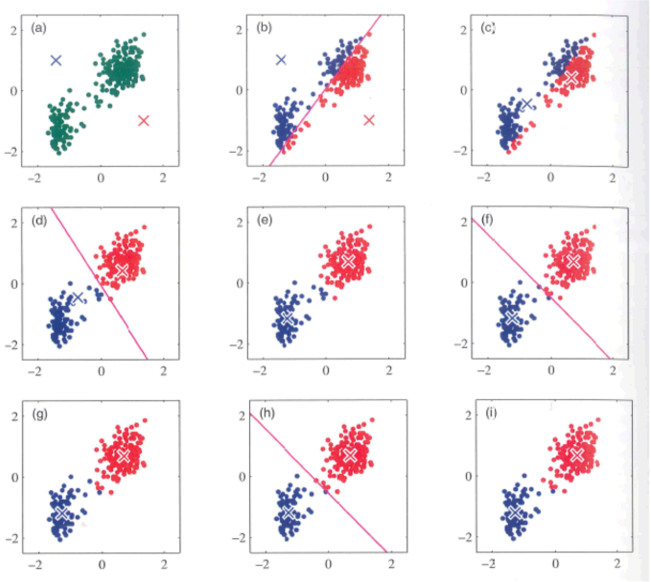
- K-Means: Optimize intra cluster coherence:
  - Find prototypes by optimizing objective function modeling intra cluster coherence as mean square error

$$J_{\mathrm{SQE}} = \sum_{k=1}^{K} \sum_{\{n|x_n\in\mathcal{C}_k\}} ||x_n-\mu_k||^2$$

$$\frac{\mathrm{d}J_{\mathrm{SQE}}}{\mathrm{d}\mu_k} \overset{!}{=} 0 \quad \Longrightarrow \quad \mu^k = \frac{1}{|\mathcal{C}_k|}\sum_{\{n|x_n\in\mathcal{C}_k\}} x_n$$

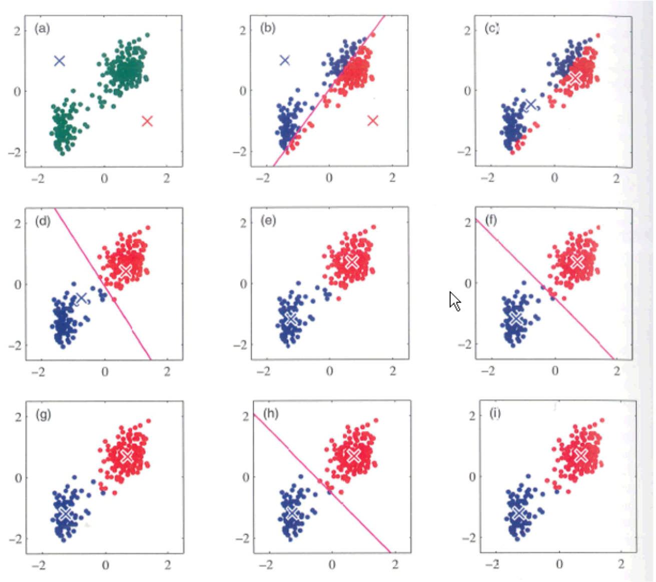
  - → cluster prototypes are barycenters („centers of gravity“) of their clusters.

K-Means Clustering



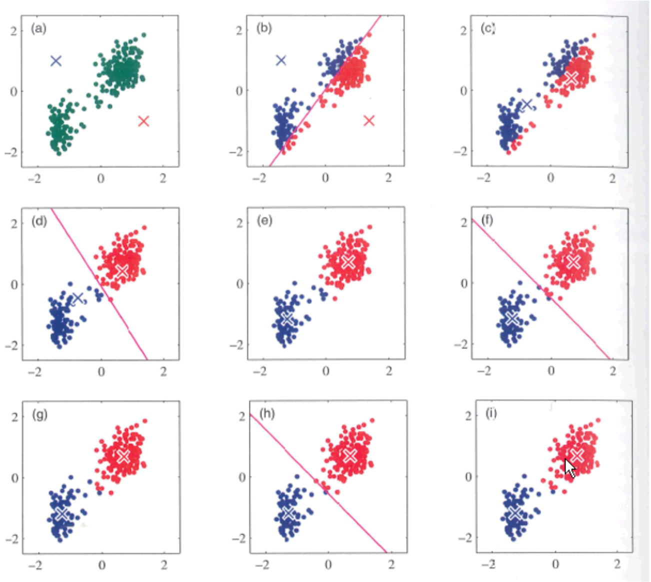
[3]

K-Means Clustering



[3]

K-Means Clustering



[3]

K-Means Clustering

- Dunn Index:

$$D = \min_{k_1 \in [1,K]} \left(\min_{k_2 \in [1,K]} \left(\frac{d_1(\mathcal{C}_{k_1}, \mathcal{C}_{k_2})}{\max_{k_3 \in [1,K]} d_2(\mathcal{C}_{k_3})} \right) \right)$$

where $d_1(\mathcal{C}_{k_1}, \mathcal{C}_{k_2})$ is the distance function between two clusters defined by

$$d_1(\mathcal{C}_{k_1}, \mathcal{C}_{k_2}) = \min_{\{(n_1,n_2) | x_{n_1} \in \mathcal{C}_{k_1} \wedge x_{n_2} \in \mathcal{C}_{k_2}\}} ||x_{n_1} - x_{n_2}||$$

(that is the single link distance from SAHN).
The "diameter" d_2 of the clusters is defined by

$$d_2(\mathcal{C}_i) = \max_{\{(n_1,n_2) | x_{n_1} \in \mathcal{C}_i \wedge x_{n_2} \in \mathcal{C}_i\}} ||x_{n_1} - x_{n_2}||$$

[7]

K-Means Clustering

- Dunn Index:

$$D = \min_{k_1 \in [1,K]} \left(\min_{k_2 \in [1,K]} \left(\frac{d_1(\mathcal{C}_{k_1}, \mathcal{C}_{k_2})}{\max_{k_3 \in [1,K]} d_2(\mathcal{C}_{k_3})} \right) \right)$$

where $d_1(\mathcal{C}_{k_1}, \mathcal{C}_{k_2})$ is the distance function between two clusters defined by

$$d_1(\mathcal{C}_{k_1}, \mathcal{C}_{k_2}) = \min_{\{(n_1,n_2) | x_{n_1} \in \mathcal{C}_{k_1} \wedge x_{n_2} \in \mathcal{C}_{k_2}\}} ||x_{n_1} - x_{n_2}||$$

(that is the single link distance from SAHN).
The "diameter" d_2 of the clusters is defined by

$$d_2(\mathcal{C}_i) = \max_{\{(n_1,n_2) | x_{n_1} \in \mathcal{C}_i \wedge x_{n_2} \in \mathcal{C}_i\}} ||x_{n_1} - x_{n_2}||$$

[7]

K-Means Clustering

- Dunn Index:

$$D = \min_{k_1 \in [1,K]} \left(\min_{k_2 \in [1,K]} \left(\frac{d_1(\mathcal{C}_{k_1}, \mathcal{C}_{k_2})}{\max_{k_3 \in [1,K]} d_2(\mathcal{C}_{k_3})} \right) \right)$$

where $d_1(\mathcal{C}_{k_1}, \mathcal{C}_{k_2})$ is the distance function between two clusters defined by

$$d_1(\mathcal{C}_{k_1}, \mathcal{C}_{k_2}) = \min_{\{(n_1,n_2) | x_{n_1} \in \mathcal{C}_{k_1} \wedge x_{n_2} \in \mathcal{C}_{k_2}\}} ||x_{n_1} - x_{n_2}||$$

(that is the single link distance from SAHN).
The "diameter" d_2 of the clusters is defined by

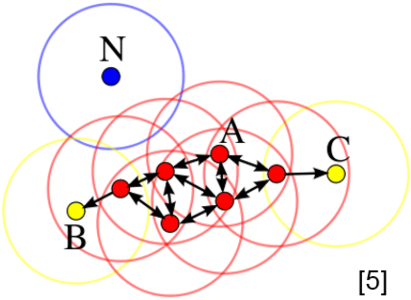
$$d_2(\mathcal{C}_i) = \max_{\{(n_1,n_2) | x_{n_1} \in \mathcal{C}_i \wedge x_{n_2} \in \mathcal{C}_i\}} ||x_{n_1} - x_{n_2}||$$

[7]

DBSCAN

- K-Means is „OK“ as cluster algorithm, but has certain disadvantages:
  - favors spherical clusters
  - need to know K
  - no notion of noise
- Alternative → DBSCAN [4]
  (used frequently in practice):
  - Idea: Two parameters: minPt, ε
  - Rough idea: iterate:

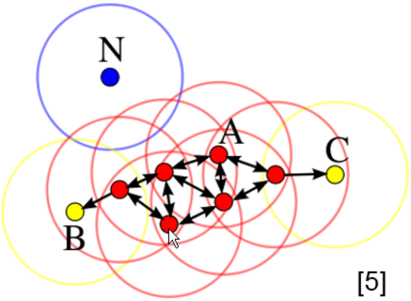
```
visit previously unseen pattern x:
  if  in ε-neighborhood {x‘} of x: |{x‘}|≥ minPt then
        start new cluster: include x and {x‘} and those of their
        ε-neighborhoods {x‘‘} that are dense enough (|{x‘‘}|≥
         minPt), etc.
  else: x is noise
```



[5]

DBSCAN

- K-Means is „OK“ as cluster algorithm, but has certain disadvantages:
  - favors spherical clusters
  - need to know K
  - no notion of noise



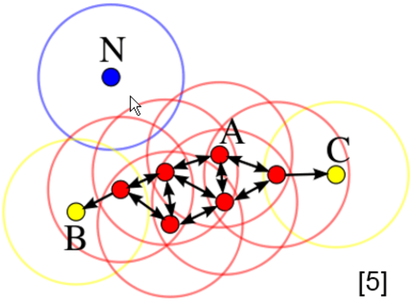
[5]

- Alternative → DBSCAN [4] (used frequently in practice):
  - Idea: Two parameters: minPt, ε
  - Rough idea: iterate:

```
visit previously unseen pattern x:
    if  in ε-neighborhood {x‘} of x: |{x‘}|≥ minPt then
            start new cluster: include x and {x‘} and those of their
            ε-neighborhoods {x‘‘} that are dense enough (|{x‘‘}|≥
             minPt), etc.
    else: x is noise
```

DBSCAN

- K-Means is „OK“ as cluster algorithm, but has certain disadvantages:
  - favors spherical clusters
  - need to know K
  - no notion of noise



[5]

- Alternative → DBSCAN [4] (used frequently in practice):
  - Idea: Two parameters: minPt, ε
  - Rough idea: iterate:

```
visit previously unseen pattern x:
    if  in ε-neighborhood {x‘} of x: |{x‘}|≥ minPt then
            start new cluster: include x and {x‘} and those of their
            ε-neighborhoods {x‘‘} that are dense enough (|{x‘‘}|≥
             minPt), etc.
    else: x is noise
```

DBSCAN

- Advantages of DBSCAN:
  - We do not need to know K in advance
  - arbitrarily shaped clusters
  - notion of noise
- Disadvantages:
  - instead of having to know K, we need to „guess“ minPt and ε instead (can be a problem for high dimensional pattern spaces (→ curse of dimensionality))
  - original DBSCAN has fixed (minPt, ε) → problems when cluster density varies

K-Means Clustering

- Interesting aspect: How do we determine correct number k of clusters? (Same problem with graph clustering: where to cut dendrogram?)
- Answer: Compute for every k clusterings; chose the best clustering with a cluster quality measure
- Cluster quality measures for metric case (countless variants exist in literature; for an overview: e.g. [2]) (Objective functions modeling clustering paradigm):
  - Dunn-Index
  - Entropy based indices
  - ....

Fuzzy C-Means Clustering

- K-Means was a crisp algorithm. Now: fuzzy variant
- Reformulate K-Means objective function with membership matrix

r_{nk}: Membership of pattern x_n in class C_k

$$J_{SQE} = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}||x_n - \mu_k||^2$$

- optimization criterion

$$\mathrm{d}J_{SQE}/\mathrm{d}\mu_k = 0$$

- together with non-overlaping constraint

$$\forall n(\exists k(r_{nk} = 1) \wedge ((k' \neq k) \rightarrow (r_{nk'} = 0)))$$

leads to well known K-Means

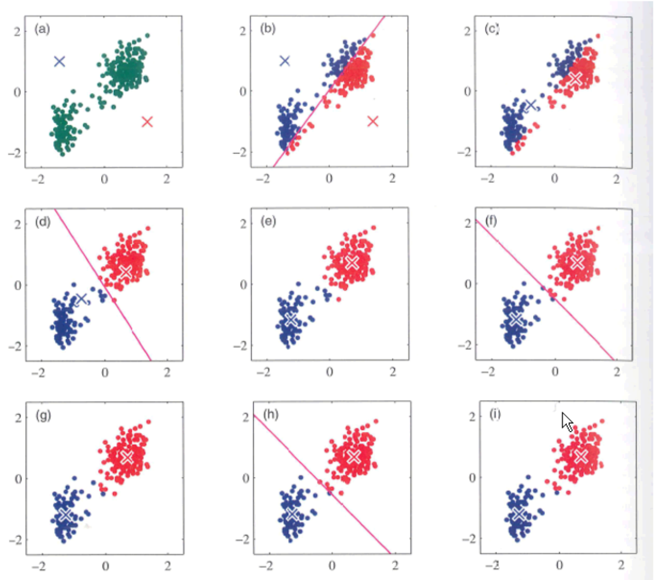
$$\mu_k = \sum_{n=1}^{N} r_{nk}x_n / \sum_{n=1}^{N} r_{nk} = (1/|\mathcal{C}_k|)\sum_{n|x_n\in\mathcal{C}_k} x_n$$

K-Means Clustering

Example Application: Clustering locations

- Problem: How do we distinguish socially relevant clusters (candidates for groups) from quasi groups?
  - Compute clusterings over period of time: Good candidates: clusters that appear over and over again, clusters that appear periodically
  - Establish threshold for distance in clusters: Human “social distance”: A few meters (if groups are very small); few tens of meters (if groups are medium sized)
  - Include velocities: If divergent → no group

K-Means Clustering



[3]

Fuzzy C-Means Clustering

- Now modify objective function to:

$$J_{GSQE} = \sum_{n=1}^{N}\sum_{k=1}^{K} (r_{nk})^m ||x_n - \mu_k||^2$$

- Exponent m models degree of fuzzyness:

  m → 1 : K-Means (crisp case);

  m → ∞ : r_{nk}→ 1/K (where K is the number of clusters)

- Optimize the obj. fct. under the conditions:

$$\forall x_n \quad : \quad \sum_{k=1}^{K} \alpha_k(x_n) = \sum_{k=1}^{K} r_{nk} = 1$$

$$\forall \mathcal{C}_k \quad : \quad \sum_{n=1}^{N} \alpha_k(x_n) = \sum_{n=1}^{N} r_{nk} > 0$$

Fuzzy C-Means Clustering

- Now modify objective function to:

$$J_{GSQE} = \sum_{n=1}^{N} \sum_{k=1}^{K} (r_{nk})^m ||x_n - \mu_k||^2$$

- Exponent m models degree of fuzzyness:
  m → 1 : K-Means (crisp case);
  m → ∞ : r_{nk} → 1/K (where K is the number of clusters)

- Optimize the obj. fct. under the conditions:

$$\begin{aligned} \forall\, x_n \quad &: \quad \sum_{k=1}^{K} \alpha_k(x_n) = \sum_{k=1}^{K} r_{nk} = 1 \\ \forall\, \mathcal{C}_k \quad &: \quad \sum_{n=1}^{N} \alpha_k(x_n) = \sum_{n=1}^{N} r_{nk} > 0 \end{aligned}$$

Fuzzy C-Means Clustering

- Now modify objective function to:

$$J_{GSQE} = \sum_{n=1}^{N} \sum_{k=1}^{K} (r_{nk})^m ||x_n - \mu_k||^2$$

- Exponent m models degree of fuzzyness:
  m → 1 : K-Means (crisp case);
  m → ∞ : r_{nk} → 1/K (where K is the number of clusters)

- Optimize the obj. fct. under the conditions:

$$\begin{aligned} \forall\, x_n \quad &: \quad \sum_{k=1}^{K} \alpha_k(x_n) = \sum_{k=1}^{K} r_{nk} = 1 \\ \forall\, \mathcal{C}_k \quad &: \quad \sum_{n=1}^{N} \alpha_k(x_n) = \sum_{n=1}^{N} r_{nk} > 0 \end{aligned}$$

Fuzzy C-Means Clustering

- Result:

$$\begin{aligned} r_{nk} &= \left(\sum_{k'=1}^{K} \left(\frac{||x_n - \mu_k||}{||x_n - \mu_{k'}||} \right)^{\frac{2}{m-1}} \right)^{-1} \qquad (☺) \\ \mu_k &= \frac{\sum_{n=1}^{N} r_{nk}^m x_n}{\sum_{n=1}^{N} r_{nk}} \qquad (☺☺) \end{aligned}$$

- the result assumes that no patterns and prototypes coincide

$$\forall\, n, k : \quad ||x_n - \mu_k|| \neq 0$$

if they do coincide, set $r_{nk} = 1$ for $x_n = \mu_k$ and $r_{nk} = 0$ for $x_n \neq \mu_k$

Fuzzy C-Means Clustering

- Limit m → ∞ gives:

$$r_{nk} \xrightarrow{m\to\infty} \frac{1}{\sum_{k'=1}^{K} 1} = \frac{1}{K}$$

- Limit m → 1 we get the nearest neighbor rule (K-Means) because:

$$r_{nk} = 1/((\sum_{k' \neq k} (\tfrac{||x_n - \mu_k||}{||x_n - \mu_{k'}||})^{\frac{2}{m-1}}) + 1)$$

in the limit m→1 the first sum in the denominator becomes ∞ if

$$||x_n - \mu_k|| \neq \min_{1 \leq k' \leq K} ||x_n - \mu_{k'}||$$

and it becomes 0 if

$$||x_n - \mu_k|| = \min_{1 \leq k' \leq K} ||x_n - \mu_{k'}||$$

Fuzzy C-Means Clustering

- Result:

$$r_{nk} = \left(\sum_{k'=1}^{K}\left(\frac{||x_n-\mu_k||}{||x_n-\mu_{k'}||}\right)^{\frac{2}{m-1}}\right)^{-1} \qquad (☺)$$

$$\mu_k = \frac{\sum_{n=1}^{N} r_{nk}^m x_n}{\sum_{n=1}^{N} r_{nk}} \qquad (☺☺)$$

- the result assumes that no patterns and prototypes coincide

$$\forall\, n,k: \quad ||x_n-\mu_k|| \neq 0$$

if they do coincide, set $r_{nk} = 1$ for $x_n = \mu_k$ and $r_{nk} = 0$ for $x_n \neq \mu_k$

Fuzzy C-Means Clustering

- Limit m → ∞ gives:

$$r_{nk} \xrightarrow{m\to\infty} \frac{1}{\sum_{k'=1}^{K} 1} = \frac{1}{K}$$

- Limit m → 1 we get the nearest neighbor rule (K-Means) because:

$$r_{nk} = 1/((\textstyle\sum_{k'\neq k}(\frac{||x_n-\mu_k||}{||x_n-\mu_{k'}||})^{\frac{2}{m-1}}) + 1)$$

in the limit m→1 the first sum in the denominator becomes ∞ if

$$||x_n-\mu_k|| \neq \min_{1\leq k'\leq K} ||x_n-\mu_{k'}||$$

and it becomes 0 if

$$||x_n-\mu_k|| = \min_{1\leq k'\leq K} ||x_n-\mu_{k'}||$$

Fuzzy C-Means Clustering

- Limit m → ∞ gives:

$$r_{nk} \xrightarrow{m\to\infty} \frac{1}{\sum_{k'=1}^{K} 1} = \frac{1}{K}$$

- Limit m → 1 we get the nearest neighbor rule (K-Means) because:

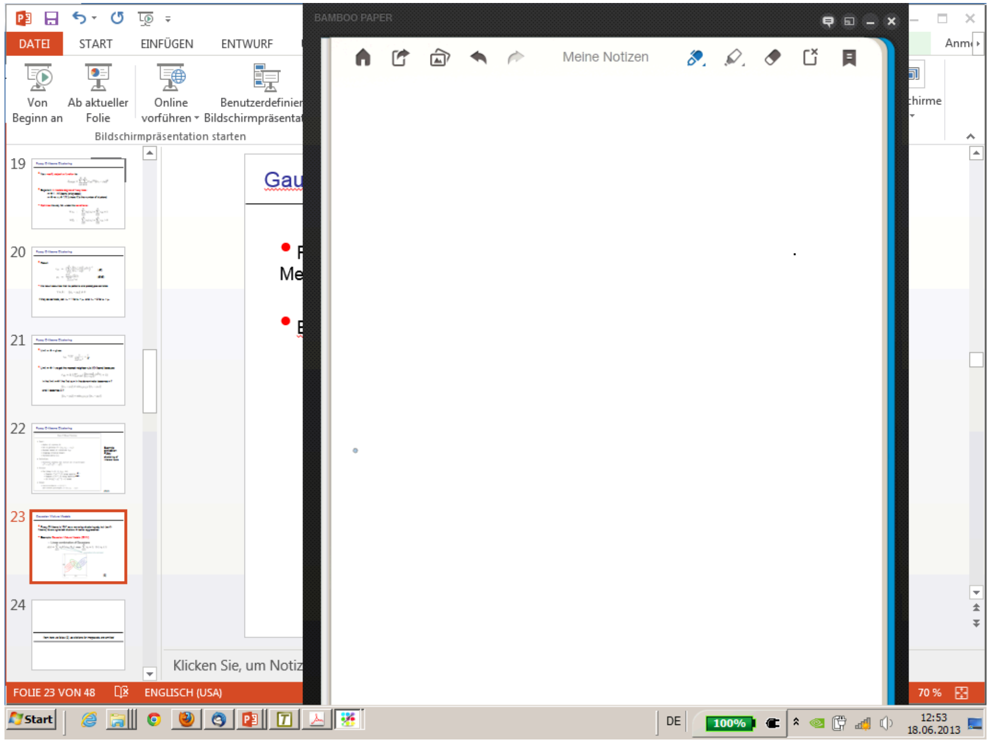
$$r_{nk} = 1/((\textstyle\sum_{k'\neq k}(\frac{||x_n-\mu_k||}{||x_n-\mu_{k'}||})^{\frac{2}{m-1}}) + 1)$$

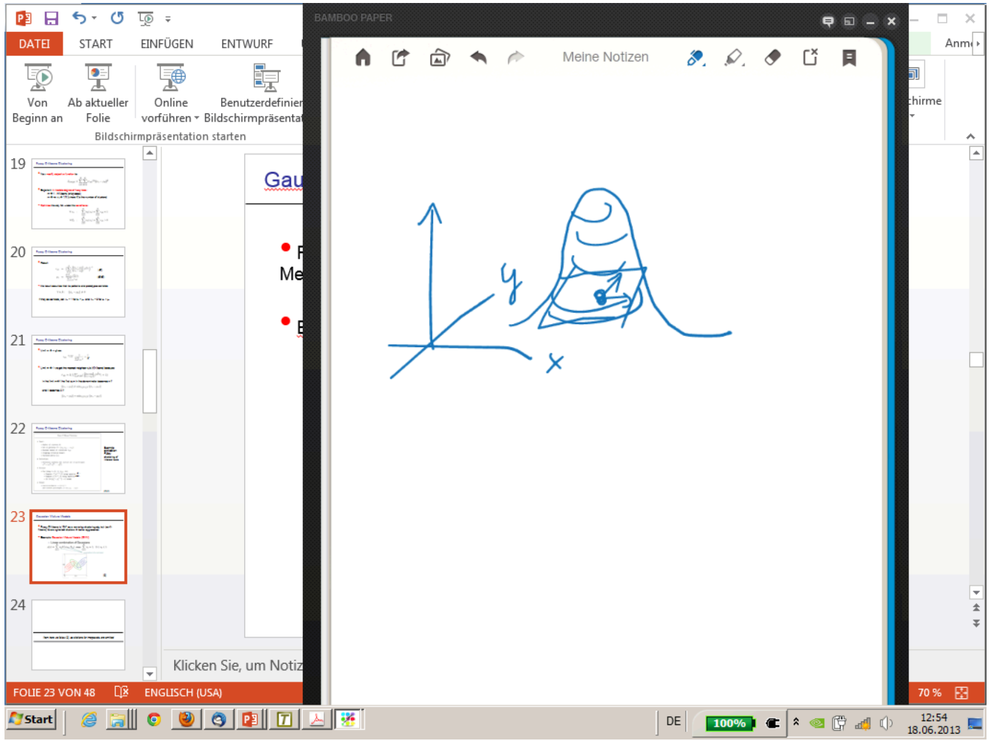
in the limit m→1 the first sum in the denominator becomes ∞ if

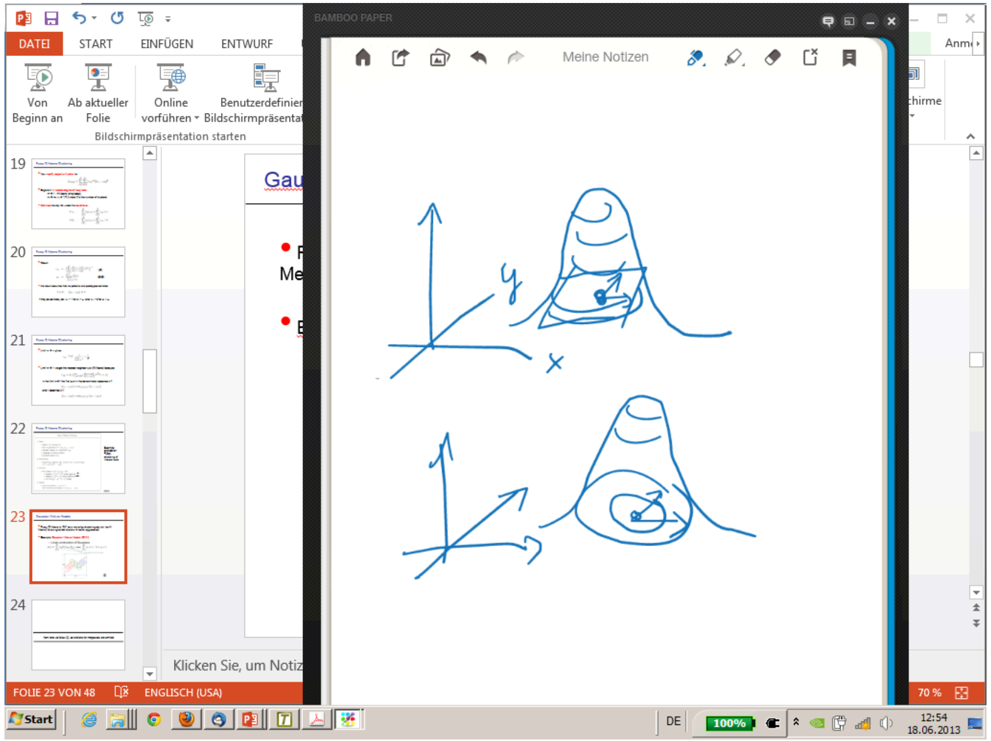
$$||x_n-\mu_k|| \neq \min_{1\leq k'\leq K} ||x_n-\mu_{k'}||$$

and it becomes 0 if

$$||x_n-\mu_k|| = \min_{1\leq k'\leq K} ||x_n-\mu_{k'}||$$







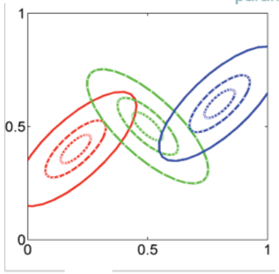
Gaussian Mixture Models

- Fuzzy C-Means is “OK” as a non-crisp clustering alg. but (as K-Means) favors spherical clusters → better approaches

- Example: Gaussian Mixture Models (GMM)
  - Linear combination of Gaussians

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \quad \text{where} \quad \sum_{k=1}^{K} \pi_k = 1, \quad 0 \leqslant \pi_k \leqslant 1$$

parameters to be estimated



[6]

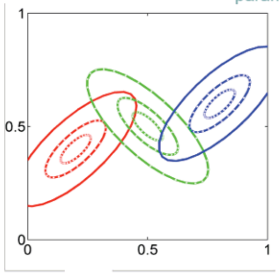
Gaussian Mixture Models

- Fuzzy C-Means is “OK” as a non-crisp clustering alg. but (as K-Means) favors spherical clusters → better approaches

- Example: Gaussian Mixture Models (GMM)
  - Linear combination of Gaussians

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \quad \text{where} \quad \sum_{k=1}^{K} \pi_k = 1, \quad 0 \leqslant \pi_k \leqslant 1$$

parameters to be estimated



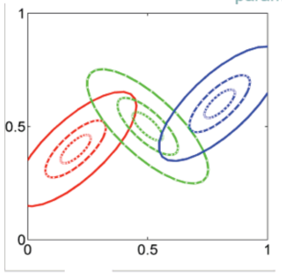
[6]

Gaussian Mixture Models

- Fuzzy C-Means is “OK” as a non-crisp clustering alg. but (as K-Means) favors spherical clusters → better approaches
- Example: Gaussian Mixture Models (GMM)
  - Linear combination of Gaussians

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \quad \text{where} \quad \sum_{k=1}^{K} \pi_k = 1, \quad 0 \leqslant \pi_k \leqslant 1$$

parameters to be estimated



[6]

Machine Learning

Learning a Generative Model for data [8]:

For a distribution $p(x|\theta)$, parameterised by θ, and data $\mathcal{X} = \{x^1, \ldots, x^N\}$ learning corresponds to inferring the θ that best explains the data $\mathcal{X}$

Bayes theorem $\rightarrow p(\theta|\mathcal{X}) \propto p(\mathcal{X}|\theta)p(\theta)$

- **Maximum A posteriori** $\theta^{MAP} = \underset{\theta}{\text{argmax}}\ p(\theta|\mathcal{X})$
- **Maximum Likelihood** $\theta^{ML} = \underset{\theta}{\text{argmax}}\ p(\mathcal{X}|\theta)$

$$= \underset{\theta}{\text{argmax}}\ L(\mathcal{X}, \theta)$$

Machine Learning

Learning a Generative Model for data [8]:

For a distribution $p(x|\theta)$, parameterised by θ, and data $\mathcal{X} = \{x^1, \ldots, x^N\}$ learning corresponds to inferring the θ that best explains the data $\mathcal{X}$

Bayes theorem $\rightarrow p(\theta|\mathcal{X}) \propto p(\mathcal{X}|\theta)p(\theta)$

- **Maximum A posteriori** $\theta^{MAP} = \underset{\theta}{\text{argmax}}\ p(\theta|\mathcal{X})$
- **Maximum Likelihood** $\theta^{ML} = \underset{\theta}{\text{argmax}}\ p(\mathcal{X}|\theta)$

$$= \underset{\theta}{\text{argmax}}\ L(\mathcal{X}, \theta)$$

Machine Learning

Learning a Generative Model for data [8]:

For a distribution $p(x|\theta)$, parameterised by θ, and data $\mathcal{X} = \{x^1, \ldots, x^N\}$ learning corresponds to inferring the θ that best explains the data $\mathcal{X}$

Bayes theorem $\rightarrow p(\theta|\mathcal{X}) \propto p(\mathcal{X}|\theta)p(\theta)$

- **Maximum A posteriori** $\theta^{MAP} = \underset{\theta}{\text{argmax}}\ p(\theta|\mathcal{X})$
- **Maximum Likelihood** $\theta^{ML} = \underset{\theta}{\text{argmax}}\ p(\mathcal{X}|\theta)$

$$= \underset{\theta}{\text{argmax}}\ L(\mathcal{X}, \theta)$$

Machine Learning

Learning a Generative Model for data [8]:

For a distribution $p(x|\theta)$, parameterised by θ, and data $\mathcal{X} = \{x^1, \dots, x^N\}$ learning corresponds to inferring the θ that best explains the data $\mathcal{X}$

Bayes theorem → $p(\theta|\mathcal{X}) \propto p(\mathcal{X}|\theta)p(\theta)$

- **Maximum A posteriori** $\theta^{MAP} = \underset{\theta}{\operatorname{argmax}}\; p(\theta|\mathcal{X})$
- **Maximum Likelihood** $\theta^{ML} = \underset{\theta}{\operatorname{argmax}}\; p(\mathcal{X}|\theta)$

$$= \underset{\theta}{\operatorname{argmax}}\;\; L(\mathcal{X}, \theta)$$

Machine Learning

Learning a Generative Model for data [8]:

For a distribution $p(x|\theta)$, parameterised by θ, and data $\mathcal{X} = \{x^1, \dots, x^N\}$ learning corresponds to inferring the θ that best explains the data $\mathcal{X}$

Bayes theorem → $p(\theta|\mathcal{X}) \propto p(\mathcal{X}|\theta)p(\theta)$

- **Maximum A posteriori** $\theta^{MAP} = \underset{\theta}{\operatorname{argmax}}\; p(\theta|\mathcal{X})$
- **Maximum Likelihood** $\theta^{ML} = \underset{\theta}{\operatorname{argmax}}\; p(\mathcal{X}|\theta)$

$$= \underset{\theta}{\operatorname{argmax}}\;\; L(\mathcal{X}, \theta)$$

Machine Learning

Learning a Generative Model for data [8]:

For a distribution $p(x|\theta)$, parameterised by θ, and data $\mathcal{X} = \{x^1, \dots, x^N\}$ learning corresponds to inferring the θ that best explains the data $\mathcal{X}$

Bayes theorem → $p(\theta|\mathcal{X}) \propto p(\mathcal{X}|\theta)p(\theta)$

- **Maximum A posteriori** $\theta^{MAP} = \underset{\theta}{\operatorname{argmax}}\; p(\theta|\mathcal{X})$
- **Maximum Likelihood** $\theta^{ML} = \underset{\theta}{\operatorname{argmax}}\; p(\mathcal{X}|\theta)$

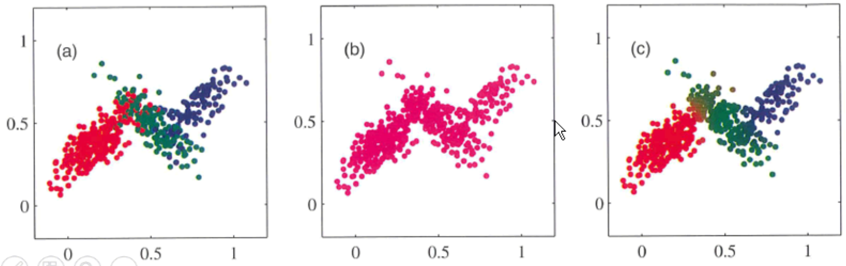
$$= \underset{\theta}{\operatorname{argmax}}\;\; L(\mathcal{X}, \theta)$$

GMM-Basics

- Responsibilities

$$\begin{aligned}\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) &= \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum\limits_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum\limits_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.\end{aligned}$$

- Example

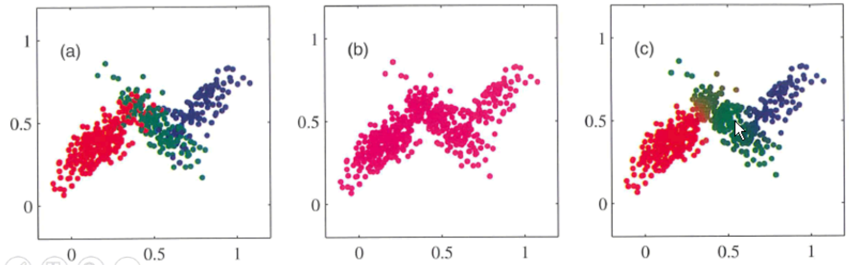


GMM-Basics

- Responsibilities

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

- Example



GMM-Basics

Maximum likelihood (one multivariate Gaussian)

$$p(\mathbf{x}|\theta) = \mathcal{N}(\mathbf{x}|\underbrace{\boldsymbol{\mu}, \boldsymbol{\Sigma}}_{\theta}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right\}$$

- Likelihood $L(\mathbf{x},\theta) = p(\mathbf{x}|\theta)$

- Maximum likelihood $\theta_{\text{best}} = \operatorname{argmax}_{\theta} L(\mathbf{x},\theta)$
  $= \operatorname{argmax}_{\theta} \ln L(\mathbf{x},\theta)$

- Pattern matrix X of N iid measurements (D-dim. pattern vectors $\mathbf{x}$),

$$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^{\mathrm{T}}$$

$$L(\mathbf{X},\Theta) = \prod_{i=1}^{N} L(\mathbf{x}_i,\Theta) \qquad \ln L(\mathbf{X},\Theta) = \sum_{i=1}^{N} \ln L(\mathbf{x}_i,\Theta)$$

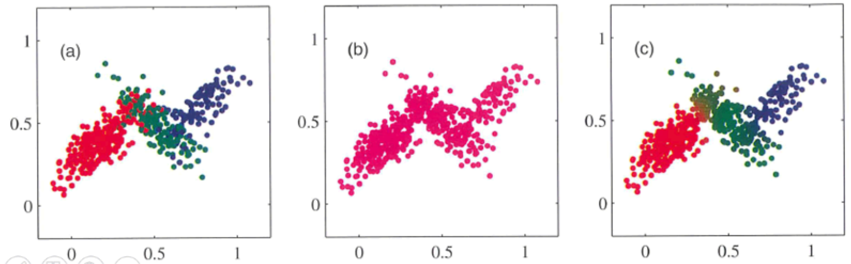
$$\ln L(\mathbf{X},\Theta) = \ln p(\mathbf{X}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{i=1}^{N} \ln N(\mathbf{x}_i\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma})$$

GMM-Basics

- Responsibilities

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

- Example

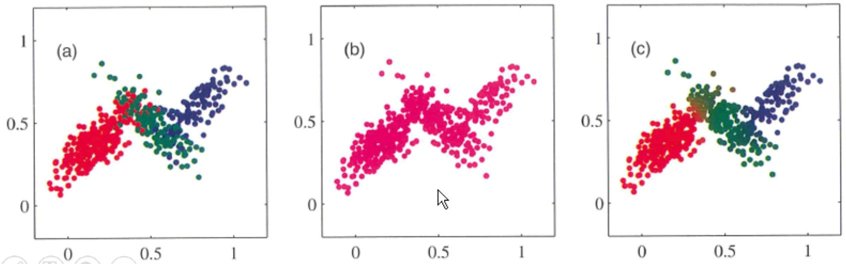


GMM-Basics

- Responsibilities

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

- Example



GMM-Basics

Maximum likelihood (one multivariate Gaussian)

$$p(\mathbf{x}|\theta)= \mathcal{N}(\mathbf{x}|\underbrace{\boldsymbol{\mu},\boldsymbol{\Sigma}}_{\theta}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- Likelihood $L(\mathbf{x},\theta) = p(\mathbf{x}|\theta)$
- Maximum likelihood $\theta_{\text{best}} = \operatorname{argmax}_{\theta} L(\mathbf{x},\theta) = \operatorname{argmax}_{\theta} \ln L(\mathbf{x},\theta)$
- Pattern matrix X of N iid measurements (D-dim. pattern vectors $\mathbf{x}$),

$\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_N)^{\mathrm{T}}$

$$L(\mathbf{X},\Theta) = \prod_{i=1}^{N} L(\mathbf{x}_i,\Theta) \qquad \ln L(\mathbf{X},\Theta) = \sum_{i=1}^{N} \ln L(\mathbf{x}_i,\Theta)$$

$$\ln L(\mathbf{X},\Theta) = \ln p(\mathbf{X}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{i=1}^{N} \ln N(\mathbf{x}_i\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma})$$

GMM-Basics

Maximum likelihood (one multivariate Gaussian)

$$p(\mathbf{x}|\theta)= \mathcal{N}(\mathbf{x}|\underbrace{\boldsymbol{\mu},\boldsymbol{\Sigma}}_{\theta}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- Likelihood $L(\mathbf{x},\theta) = p(\mathbf{x}|\theta)$
- Maximum likelihood $\theta_{\text{best}} = \operatorname{argmax}_{\theta} L(\mathbf{x},\theta) = \operatorname{argmax}_{\theta} \ln L(\mathbf{x},\theta)$
- Pattern matrix X of N iid measurements (D-dim. pattern vectors $\mathbf{x}$),

$\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_N)^{\mathrm{T}}$

$$L(\mathbf{X},\Theta) = \prod_{i=1}^{N} L(\mathbf{x}_i,\Theta) \qquad \ln L(\mathbf{X},\Theta) = \sum_{i=1}^{N} \ln L(\mathbf{x}_i,\Theta)$$

$$\ln L(\mathbf{X},\Theta) = \ln p(\mathbf{X}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{i=1}^{N} \ln N(\mathbf{x}_i\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma})$$

GMM-Basics

Maximum likelihood (one multivariate Gaussian)

$$p(\mathbf{x}|\theta)= \mathcal{N}(\mathbf{x}|\underbrace{\boldsymbol{\mu},\boldsymbol{\Sigma}}_{\theta}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- Likelihood $L(\mathbf{x},\theta) = p(\mathbf{x}|\theta)$
- Maximum likelihood $\theta_{\text{best}} = \operatorname{argmax}_{\theta} L(\mathbf{x},\theta) = \operatorname{argmax}_{\theta} \ln L(\mathbf{x},\theta)$
- Pattern matrix X of N iid measurements (D-dim. pattern vectors $\mathbf{x}$),

$\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_N)^{\mathrm{T}}$

$$L(\mathbf{X},\Theta) = \prod_{i=1}^{N} L(\mathbf{x}_i,\Theta) \qquad \ln L(\mathbf{X},\Theta) = \sum_{i=1}^{N} \ln L(\mathbf{x}_i,\Theta)$$

$$\ln L(\mathbf{X},\Theta) = \ln p(\mathbf{X}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{i=1}^{N} \ln N(\mathbf{x}_i\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma})$$

GMM-Basics

Maximum likelihood (one multivariate Gaussian)

$$\ln L(\mathbf{X},\Theta) = \ln L(\mathbf{X},\boldsymbol{\mu},\boldsymbol{\Sigma}) =$$

$$\ln p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})$$

$$\theta_{\text{best}} = \operatorname{argmax}_{\theta} \ln L(\mathbf{x},\theta) \rightarrow \left.\begin{aligned} \boldsymbol{\mu}_{\text{best}}&: \frac{\partial}{\partial\boldsymbol{\mu}}\ln L(\mathbf{X},\boldsymbol{\mu},\boldsymbol{\Sigma}) = 0 \\ \boldsymbol{\Sigma}_{\text{best}}&: \frac{\partial}{\partial\boldsymbol{\Sigma}}\ln L(\mathbf{X},\boldsymbol{\mu},\boldsymbol{\Sigma}) = 0 \end{aligned}\right\}$$

$$\left\{\begin{aligned} \boldsymbol{\mu}_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \\ \boldsymbol{\Sigma}_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}} \end{aligned}\right.$$

GMM-Basics

Maximum likelihood (one multivariate Gaussian)

$$\ln L(\mathbf{X},\Theta) = \ln L(\mathbf{X},\boldsymbol{\mu},\boldsymbol{\Sigma}) =$$

$$\ln p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})$$

$$\theta_{\text{best}} = \operatorname{argmax}_{\theta}\ \ln L(\mathbf{x},\theta) \rightarrow \quad \begin{cases} \boldsymbol{\mu}_{\text{best}}: & \dfrac{\partial}{\partial\boldsymbol{\mu}}\ln L(\mathbf{X},\boldsymbol{\mu},\boldsymbol{\Sigma}) = 0 \\ \boldsymbol{\Sigma}_{\text{best}}: & \dfrac{\partial}{\partial\boldsymbol{\Sigma}}\ln L(\mathbf{X},\boldsymbol{\mu},\boldsymbol{\Sigma}) = 0 \end{cases}$$

$$\begin{cases} \boldsymbol{\mu}_{\mathrm{ML}} = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N}\mathbf{x}_n \\ \boldsymbol{\Sigma}_{\mathrm{ML}} = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}} \end{cases}$$

GMM-Basics

- GMM

$$p(\mathbf{x}|\theta) = \quad p(\mathbf{x}) = \sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k). \qquad 0 \leqslant \pi_k \leqslant 1 \qquad \sum_{k=1}^{K}\pi_k = 1$$

- 1 of K representation

K-dimensional binary random variable $\mathbf{z}$

$$z_k \in \{0,1\} \text{ and } \textstyle\sum_k z_k = 1$$

$$p(z_k = 1) = \pi_k$$

$$p(\mathbf{z}) = \prod_{k=1}^{K}\pi_k^{z_k}$$

- conditional probability

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \qquad p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)^{z_k}$$

$$p(\mathbf{x}) = \sum_{\mathbf{z}}\underbrace{p(\mathbf{z})p(\mathbf{x}|\mathbf{z})}_{p(x,z)} = \sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$$

GMM-Basics

- GMM

$$p(\mathbf{x}|\theta) = \quad p(\mathbf{x}) = \sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k). \qquad 0 \leqslant \pi_k \leqslant 1 \qquad \sum_{k=1}^{K}\pi_k = 1$$

- 1 of K representation

K-dimensional binary random variable $\mathbf{z}$

$$z_k \in \{0,1\} \text{ and } \textstyle\sum_k z_k = 1$$

$$p(z_k = 1) = \pi_k$$

$$p(\mathbf{z}) = \prod_{k=1}^{K}\pi_k^{z_k}$$

- conditional probability

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \qquad p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)^{z_k}$$

$$p(\mathbf{x}) = \sum_{\mathbf{z}}\underbrace{p(\mathbf{z})p(\mathbf{x}|\mathbf{z})}_{p(x,z)} = \sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$$

GMM-Basics

- GMM

$$p(\mathbf{x}|\theta) = \quad p(\mathbf{x}) = \sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k). \qquad 0 \leqslant \pi_k \leqslant 1 \qquad \sum_{k=1}^{K}\pi_k = 1$$

- 1 of K representation

K-dimensional binary random variable $\mathbf{z}$

$$z_k \in \{0,1\} \text{ and } \textstyle\sum_k z_k = 1$$

$$p(z_k = 1) = \pi_k$$

$$p(\mathbf{z}) = \prod_{k=1}^{K}\pi_k^{z_k}$$

remark: If we have several observations $\mathbf{x}_1,\ldots,\mathbf{x}_N$, then, because we have represented the marginal distribution in the form $p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z})$, it follows that for every observed data point $\mathbf{x}_n$ there is a corresponding latent variable $\mathbf{z}_n$.

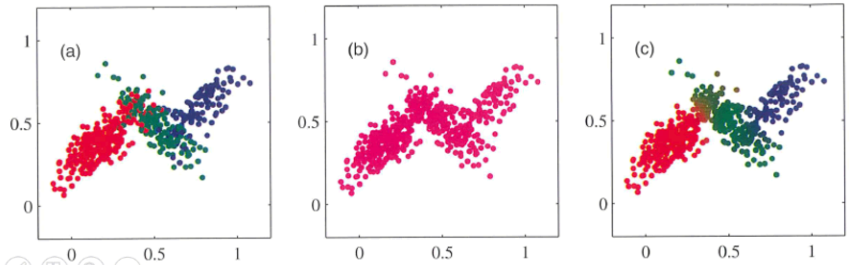
$$p(\mathbf{x}) = \sum_{\mathbf{z}}\underbrace{p(\mathbf{z})p(\mathbf{x}|\mathbf{z})}_{p(x,z)} = \sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$$

GMM-Basics

- Responsibilities

$$\gamma(z_k) \equiv p(z_k=1|\mathbf{x}) = \frac{p(z_k=1)p(\mathbf{x}|z_k=1)}{\sum_{j=1}^{K} p(z_j=1)p(\mathbf{x}|z_j=1)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}.$$

- Example



GMM-Basics

Maximum likelihood (GMM)

$$\ln L(\mathbf{X},\Theta) = \ln L(\mathbf{X},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \ln p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln\left\{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right\}$$

Vector of K *D*-dim. means $\boldsymbol{\mu}_k$

Vector of K *DxD* covariances $\boldsymbol{\Sigma}_k$

- maximizing w.r.t $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ →

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n \qquad \boldsymbol{\Sigma}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n-\boldsymbol{\mu}_k)(\mathbf{x}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}$$

$$\left(N_k = \sum_{n=1}^{N}\gamma(z_{nk}). \right) \qquad \pi_k = \frac{N_k}{N}$$

GMM-Basics

Maximum likelihood (GMM)

$$\ln L(\mathbf{X},\Theta) = \ln L(\mathbf{X},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \ln p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln\left\{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right\}$$

Vector of K *D*-dim. means $\boldsymbol{\mu}_k$

Vector of K *DxD* covariances $\boldsymbol{\Sigma}_k$

- maximizing w.r.t $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ →

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n \qquad \boldsymbol{\Sigma}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n-\boldsymbol{\mu}_k)(\mathbf{x}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}$$

$$\left(N_k = \sum_{n=1}^{N}\gamma(z_{nk}). \right) \qquad \pi_k = \frac{N_k}{N}$$

GMM-Basics

Maximum likelihood (GMM)

$$\ln L(\mathbf{X},\Theta) = \ln L(\mathbf{X},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \ln p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln\left\{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right\}$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n \qquad \boldsymbol{\Sigma}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n-\boldsymbol{\mu}_k)(\mathbf{x}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}$$

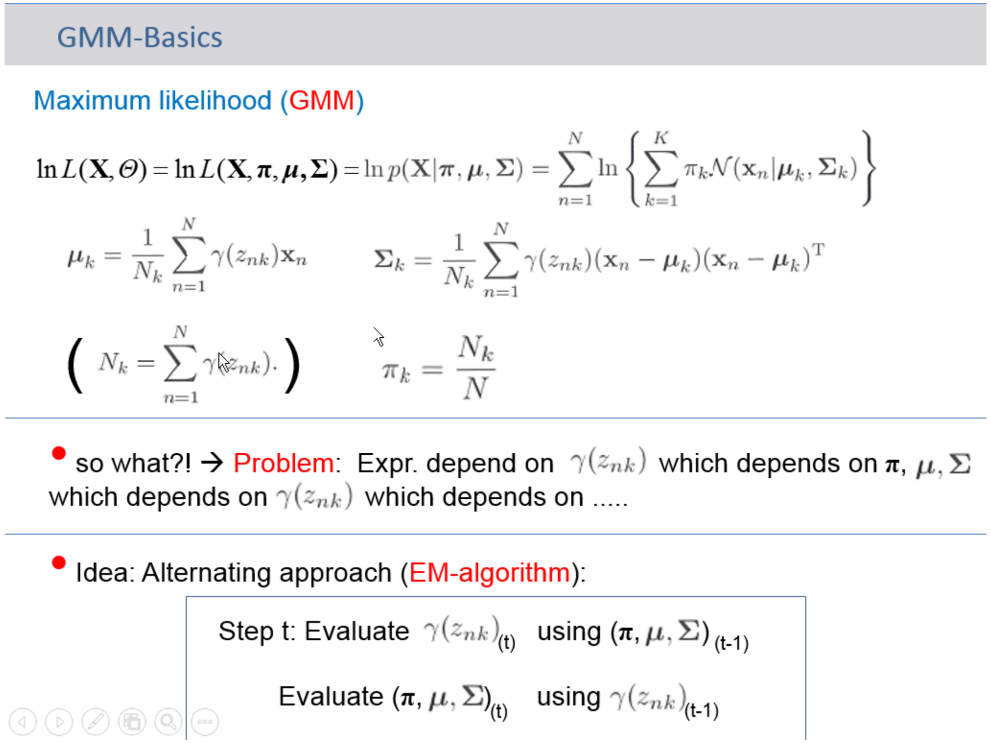
$$\left(N_k = \sum_{n=1}^{N}\gamma(z_{nk}). \right) \qquad \pi_k = \frac{N_k}{N}$$

- so what?! → Problem: Expr. depend on $\gamma(z_{nk})$ which depends on $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ which depends on $\gamma(z_{nk})$ which depends on .....

- Idea: Alternating approach (EM-algorithm):

Step t: Evaluate $\gamma(z_{nk})_{(t)}$ using $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})_{(t-1)}$

Evaluate $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})_{(t)}$ using $\gamma(z_{nk})_{(t-1)}$



GMM-Basics

Maximum likelihood (GMM)

$$\ln L(\mathbf{X},\Theta)=\ln L(\mathbf{X},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma})=\ln p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right\}$$

$$\boldsymbol{\mu}_k=\frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n \qquad \boldsymbol{\Sigma}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n-\boldsymbol{\mu}_k)(\mathbf{x}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}$$

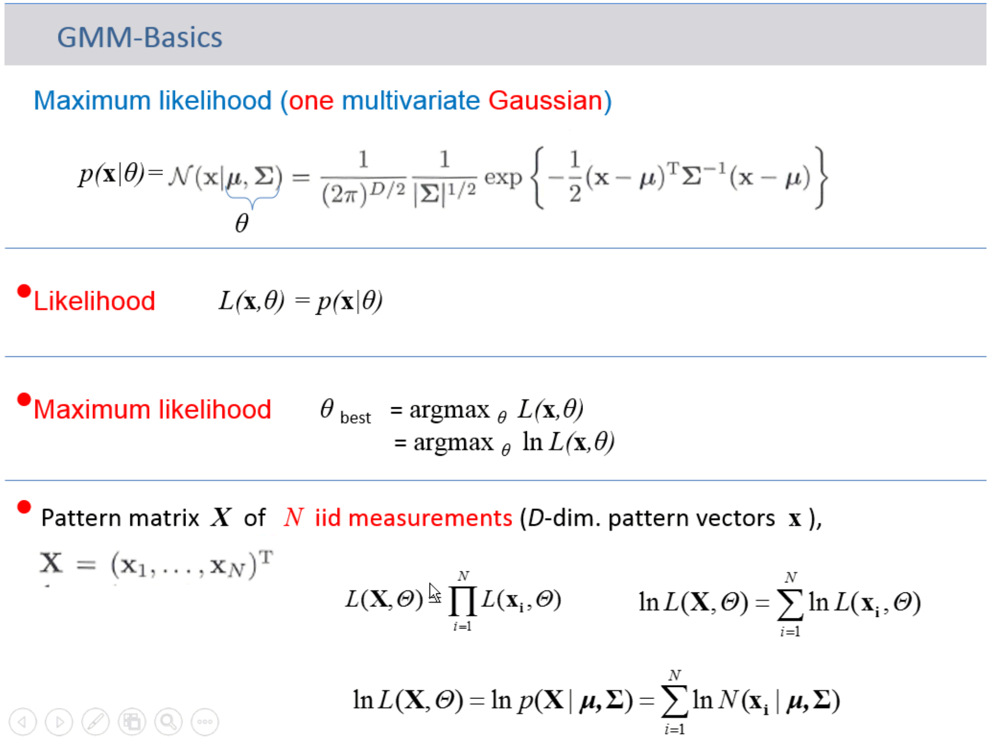
$$\left(N_k=\sum_{n=1}^{N}\gamma(z_{nk}). \right) \qquad \pi_k=\frac{N_k}{N}$$

- so what?! → Problem: Expr. depend on $\gamma(z_{nk})$ which depends on $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ which depends on $\gamma(z_{nk})$ which depends on .....

- Idea: Alternating approach (EM-algorithm):

Step t: Evaluate $\gamma(z_{nk})_{(t)}$ using $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})_{(t-1)}$

Evaluate $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})_{(t)}$ using $\gamma(z_{nk})_{(t-1)}$



GMM-Basics

Maximum likelihood (one multivariate Gaussian)

$$p(\mathbf{x}|\theta)=\mathcal{N}(\mathbf{x}|\underbrace{\boldsymbol{\mu},\boldsymbol{\Sigma}}_{\theta}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- Likelihood $L(\mathbf{x},\theta) = p(\mathbf{x}|\theta)$

- Maximum likelihood $\theta_{\text{best}} = \text{argmax}_{\theta}\, L(\mathbf{x},\theta)$ $= \text{argmax}_{\theta}\, \ln L(\mathbf{x},\theta)$

- Pattern matrix X of N iid measurements (*D*-dim. pattern vectors $\mathbf{x}$),

$\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_N)^{\mathrm{T}}$

$$L(\mathbf{X},\Theta)=\prod_{i=1}^{N}L(\mathbf{x}_i,\Theta) \qquad \ln L(\mathbf{X},\Theta)=\sum_{i=1}^{N}\ln L(\mathbf{x}_i,\Theta)$$

$$\ln L(\mathbf{X},\Theta)=\ln p(\mathbf{X}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma})=\sum_{i=1}^{N}\ln N(\mathbf{x}_i\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma})$$