Unsupervised Classification

A second partitionning-based algorithm: Gaussian Mixture Models (GMM)

Data Analysis – juliette.chevallier@insa-toulouse.fr

INSA Toulouse, Applied Mathematics, 4th year

1. Finite Mixture Models

1.1 General Principle

1.2 Step 1: Modeling and Choice of Distributions

1.3 Step 2: Parameters Estimation

1.4 Step 3: Maximum a Posteriori & Classification

1.5 Step 4: Model Selection Criteria

2. Gaussian Mixture Models

2.1 Multivariate Gaussian Mixtures

Finite Mixture Models

1.1 General Principle

1.2 Step 1: Modeling and Choice of Distributions

1.3 Step 2: Parameters Estimation

1.4 Step 3: Maximum a Posteriori & Classification

1.5 Step 4: Model Selection Criteria

Introduction

We observe n individuals described by p variables: $x_i = \left(x_{i1}, x_{i2}, \dots, x_{ip}\right) \in \mathcal{X}$

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

$$\mathcal{X} = \mathbb{R}^p\,,\ \{0,1\}^p\,,\ ]-\pi,\pi]^p\,,\ \mathbb{R}^q \times \{0,1\}^{p-q}\,, \dots$$

- Initial measurements,
- Transformed measurements,
- Coordinates after dimension reduction.

- **Assumption**: The data come from a population composed of several sub-populations.

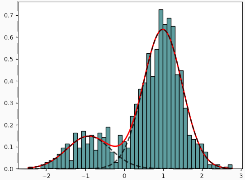
Introduction

We observe n individuals described by p variables: $x_i = \left(x_{i1}, x_{i2}, \ldots, x_{ip}\right) \in \mathcal{X}$

$$X = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

$$\mathcal{X} = \mathbb{R}^p\,,\ \{0,1\}^p\,,\ ]-\pi,\pi]^p\,,\ \mathbb{R}^q \times \{0,1\}^{p-q}\,, \ldots$$

- Initial measurements,
- Transformed measurements,
- Coordinates after dimension reduction.



- **Assumption**: The data come from a population composed of several sub-populations.

- **Modeling**:
  - Each sub-population is modeled independently of the others,

    ⇝ Choice of a distribution law for each sub-population.
  - Total population seen as a mixture of these sub-populations,

    ⇝ Finite mixture model.

Finite Mixture Model

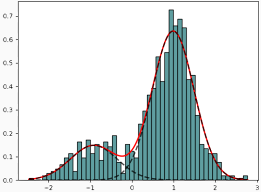
Mixture model of K components: *Given:*

- $(\alpha_k)_{k \in [\![1,K]\!]}$ the *proportions* of the mixture

$$\forall k \in [\![1,K]\!]\,, \quad \alpha_k \in [0,1] \qquad \text{and} \qquad \sum_{k=1}^{K} \alpha_k = 1\,;$$

- For all k, $f_k\big(\,\cdot\,;\,\omega_k\big)$ the *density* of the k-th sub-population, which (possibly) depends on a parameter ω_k; *and*

- $\theta = \big(\alpha_k, \omega_k\big)_{k \in [\![1,K]\!]}$ the whole *parameters* of the mixture model.



Finite Mixture Model

Mixture model of K components: *Given:*

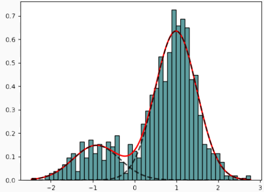
- $(\alpha_k)_{k\in[\![1,K]\!]}$ the *proportions* of the mixture

$$\forall k \in [\![1, K]\!]\,, \quad \alpha_k \in [0,1] \qquad \text{and} \qquad \sum_{k=1}^{K} \alpha_k = 1\,;$$

- For all k, $f_k\big(\,\cdot\,;\,\omega_k\big)$ the *density* of the k-th sub-population, which (possibly) depends on a parameter ω_k; *and*

- $\theta = \big(\alpha_k, \omega_k\big)_{k\in[\![1,K]\!]}$ the whole *parameters* of the mixture model.

We define: $$\boxed{q\big(\,\cdot\,;\theta\big) \;=\; \sum_{k=1}^{K} \alpha_k\, f_k\big(\,\cdot\,;\,\omega_k\big)}$$



Mixture Models *vs.* Latent Variables

Mixture model: $q\big(\cdot\,;\theta\big) = \sum_{k=1}^{K} \alpha_k \, f_k\big(\cdot\,;\omega_k\big) \quad \longleftrightarrow \quad \begin{cases} \text{Proportions } \alpha_k \\ \text{Densities } f_k\big(\cdot\,;\omega_k\big) \end{cases}$

Question: *How to generate data according to such a model?* $(x_i)_{i \in [\![1,n]\!]}$

Mixture Models *vs.* Latent Variables

Mixture model: $q\left(\cdot\,;\theta\right) = \sum_{k=1}^{K} \alpha_k\, f_k\left(\cdot\,;\,\omega_k\right) \quad \longleftrightarrow \quad \begin{cases} \text{Proportions } \alpha_k \\ \text{Densities } f_k\left(\cdot\,;\,\omega_k\right) \end{cases}$

Question: *How to generate data according to such a model?* $(x_i)_{i\in[\![1,n]\!]}$

For all individual $i \in [\![1,n]\!]$,

(i) Let $\mathcal{P}_K = \{\mathcal{C}_1,\ldots,\mathcal{C}_K\}$ be a partition of $[\![1,n]\!]$ into K classes, such that the individual i belongs to $\mathcal{C}_k$ with probability α_k: $\mathbb{P}\left(i\in\mathcal{C}_k;\,\alpha_k\right) = \alpha_k$;

(ii) Then, x_i is generated according to the density $f_k\left(\cdot\,;\,\omega_k\right)$ if $i\in\mathcal{C}_k$.

Mixture Models *vs.* Latent Variables

Mixture model: $q\big(\cdot\,;\theta\big) = \sum_{k=1}^{K} \alpha_k\, f_k\big(\cdot\,;\omega_k\big) \quad \longleftrightarrow \quad \begin{cases} \text{Proportions } \alpha_k \\ \text{Densities } f_k\big(\cdot\,;\omega_k\big) \end{cases}$

Question: *How to generate data according to such a model?* $(x_i)_{i\in[\![1,n]\!]}$

For all individual $i \in [\![1,n]\!]$,

(i) Let $\mathcal{P}_K = \{\mathcal{C}_1,\dots,\mathcal{C}_K\}$ be a partition of $[\![1,n]\!]$ into K classes, such that the individual i belongs to $\mathcal{C}_k$ with probability α_k: $\mathbb{P}\left(i \in \mathcal{C}_k;\, \alpha_k\right) = \alpha_k$;

(ii) Then, x_i is generated according to the density $f_k\big(\cdot\,;\omega_k\big)$ if $i \in \mathcal{C}_k$.

$\rightsquigarrow$ Latent variables $(z_i)_{i\in[\![1,n]\!]}$ to encode the classes.

For all individual $i \in [\![1,n]\!]$,

$$\boxed{\begin{cases} z_i \,|\, (\alpha_k)_{k\in[\![1,K]\!]} \sim \displaystyle\sum_{k=1}^{K} \alpha_k \delta_k \\ x_i \,|\, z_i,\, \theta = (\alpha_k, \omega_k)_{k\in[\![1,K]\!]} \sim f_{z_i}\big(\cdot\,;\omega_{z_i}\big) \end{cases}}$$

Hierarchical Writing of Mixture Models

- **Mixture Models**: For all $i \in [\![1, n]\!]$, $\begin{cases} z_i \,|\, (\alpha_k)_{k \in [\![1,K]\!]} \sim \displaystyle\sum_{k=1}^{K} \alpha_k \delta_k \,, \\ x_i \,|\, z_i,\, (\alpha_k, \omega_k)_{k \in [\![1,K]\!]} \sim f_{z_i}\big(\,\cdot\,;\, \omega_{z_i}\big). \end{cases}$

Hierarchical Writing of Mixture Models

- **Mixture Models**: For all $i \in [\![1, n]\!]$, $\begin{cases} z_i \,|\, (\alpha_k)_{k \in [\![1,K]\!]} \sim \displaystyle\sum_{k=1}^{K} \alpha_k \delta_k \,, \\ x_i \,|\, z_i,\, (\alpha_k, \omega_k)_{k \in [\![1,K]\!]} \sim f_{z_i}\big(\,\cdot\,;\, \omega_{z_i}\big). \end{cases}$

- **Complete likelihood**: For all $x = (x_i)_{i \in [\![1,n]\!]}$ and $z = (z_i)_{i \in [\![1,n]\!]}$,

$$q(x, z\,;\,\theta) = \prod_{i=1}^{n} q(x_i, z_i\,;\,\theta) = \prod_{i=1}^{n} q(x_i | z_i\,;\,\theta)\, q(z_i\,;\,\theta) = \prod_{i=1}^{n} \alpha_{z_i}\, f_{z_i}\big(x_i\,;\,\omega_{z_i}\big)\,.$$

Hierarchical Writing of Mixture Models

- **Mixture Models**: For all $i \in [\![1, n]\!]$, $\begin{cases} z_i \mid (\alpha_k)_{k \in [\![1,K]\!]} \sim \sum_{k=1}^{K} \alpha_k \delta_k \,, \\ x_i \mid z_i,\, (\alpha_k, \omega_k)_{k \in [\![1,K]\!]} \sim f_{z_i}\big(\cdot\,;\, \omega_{z_i}\big). \end{cases}$

- **Complete likelihood**: For all $x = (x_i)_{i \in [\![1,n]\!]}$ and $z = (z_i)_{i \in [\![1,n]\!]}$,

$$q(x, z\,;\, \theta) = \prod_{i=1}^{n} q(x_i, z_i\,;\, \theta) = \prod_{i=1}^{n} q(x_i | z_i\,;\, \theta)\, q(z_i\,;\, \theta) = \prod_{i=1}^{n} \alpha_{z_i}\, f_{z_i}\big(x_i\,;\, \omega_{z_i}\big)\,.$$

- **Posterior distribution**: For all $i \in [\![1, n]\!]$,

$$\begin{aligned} q(x_i\,;\, \theta) &= \sum_{k=1}^{K} q(x_i, \{z_i = k\}\,;\, \theta) \\ &= \sum_{k=1}^{K} q(x_i \mid \{z_i = k\}\,;\, \theta)\, q(\{z_i = k\}\,;\, \theta) = \sum_{k=1}^{K} \alpha_k\, f_k\big(x_i\,;\, \omega_k\big) \end{aligned}$$

Steps to Define a Mixture Model

1. **Modeling and Choice of Distributions**

2. **Parameters Estimation**

3. **Maximum a Posteriori & Classification**

4. **Model Selection Criteria**

Steps to Define a Mixture Model

1. **Modeling and Choice of Distributions**
   - Initial choice of the model,
   - Selection of a suitable density family $f_k(\,\cdot\,;\,\omega_k)$,
   - Choice to be made according to the problem/data studied,

   $\rightsquigarrow$ Collection of models: One model for each fixed number of classes $K \in \mathbb{N}^\star$:

$$\mathcal{M}_K := \left\{ x \in \mathbb{R}^d \mapsto q_K(x\,;\,\theta) = \sum_{k=1}^{K} \alpha_k\, f_k\Big(x\,;\omega_k\Big) \right\} .$$

2. **Parameters Estimation**

3. **Maximum a Posteriori & Classification**

4. **Model Selection Criteria**

Steps to Define a Mixture Model

1. **Modeling and Choice of Distributions**
   - Initial choice of the model,
   - Selection of a suitable density family $f_k(\,\cdot\,;\,\omega_k)$,
   - Choice to be made according to the problem/data studied,

     $\rightsquigarrow$ Collection of models: One model for each fixed number of classes $K \in \mathbb{N}^\star$:

$$\mathcal{M}_K := \left\{ x \in \mathbb{R}^d \mapsto q_K(x\,;\,\theta) = \sum_{k=1}^{K} \alpha_k\, f_k\Big(x\,;\omega_k\Big) \right\} .$$

2. **Parameters Estimation**
   - In each $\mathcal{M}_K$ model, we identify the mixture that best fits the data: $q_K(\,\cdot\,;\,\hat{\theta})$.

     $\rightsquigarrow$ Parameter estimation algorithm.

3. **Maximum a Posteriori & Classification**

4. **Model Selection Criteria**

Steps to Define a Mixture Model

1. **Modeling and Choice of Distributions**
   - Initial choice of the model,
   - Selection of a suitable density family $f_k(\,\cdot\,;\,\omega_k)$,
   - Choice to be made according to the problem/data studied,

   $\rightsquigarrow$ Collection of models: One model for each fixed number of classes $K \in \mathbb{N}^\star$:

$$\mathcal{M}_K \ := \left\{ x \in \mathbb{R}^d \ \mapsto \ q_K(x\,;\,\theta) = \sum_{k=1}^{K} \alpha_k \, f_k\Big(\,x\,;\omega_k\Big) \right\} .$$

2. **Parameters Estimation**
   - In each $\mathcal{M}_K$ model, we identify the mixture that best fits the data: $q_K(\,\cdot\,;\,\hat{\theta})$.

   $\rightsquigarrow$ Parameter estimation algorithm.

3. **Maximum a Posteriori & Classification**
   - Maximum A Posteriori (MAP) rule to derive a classification of the data.

4. **Model Selection Criteria**

Steps to Define a Mixture Model

1. **Modeling and Choice of Distributions**
   - Initial choice of the model,
   - Selection of a suitable density family $f_k(\cdot\,;\omega_k)$,
   - Choice to be made according to the problem/data studied,

   $\rightsquigarrow$ Collection of models: One model for each fixed number of classes $K \in \mathbb{N}^\star$:

$$\mathcal{M}_K := \left\{ x \in \mathbb{R}^d \mapsto q_K(x\,;\theta) = \sum_{k=1}^{K} \alpha_k\, f_k\Big(x\,;\omega_k\Big) \right\}.$$

2. **Parameters Estimation**
   - In each $\mathcal{M}_K$ model, we identify the mixture that best fits the data: $q_K(\cdot\,;\hat{\theta})$.

   $\rightsquigarrow$ Parameter estimation algorithm.

3. **Maximum a Posteriori & Classification**
   - Maximum A Posteriori (MAP) rule to derive a classification of the data.

4. **Model Selection Criteria**
   - Choose the "best" mixture model among $q_2(\cdot\,;\hat{\theta})$, $q_3(\cdot\,;\hat{\theta})$,. . . , $q_{K_{\mathsf{max}}}(\cdot\,;\hat{\theta})$

   $\rightsquigarrow$ Model selection criterion to determine $\hat{K}$ and thus choose $q_{\hat{K}}(\cdot\,;\hat{\theta})$.

Finite Mixture Models

1.1 General Principle

1.2 Step 1: Modeling and Choice of Distributions

1.3 Step 2: Parameters Estimation

1.4 Step 3: Maximum a Posteriori & Classification

1.5 Step 4: Model Selection Criteria

Step 1: Modeling and Choice of Distributions

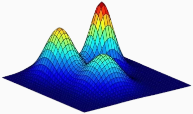
- **Aim**: Define a collection of models $(\mathcal{M}_K)_{K \in \mathbb{R}^\star}$, one model for each fixed number of classes $K \in \mathbb{N}^\star$:
  - A priori modeling, made according to the problem/data studied,
  - Selection of a suitable density family $f_k(\cdot\,;\,\omega_k)$,
  - $\mathcal{M}_K \equiv K$-class (parametric) models, with K fixed:

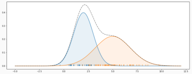
$$\forall K \in \mathbb{N}^\star\,, \quad \mathcal{M}_K := \left\{ \theta \,\middle|\, x \in \mathbb{R}^d \mapsto q_K(x\,;\,\theta) = \sum_{k=1}^{K} \alpha_k\, f_k\Big(\, x\,;\omega_k \Big) \right\}\,.$$

Step 1: Modeling and Choice of Distributions

- **Aim**: Define a collection of models $(\mathcal{M}_K)_{K \in \mathbb{R}^\star}$, one model for each fixed number of classes $K \in \mathbb{N}^\star$:
  - A priori modeling, made according to the problem/data studied,
  - Selection of a suitable density family $f_k(\cdot\,;\omega_k)$,
  - $\mathcal{M}_K \equiv K$-class (parametric) models, with K fixed:

$$\forall K \in \mathbb{N}^\star\,,\quad \mathcal{M}_K := \left\{ \theta \,\middle|\, x \in \mathbb{R}^d \mapsto q_K(x\,;\theta) = \sum_{k=1}^{K} \alpha_k\, f_k\Big(x\,;\omega_k\Big) \right\}.$$





- **Quantitative data**:
  - Gaussian mixtures,
  - Student mixtures, *etc.*
- **Counting data**:
  - Poisson mixtures,
  - Negative binomials, *etc.*
- **Qualitative data**:
  - Mixtures of multinomials, *etc.*
- **Compositional data**:
  - Dirichlet mixtures, *etc.*

Finite Mixture Models

1.1 General Principle

1.2 Step 1: Modeling and Choice of Distributions

1.3 Step 2: Parameters Estimation

1.4 Step 3: Maximum a Posteriori & Classification

1.5 Step 4: Model Selection Criteria

Step 2: Parameters Estimation – Maximum Likelihood Estimation

Assume a fixed number of classes $K \in \mathbb{N}^*$. ***Shortcut**: Let us note q for q_K.*

- **Aim**: Maximize in $\theta = (\alpha_k, \omega_k)_{k \in [\![1,K]\!]}$ the log-likelihood ℓ of the model.

  (i) Recall that: $\forall i \in [\![1,n]\!]$, $q(x_i\,;\,\theta) = \sum_{k=1}^{K} \alpha_k\, f_k\big(x_i\,;\omega_k\big)$

  (ii) Hence, $\boxed{\log q(x\,;\,\theta) = \sum_{i=1}^{n} \log\left[\sum_{k=1}^{K} \alpha_k\, f_k\big(x_i\,;\omega_k\big)\right]}$

Step 2: Parameters Estimation – Maximum Likelihood Estimation

Assume a fixed number of classes $K \in \mathbb{N}^*$. ***Shortcut**: Let us note q for q_K.*

- **Aim**: Maximize in $\theta = (\alpha_k, \omega_k)_{k \in [\![1,K]\!]}$ the log-likelihood ℓ of the model.

  (i) Recall that: $\forall i \in [\![1,n]\!]$, $q(x_i\,;\theta) = \sum_{k=1}^{K} \alpha_k\, f_k\big(x_i\,;\omega_k\big)$

  (ii) Hence, $\boxed{\log q(x\,;\theta) = \sum_{i=1}^{n} \log\left[\sum_{k=1}^{K} \alpha_k\, f_k\big(x_i\,;\omega_k\big)\right]}$

- **Challenge**: This maximization problem usually has no analytical solution.

Step 2: Parameters Estimation – Maximum Likelihood Estimation

Assume a fixed number of classes $K \in \mathbb{N}^*$. ***Shortcut**: Let us note q for q_K.*

- **Aim**: Maximize in $\theta = (\alpha_k, \omega_k)_{k \in [\![1,K]\!]}$ the log-likelihood ℓ of the model.

  (i) Recall that: $\forall i \in [\![1, n]\!]$, $q(x_i\,;\theta) = \sum_{k=1}^{K} \alpha_k\, f_k\big(x_i\,;\omega_k\big)$

  (ii) Hence, $\boxed{\log q(x\,;\theta) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \alpha_k\, f_k\big(x_i\,;\omega_k\big) \right]}$

- **Challenge**: This maximization problem usually has no analytical solution.

  $\rightsquigarrow$ *Optimization algorithm to approximate $\hat{\theta}_{MLE}$.*

  $\Longrightarrow$ Expectation-Maximization (EM) algorithm:
  A core tool for estimation in latent variable models.

Step 2: Parameters Estimation – The EM Algorithm [Dempster et al., 1977]

Assume a fixed number of classes $K \in \mathbb{N}^*$. ***Shortcut***: *Let us note q for q_K.*

- **Aim**: Given a latent variables model, find the MLE: $\boxed{\hat{\theta}_{MLE} \in \underset{\theta \in \Theta}{\operatorname{argmax}}\, q(x\,;\,\theta)}$

Observations: $X \,|\, x = (x_i)_{i \in [\![1,n]\!]} \in \mathcal{X}$.

Latent variables: $Z \,|\, z = (z_i)_{i \in [\![1,n]\!]} \in \mathcal{Z}$,

Here $z_i \in [\![1, K]\!]$ denotes the class of the i-th individual.

Parameter: $\theta \in \Theta$, where Θ set of admissible parameters,

Here $\theta = (\alpha_k, \omega_k)_{k \in [\![1,K]\!]}$ and $\Theta = \left\{[0,1]^K \times \Omega_K \middle| \sum_{k=1}^K \alpha_k = 1\right\}$.

Step 2: Parameters Estimation – The EM Algorithm [Dempster et al., 1977]

Assume a fixed number of classes $K \in \mathbb{N}^*$. ***Shortcut**: Let us note q for q_K.*

- **Aim**: Given a latent variables model, find the MLE: $\boxed{\hat{\theta}_{MLE} \in \underset{\theta \in \Theta}{\operatorname{argmax}}\, q(x\,;\,\theta)}$

  Observations: $X \,|\, x = (x_i)_{i \in [\![1,n]\!]} \in \mathcal{X}$.

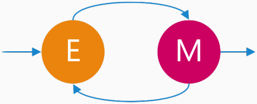
  Latent variables: $Z \,|\, z = (z_i)_{i \in [\![1,n]\!]} \in \mathcal{Z}$,

  Here $z_i \in [\![1, K]\!]$ denotes the class of the i-th individual.

  Parameter: $\theta \in \Theta$, where Θ set of admissible parameters,

  Here $\theta = (\alpha_k, \omega_k)_{k \in [\![1,K]\!]}$ and $\Theta = \left\{[0,1]^K \times \Omega_K \middle| \sum_{k=1}^K \alpha_k = 1\right\}$.

- **Principle**: Alternation of an *Expectation step* and a *Maximization step* until convergence.



Denote $\theta^{(t)}$ the current value of the θ parameter.

Step 2: Parameters Estimation – The EM Algorithm [Dempster et al., 1977]

E-step: Compute the conditional expected log-likelihood

$$Q(\theta|\theta^{(t)}) = \int_{\mathcal{Z}} \log q\big(x, z\,;\,\theta\big)\, q\big(z|x\,;\,\theta^{(t)}\big)\,\mathrm{d}z$$

$$= \mathbb{E}_{Z\sim q(\,\cdot\,|x\,;\,\theta^{(t)})}\left[\log q\big(x, Z\,;\,\theta\big)\right] = \mathbb{E}\left[\log q\big(x, Z\,;\,\theta\big)\Big|x\,;\,\theta^{(t)})\right] .$$

M-step: Maximize $Q(\,\cdot\,|\theta^{(t)})$ in the feasible set Θ: $\theta^{(t+1)} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\, Q(\theta|\theta^{(t)})$

Step 2: Parameters Estimation – The EM Algorithm [Dempster et al., 1977]

E-step: Compute the conditional expected log-likelihood

$$Q(\theta|\theta^{(t)}) = \int_{\mathcal{Z}} \log q\big(x, z\,;\,\theta\big)\, q\big(z|x\,;\,\theta^{(t)}\big)\,\mathrm{d}z$$

$$= \mathbb{E}_{Z\sim q(\,\cdot\,|x\,;\,\theta^{(t)})}\left[\log q\big(x, Z\,;\,\theta\big)\right] = \mathbb{E}\left[\log q\big(x, Z\,;\,\theta\big)\big|x\,;\,\theta^{(t)}\big)\right].$$

Here: $\boxed{Q(\theta|\theta^{(t)}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\left[\log(\alpha_k) + \log\left(f_k(x_i\,;\,\omega_k)\right)\right]\tau_{ik}^{(t)}}$

where $\tau_{ik}^{(t)} = \mathbb{P}\left(Z_i = k\,|\,x_i\,;\,\theta^{(t)}\right) = \dfrac{\alpha_k^{(t)} f_k(x_i\,;\,\omega_k^{(t)})}{\sum_{\ell=1}^{K}\alpha_\ell^{(t)} f_\ell(x_i\,;\,\omega_\ell^{(t)})}$.

M-step: Maximize $Q(\,\cdot\,|\theta^{(t)})$ in the feasible set Θ: $\theta^{(t+1)} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\, Q(\theta|\theta^{(t)})$

Step 2: Parameters Estimation – The EM Algorithm [Dempster et al., 1977]

E-step: Compute the conditional expected log-likelihood

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= \int_{\mathcal{Z}} \log q\big(x, z\,;\,\theta\big)\, q\big(z|x\,;\,\theta^{(t)}\big)\,\mathrm{d}z \\ &= \mathbb{E}_{Z\sim q(\,\cdot\,|x\,;\,\theta^{(t)})}\left[\log q\big(x, Z\,;\,\theta\big)\right] = \mathbb{E}\left[\log q\big(x, Z\,;\,\theta\big)\Big|x\,;\,\theta^{(t)}\big)\right] . \end{aligned}$$

Here: $$\boxed{Q(\theta|\theta^{(t)}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\Big[\log(\alpha_k) + \log\big(f_k(x_i\,;\,\omega_k)\big)\Big]\,\tau_{ik}^{(t)}}$$

where $\tau_{ik}^{(t)} = \mathbb{P}\left(Z_i = k\,|\,x_i\,;\,\theta^{(t)}\right) = \dfrac{\alpha_k^{(t)} f_k(x_i\,;\,\omega_k^{(t)})}{\sum_{\ell=1}^{K}\alpha_\ell^{(t)} f_\ell(x_i\,;\,\omega_\ell^{(t)})}$.

M-step: Maximize $Q(\,\cdot\,|\theta^{(t)})$ in the feasible set Θ: $\theta^{(t+1)} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\, Q(\theta|\theta^{(t)})$

Here: $$\boxed{\begin{cases} \forall k \in [\![1, K]\!], \quad \alpha_k^{(t+1)} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\tau_{ik}^{(t)} \\ (\omega_k^{(t+1)})_{k\in[\![1,K]\!]} \in \underset{\omega=(\omega_k)\in\Omega}{\operatorname{argmax}} \displaystyle\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(t)}\log\big(f_k(x_i\,;\,\omega_k)\big) \end{cases}}$$

The EM Algorithm [Dempster et al., 1977, Delyon et al., 1999]

The EM algorithm

E-step: Conditional expected log-likelihood

$$Q(\theta|\theta^{(t)}) = \mathbb{E}\left[\log q\big(x, Z\,;\,\theta\big)\Big|x\,;\,\theta^{(t)}\right]\;;$$

M-step: Maximize $Q(\,\cdot\,|\theta^{(t)})$ in Θ:

$$\theta^{(t+1)} \in \operatorname*{argmax}_{\theta\in\Theta} Q(\theta|\theta^{(t)})\,.$$

The EM Algorithm [Dempster et al., 1977, Delyon et al., 1999]

The EM algorithm

E-step: Conditional expected log-likelihood

$$Q(\theta|\theta^{(t)}) = \mathbb{E}\left[\log q\big(x, Z\,;\,\theta\big)\Big|x\,;\,\theta^{(t)}\right]\,;$$

M-step: Maximize $Q(\,\cdot\,|\theta^{(t)})$ in Θ:

$$\theta^{(t+1)} \in \operatorname*{argmax}_{\theta\in\Theta} Q(\theta|\theta^{(t)})\,.$$

Convergence of the EM algorithm

Suppose that:

- The f_k laws belong to the exponential family, and are sufficiently regular,
- We can conduct the M-step at each iteration.

1. At every iteration of the EM algorithm, the log likelihood increases.
2. The EM algorithm converges, but not necessarily to the global maximum of the log-likelihood, nor necessarily in finite time.

The EM Algorithm [Dempster et al., 1977, Delyon et al., 1999]

The EM algorithm

E-step: Conditional expected log-likelihood

$$Q(\theta|\theta^{(t)}) = \mathbb{E}\left[\log q\Big(x, Z\,;\,\theta\Big)\Big|x\,;\,\theta^{(t)}\right]\,;$$

M-step: Maximize $Q(\,\cdot\,|\theta^{(t)})$ in Θ:

$$\theta^{(t+1)} \in \operatorname*{argmax}_{\theta\in\Theta} Q(\theta|\theta^{(t)})\,.$$

In practice:

- Easy to implement,
- Sometimes slow to converge (especially when components are very mixed),
- Sensitive to the initialization, *i.e.* to the choice of $\theta^{(0)}$

Convergence of the EM algorithm

Suppose that:

- The f_k laws belong to the exponential family, and are sufficiently regular,
- We can conduct the M-step at each iteration.

1. At every iteration of the EM algorithm, the log likelihood increases.
2. The EM algorithm converges, but not necessarily to the global maximum of the log-likelihood, nor necessarily in finite time.

Variants of the EM Algorithm

1. **Speeding-up** the EM Algorithm.

Variants of the EM Algorithm

1. **Speeding-up** the EM Algorithm.

2. Limitations concerning the **M-step**.

   GEM: Generalized EM Algorithm [Delyon et al., 1999, Lange, 1995]

Variants of the EM Algorithm

1. **Speeding-up** the EM Algorithm.

2. Limitations concerning the **M-step**.

   GEM: Generalized EM Algorithm [Delyon et al., 1999, Lange, 1995]

3. Limitations concerning the **E-step**.

   SEM: Stochastic EM Algorithm [Celeux et al., 1996]

   MCEM: Monte-Carlo EM Algorithm [Wei and Tanner, 1990]

   SAEM: Stochastic-Approximation EM Algorithm [Delyon et al., 1999]

Variants of the EM Algorithm

1. **Speeding-up** the EM Algorithm.

2. Limitations concerning the **M-step**.

   GEM: Generalized EM Algorithm [Delyon et al., 1999, Lange, 1995]

3. Limitations concerning the **E-step**.

   SEM: Stochastic EM Algorithm [Celeux et al., 1996]

   MCEM: Monte-Carlo EM Algorithm [Wei and Tanner, 1990]

   SAEM: Stochastic-Approximation EM Algorithm [Delyon et al., 1999]

SEM – Stochastic EM

S-step: Draw **an** un observed sample

$$z^{(t)} \sim q\Big(\cdot |x; \theta^{(t)}\Big)$$

"E"-step: Estim. of $Q(\cdot|\theta^{(t)})$

$$Q_t(\theta) = \log q(x, z^{(t)}; \theta)$$

M-step: Maximize Q_{t+1}:

$$\theta^{(t+1)} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\, Q_{t+1}(\theta)$$

MCEM – Monte Carlo EM

S-step: Draw m samples

$$z_j^{(t)} \sim q\Big(\cdot |x; \theta^{(t)}\Big)$$

"E"-step: Monte-Carlo estim.

$$Q_t(\theta) = \frac{1}{m} \sum_{j=1}^{m} \log q(x, z_j^{(t)}; \theta)$$

M-step: Maximize Q_{t+1}:

$$\theta^{(t+1)} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\, Q_{t+1}(\theta)$$

Variants of the EM Algorithm

1. **Speeding-up** the EM Algorithm.

2. Limitations concerning the **M-step**.

   GEM: Generalized EM Algorithm [Delyon et al., 1999, Lange, 1995]

3. Limitations concerning the **E-step**.

   SEM: Stochastic EM Algorithm [Celeux et al., 1996]

   MCEM: Monte-Carlo EM Algorithm [Wei and Tanner, 1990]

   SAEM: Stochastic-Approximation EM Algorithm [Delyon et al., 1999]

SEM – Stochastic EM

S-step: Draw **an** un observed sample

$$z^{(t)} \sim q\Big(\cdot |x; \theta^{(t)}\Big)$$

"E"-step: Estim. of $Q(\cdot|\theta^{(t)})$

$$Q_t(\theta) = \log q(x, z^{(t)}; \theta)$$

M-step: Maximize Q_{t+1}:

$$\theta^{(t+1)} \in \operatorname*{argmax}_{\theta\in\Theta} Q_{t+1}(\theta)$$

MCEM – Monte Carlo EM

S-step: Draw m samples

$$z_j^{(t)} \sim q\Big(\cdot |x; \theta^{(t)}\Big)$$

"E"-step: Monte-Carlo estim.

$$Q_t(\theta) = \frac{1}{m} \sum_{j=1}^{m} \log q(x, z_j^{(t)}; \theta)$$

M-step: Maximize Q_{t+1}:

$$\theta^{(t+1)} \in \operatorname*{argmax}_{\theta\in\Theta} Q_{t+1}(\theta)$$

SAEM – Stochastic Approx.

S-step: Draw **a** sample

$$z^{(t)} \sim q\Big(\cdot |x; \theta^{(t)}\Big)$$

SA-step: Update $Q_t(\theta)$ as

$$Q_{t+1}(\theta) = Q_t(\theta) + \gamma_t \Big(\log q(x, z^{(t)}; \theta) - Q_t(\theta)\Big)$$

M-step: Maximize Q_{t+1}:

$$\theta^{(t+1)} \in \operatorname*{argmax}_{\theta\in\Theta} Q_{t+1}(\theta)$$

Initialization

- EM type algorithms are sensitive to initialization.

- Search/Run/Select Strategy:
  - Choice of M initial positions,
  - Some iterations of the algorithm for each position,
  - Selection of the position with the highest likelihood.

- Initialize on the output of a k-means.

- Stochastic variants of the EM.

- More complicated strategies, *etc.*

Finite Mixture Models

1.1 General Principle

1.2 Step 1: Modeling and Choice of Distributions

1.3 Step 2: Parameters Estimation

1.4 Step 3: Maximum a Posteriori & Classification

1.5 Step 4: Model Selection Criteria

Step 3: Maximum a Posteriori & Classification

- **Principle**: Each individual is assigned to the class to which it has the highest probability of belonging, given the estimated parameter $\hat{\theta}_{MLE}$.

Step 3: Maximum a Posteriori & Classification

- **Principle**: Each individual is assigned to the class to which it has the highest probability of belonging, given the estimated parameter $\hat{\theta}_{MLE}$.

- **Posterior distribution**: Probability for an individual $i \in [\![1, n]\!]$ to belong to the class $\mathcal{C}_k$, given the parameter θ :

$$\tau_{ik}(\theta) = \mathbb{P}\left(Z_i = k | x_i\,;\,\theta\right)$$

Step 3: Maximum a Posteriori & Classification

- **Principle**: Each individual is assigned to the class to which it has the highest probability of belonging, given the estimated parameter $\hat{\theta}_{MLE}$.

- **Posterior distribution**: Probability for an individual $i \in [\![1, n]\!]$ to belong to the class $\mathcal{C}_k$, given the parameter θ :

$$\tau_{ik}(\theta) = \mathbb{P}\left(Z_i = k | x_i\,;\,\theta\right)$$

According to Bayes' rule, $\tau_{ik}(\theta) \propto \mathbb{P}\left(Z_i = k\,;\,\theta\right) \times q\left(x_i \,|\, \{Z_i = k\}\,;\,\theta\right)$

Step 3: Maximum a Posteriori & Classification

- **Principle**: Each individual is assigned to the class to which it has the highest probability of belonging, given the estimated parameter $\hat{\theta}_{MLE}$.

- **Posterior distribution**: Probability for an individual $i \in [\![1, n]\!]$ to belong to the class $\mathcal{C}_k$, given the parameter θ :

$$\tau_{ik}(\theta) = \mathbb{P}\left(Z_i = k | x_i \, ; \, \theta\right)$$

According to Bayes' rule, $\tau_{ik}(\theta) \propto \underbrace{\mathbb{P}\left(Z_i = k \, ; \, \theta\right)}_{\alpha_k} \times \underbrace{q\left(x_i \,|\, \{Z_i = k\} \, ; \, \theta\right)}_{f_k(x_i \, ; \, \omega_k)}$

Hence: $\boxed{\tau_{ik}(\theta) = \dfrac{\alpha_k \, f_k(x_i \, ; \, \omega_k)}{\sum_{\ell=1}^{K} \alpha_\ell \, f_\ell(x_i \, ; \, \omega_\ell)}}$

Step 3: Maximum a Posteriori & Classification

- **Principle**: Each individual is assigned to the class to which it has the highest probability of belonging, given the estimated parameter $\hat{\theta}_{MLE}$.

- **Posterior distribution**: Probability for an individual $i \in [\![1, n]\!]$ to belong to the class $\mathcal{C}_k$, given the parameter θ :

$$\tau_{ik}(\theta) = \mathbb{P}\left(Z_i = k | x_i \,;\, \theta\right)$$

According to Bayes' rule, $\tau_{ik}(\theta) \propto \underbrace{\mathbb{P}\left(Z_i = k \,;\, \theta\right)}_{\alpha_k} \times \underbrace{q\left(x_i \,|\, \{Z_i = k\} \,;\, \theta\right)}_{f_k(x_i\,;\,\omega_k)}$

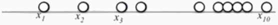
Hence: $\boxed{\tau_{ik}(\theta) = \dfrac{\alpha_k\, f_k(x_i\,;\,\omega_k)}{\sum_{\ell=1}^{K} \alpha_\ell\, f_\ell(x_i\,;\,\omega_\ell)}}$

- **Maximum a Posteriori**: Let the estimate $\hat{\theta}_{MLE}$ of the parameter.

$$i \in \mathcal{C}_k \qquad \textit{iff} \qquad \forall \ell \neq k\,, \quad \tau_{ik}(\hat{\theta}_{MLE}) > \tau_{i\ell}(\hat{\theta}_{MLE})$$

Example: One-dimensional Gaussian Mixture Model

Images from Victor Lavrenko.

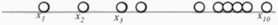


Likelihood: $\theta = \left(\alpha, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right)$

$$\begin{cases} q(x\,;\,\theta) \;=\; \alpha\,\phi(x\,;\,\mu_1,\sigma_1^2) \;+\; (1-\alpha)\,\phi(x\,;\,\mu_2,\sigma_2^2) \\ \phi(x\,;\,\mu,\sigma^2) \;=\; \dfrac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\dfrac{(x-\mu)^2}{2\sigma^2}\right) \end{cases}$$

Example: One-dimensional Gaussian Mixture Model

Images from Victor Lavrenko.



Likelihood: $\theta = \left(\alpha, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right)$

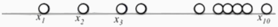
$$\begin{cases} q(x\,;\,\theta) = \alpha\,\phi(x\,;\,\mu_1,\sigma_1^2) + (1-\alpha)\,\phi(x\,;\,\mu_2,\sigma_2^2) \\ \phi(x\,;\,\mu,\sigma^2) = \dfrac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\dfrac{(x-\mu)^2}{2\sigma^2}\right) \end{cases}$$

E-step: $Q(\theta|\theta^{(t)}) \quad \longleftrightarrow \quad \begin{cases} \tau_{i1}^{(t)} \\ \tau_{i2}^{(t)} = 1 - \tau_{i1}^{(t)} \end{cases}$

$$\tau_{i1}^{(t)} = \frac{\alpha^{(t)}\,\phi(x;\mu_1^{(t)},\sigma_1^{2(t)})}{\alpha^{(t)}\,\phi(x;\mu_1^{(t)},\sigma_1^{2(t)}) + (1-\alpha^{(t)})\,\phi(x;\mu_2^{(t)},\sigma_2^{2(t)})}$$

Example: One-dimensional Gaussian Mixture Model

Images from Victor Lavrenko.



Likelihood: $\theta = \left(\alpha, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right)$

$$\begin{cases} q(x\,;\,\theta) \;=\; \alpha\,\phi(x\,;\,\mu_1, \sigma_1^2) \;+\; (1-\alpha)\,\phi(x\,;\,\mu_2, \sigma_2^2) \\ \phi(x\,;\,\mu, \sigma^2) \;=\; \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\dfrac{(x-\mu)^2}{2\sigma^2}\right) \end{cases}$$

E-step: $Q(\theta|\theta^{(t)}) \quad \longleftrightarrow \quad \begin{cases} \tau_{i1}^{(t)} \\ \tau_{i2}^{(t)} = 1 - \tau_{i1}^{(t)} \end{cases}$

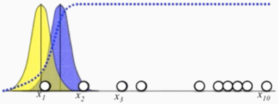
$$\tau_{i1}^{(t)} \;=\; \frac{\alpha^{(t)}\,\phi(x; \mu_1^{(t)}, \sigma_1^{2(t)})}{\alpha^{(t)}\,\phi(x; \mu_1^{(t)}, \sigma_1^{2(t)}) + (1-\alpha^{(t)})\,\phi(x; \mu_2^{(t)}, \sigma_2^{2(t)})}$$

M-step: $\alpha^{(t+1)} \;=\; \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \tau_{i1}^{(t)}$

$$\forall k \in \{1,2\}\,, \quad \mu_k^{(t+1)} \;=\; \frac{\sum_{i=1}^n \tau_{ik}^{(t)} x_i}{\sum_{i=1}^n \tau_{ik}^{(t)}} \quad \& \quad \sigma_k^{2(t+1)} \;=\; \frac{\sum_{i=1}^n \tau_{ik}^{(t)} \left(x_i - \mu_k^{(t+1)}\right)^2}{\sum_{i=1}^n \tau_{ik}^{(t)}}$$

Example: One-dimensional Gaussian Mixture Model

Images from Victor Lavrenko.



Likelihood: $\theta = \left(\alpha, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right)$

$$\begin{cases} q(x\,;\,\theta) = \alpha\,\phi(x\,;\,\mu_1, \sigma_1^2) + (1-\alpha)\,\phi(x\,;\,\mu_2, \sigma_2^2) \\ \phi(x\,;\,\mu, \sigma^2) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\dfrac{(x-\mu)^2}{2\sigma^2}\right) \end{cases}$$

E-step: $Q(\theta|\theta^{(t)}) \quad \longleftrightarrow \quad \begin{cases} \tau_{i1}^{(t)} \\ \tau_{i2}^{(t)} = 1 - \tau_{i1}^{(t)} \end{cases}$

$$\tau_{i1}^{(t)} = \frac{\alpha^{(t)}\,\phi(x; \mu_1^{(t)}, \sigma_1^{2(t)})}{\alpha^{(t)}\,\phi(x; \mu_1^{(t)}, \sigma_1^{2(t)}) + (1-\alpha^{(t)})\,\phi(x; \mu_2^{(t)}, \sigma_2^{2(t)})}$$

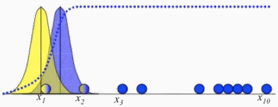
M-step: $\alpha^{(t+1)} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \tau_{i1}^{(t)}$

$$\forall k \in \{1, 2\}\,, \quad \mu_k^{(t+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(t)} x_i}{\sum_{i=1}^{n} \tau_{ik}^{(t)}} \quad \& \quad \sigma_k^{2(t+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(t)} \left(x_i - \mu_k^{(t+1)}\right)^2}{\sum_{i=1}^{n} \tau_{ik}^{(t)}}$$

Example: One-dimensional Gaussian Mixture Model

Likelihood: $\theta = \left(\alpha, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right)$

Images from Victor Lavrenko.



$$\begin{cases} q(x\,;\,\theta) \;=\; \alpha\,\phi(x\,;\,\mu_1,\sigma_1^2) \;+\; (1-\alpha)\,\phi(x\,;\,\mu_2,\sigma_2^2) \\ \phi(x\,;\,\mu,\sigma^2) \;=\; \dfrac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\dfrac{(x-\mu)^2}{2\sigma^2}\right) \end{cases}$$

E-step: $Q(\theta|\theta^{(t)}) \quad \longleftrightarrow \quad \begin{cases} \tau_{i1}^{(t)} \\ \tau_{i2}^{(t)} = 1 - \tau_{i1}^{(t)} \end{cases}$

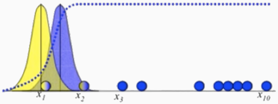
$$\tau_{i1}^{(t)} \;=\; \frac{\alpha^{(t)}\,\phi(x;\mu_1^{(t)},\sigma_1^{2(t)})}{\alpha^{(t)}\,\phi(x;\mu_1^{(t)},\sigma_1^{2(t)}) + (1-\alpha^{(t)})\,\phi(x;\mu_2^{(t)},\sigma_2^{2(t)})}$$

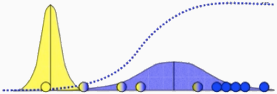
M-step: $\alpha^{(t+1)} \;=\; \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\tau_{i1}^{(t)}$

$$\forall k \in \{1,2\}\,, \quad \mu_k^{(t+1)} \;=\; \frac{\sum_{i=1}^n \tau_{ik}^{(t)} x_i}{\sum_{i=1}^n \tau_{ik}^{(t)}} \quad \& \quad \sigma_k^{2(t+1)} \;=\; \frac{\sum_{i=1}^n \tau_{ik}^{(t)}\left(x_i - \mu_k^{(t+1)}\right)^2}{\sum_{i=1}^n \tau_{ik}^{(t)}}$$

Example: One-dimensional Gaussian Mixture Model

Images from Victor Lavrenko.





Likelihood: $\theta = \left(\alpha, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right)$

$$\begin{cases} q(x\,;\,\theta) \;=\; \alpha\,\phi(x\,;\,\mu_1, \sigma_1^2) \;+\; (1-\alpha)\,\phi(x\,;\,\mu_2, \sigma_2^2) \\ \phi(x\,;\,\mu, \sigma^2) \;=\; \dfrac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\dfrac{(x-\mu)^2}{2\sigma^2}\right) \end{cases}$$

E-step: $Q(\theta|\theta^{(t)}) \quad \longleftrightarrow \quad \begin{cases} \tau_{i1}^{(t)} \\ \tau_{i2}^{(t)} = 1 - \tau_{i1}^{(t)} \end{cases}$

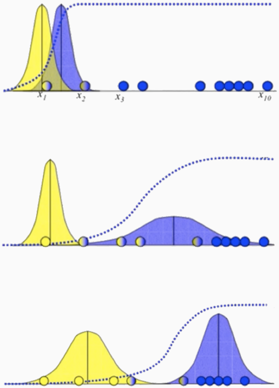
$$\tau_{i1}^{(t)} \;=\; \frac{\alpha^{(t)}\,\phi(x; \mu_1^{(t)}, \sigma_1^{2(t)})}{\alpha^{(t)}\,\phi(x; \mu_1^{(t)}, \sigma_1^{2(t)}) + (1-\alpha^{(t)})\,\phi(x; \mu_2^{(t)}, \sigma_2^{2(t)})}$$

M-step: $\alpha^{(t+1)} \;=\; \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \tau_{i1}^{(t)}$

$$\forall k \in \{1, 2\}\,, \quad \mu_k^{(t+1)} \;=\; \frac{\sum_{i=1}^{n} \tau_{ik}^{(t)} x_i}{\sum_{i=1}^{n} \tau_{ik}^{(t)}} \quad \& \quad \sigma_k^{2(t+1)} \;=\; \frac{\sum_{i=1}^{n} \tau_{ik}^{(t)}\left(x_i - \mu_k^{(t+1)}\right)^2}{\sum_{i=1}^{n} \tau_{ik}^{(t)}}$$

Example: One-dimensional Gaussian Mixture Model

Images from Victor Lavrenko.



Likelihood: $\theta = \left(\alpha, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right)$

$$\begin{cases} q(x\,;\,\theta) = \alpha\,\phi(x\,;\,\mu_1,\sigma_1^2) + (1-\alpha)\,\phi(x\,;\,\mu_2,\sigma_2^2) \\ \phi(x\,;\,\mu,\sigma^2) = \dfrac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\dfrac{(x-\mu)^2}{2\sigma^2}\right) \end{cases}$$

E-step: $Q(\theta|\theta^{(t)}) \quad \longleftrightarrow \quad \begin{cases} \tau_{i1}^{(t)} \\ \tau_{i2}^{(t)} = 1 - \tau_{i1}^{(t)} \end{cases}$

$$\tau_{i1}^{(t)} = \frac{\alpha^{(t)}\,\phi(x;\mu_1^{(t)},\sigma_1^{2(t)})}{\alpha^{(t)}\,\phi(x;\mu_1^{(t)},\sigma_1^{2(t)}) + (1-\alpha^{(t)})\,\phi(x;\mu_2^{(t)},\sigma_2^{2(t)})}$$

M-step: $\alpha^{(t+1)} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\tau_{i1}^{(t)}$

$$\forall k \in \{1,2\}\,, \quad \mu_k^{(t+1)} = \frac{\sum_{i=1}^{n}\tau_{ik}^{(t)}x_i}{\sum_{i=1}^{n}\tau_{ik}^{(t)}} \quad \& \quad \sigma_k^{2(t+1)} = \frac{\sum_{i=1}^{n}\tau_{ik}^{(t)}\left(x_i - \mu_k^{(t+1)}\right)^2}{\sum_{i=1}^{n}\tau_{ik}^{(t)}}$$

Alternative to EM: Classification EM Algorithm [Celeux and Govaert, 1992]

CEM – Classification EM

E-step: Compute the $\tau_{ik}^{(t)}$

C-step: Determination of a partition of the data x by the MAP rule:

$$\hat{z}_i^{(t)} \in \underset{\ell \in [\![1,K]\!]}{\operatorname{argmax}} \, \tau_{i\ell}^{(t)}$$

M-step: We update the parameters by replacing $\tau_{ik}^{(t)}$ by $\hat{z}_i^{(t)}$

- CEM converges in a finite number of iterations, contrary to EM,
- CEM produces biased estimators of the mixture parameters,
- We can prove that the k-means algorithm is a Gaussian mixture model with constant variance (ellipse $\equiv$ circle), estimated by CEM.

Finite Mixture Models

1.1 General Principle

1.2 Step 1: Modeling and Choice of Distributions

1.3 Step 2: Parameters Estimation

1.4 Step 3: Maximum a Posteriori & Classification

1.5 Step 4: Model Selection Criteria

Step 4: Model Selection Criteria

- **Aim**: Determine the optimal number of classes K,

$$\mathcal{M}_K = \left\{ x \in \mathbb{R}^d \mapsto q_K(x\,;\,\theta) = \sum_{k=1}^{K} \alpha_k \, f_k\big(\, x\,;\omega_k\big) \right\}.$$

Step 4: Model Selection Criteria

- **Aim**: Determine the optimal number of classes K,

$$\mathcal{M}_K = \left\{ x \in \mathbb{R}^d \mapsto q_K(x\,;\,\theta) = \sum_{k=1}^{K} \alpha_k\, f_k\big(x\,;\omega_k\big) \right\}.$$

- **Probabilistic model**

  ⇝ We can use the criteria from the information theory:

$$\hat{K} = \underset{K}{\operatorname{argmin}}\ \mathtt{CRIT}(K) = \underset{K}{\operatorname{argmin}} \{-\log q_K(x\,;\,\theta) + pen(K)\}.$$

AIC: Akaike Information Criterion [Akaike et al., 1973]

BIC: Bayesian Information Criterion [Schwarz, 1978]

ICL: Integrated Completed Likelihood [Biernacki et al., 2000]

Step 4: Model Selection Criteria

$$\hat{K} = \underset{K}{\text{argmin}}\ \texttt{CRIT}(K) = \underset{K}{\text{argmin}} \{ -\log q_K(x\,;\,\hat{\theta}_{MLE}) + pen(K) \}\,.$$

AIC

$$\boxed{\begin{aligned}\texttt{AIC}(K) = -\log q_K(x\,;\,\hat{\theta})\\ +\ \nu_k\end{aligned}}$$

- Achieves a bias-variance trade-off
- Asymptotically, the AIC retains the model minimizing the *mean* Kullback deviation with the true unknown law
- In the context of finite mixture models, AIC tends to under-penalize

BIC

$$\boxed{\begin{aligned}\texttt{BIC}(K) = -\log q_K(x\,;\,\hat{\theta})\\ + \frac{\nu_k}{2}\log(n)\end{aligned}}$$

- Asymptotically, BIC selects the model minimizing the Kullback deviation from the true law.

  $\rightsquigarrow$ The BIC is convergent if the true model is in the list of models.

ICL

$$\boxed{\begin{aligned}\texttt{ICL}(K) = -\log q_K(x, \hat{z}\,;\,\hat{\theta})\\ + \frac{\nu_k}{2}\log(n)\end{aligned}}$$

- Where $\hat{z}$ is the MAP of $\hat{\theta}_{MLE}$

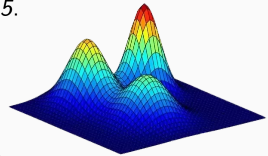
where ν_K is the number of free parameters of the mixtures $\mathcal{M}_K$.

Gaussian Mixture Models

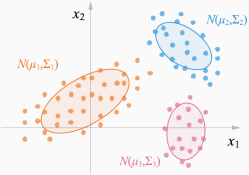
2.1 Multivariate Gaussian Mixtures

Multivariate Gaussian Mixtures

- **Quantitative data**: $x_i \in \mathbb{R}^d$, *Generalization of slide 15.*



- **Likelihood**: $\theta = (\alpha_k, \mu_k, \Sigma_k)_{k \in [\![1, K]\!]}$



$$\begin{cases} q_K(x\,;\,\theta) = \displaystyle\sum_{k=1}^{K} \alpha_k\,\phi\,(x\,;\,\mu_k, \Sigma_k)\ . \\[2ex] \phi\,(x\,;\,\mu, \Sigma) = \dfrac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\dfrac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right)\ . \end{cases}$$

- **Gaussian Mixture Model**:

$$\begin{cases} \Theta_K = \left\{ (\alpha_k, \mu_k, \Sigma_k)_{k \in [\![1, K]\!]} \in \left([0,1] \times \mathbb{R}^d \times \mathcal{S}_d\mathbb{R}\right)^K \,\middle|\, \displaystyle\sum_{k=1}^{K} \alpha_k = 1 \right\} \\[2ex] \mathcal{M}_K = \left\{ \theta \in \Theta_K \,\middle|\, x \in \mathbb{R}^d \mapsto q_K(x\,;\,\theta) \right\} \end{cases}$$

- **Estimation** through the EM algorithm *(See tutorials)*

Estimation under Constraints

- Without constraint, $\dim(\Theta_K) = (K-1) + Kd + K\dfrac{d(d+1)}{2}$
  - $\rightsquigarrow$ In large dimensions, this can lead to an over-parameterized model.
    - $\rightsquigarrow$ Constraints on the type of covariance matrix.

Estimation under Constraints

- Without constraint, $\dim(\Theta_K) = (K-1) + Kd + K\dfrac{d(d+1)}{2}$

  $\rightsquigarrow$ In large dimensions, this can lead to an over-parameterized model.

  $\rightsquigarrow$ Constraints on the type of covariance matrix.

- **Forms of Gaussian mixtures**:

  Eigenvalue decomposition of covariance matrices: $\boxed{\Sigma_k \; = \; L_k D_k A_k D_k^\top}$

  - *Volume:* $L_k = |\Sigma_k|^{1/d}$, *constant, or not.*
  - *Orientation:* D_k matrix of eigenvectors of Σ_k, *constant, or not.*
  - *Form:* A_k diagonal matrix of normalized eigenvectors of Σ_k, *spherical, diagonal or full.*

Estimation under Constraints

- Without constraint, $\dim(\Theta_K) = (K-1) + Kd + K\frac{d(d+1)}{2}$

  ⤳ In large dimensions, this can lead to an over-parameterized model.

  ⤳ Constraints on the type of covariance matrix.

- **Forms of Gaussian mixtures**:

  Eigenvalue decomposition of covariance matrices: $\boxed{\Sigma_k = L_k D_k A_k D_k^\top}$

  - *Volume:* $L_k = |\Sigma_k|^{1/d}$, *constant, or not.*
  - *Orientation:* D_k matrix of eigenvectors of Σ_k, *constant, or not.*
  - *Form:* A_k diagonal matrix of normalized eigenvectors of Σ_k, *spherical, diagonal or full.*

  ⤳ **14** type of covariance matrix:

  - *Spherical:* Equal volume, or not (2),
  - *Diagonal:* Equal volume, or not ; Equal shape, or not (4),
  - *Full:* Equal volume, or not ; Equal shape, or not ; Equal orientation, or not (8).

Estimation under Constraints

- Without constraint, $\dim(\Theta_K) = (K-1) + Kd + K\frac{d(d+1)}{2}$

  ⤳ In large dimensions, this can lead to an over-parameterized model.

  ⤳ Constraints on the type of covariance matrix.

- **Forms of Gaussian mixtures**:
  Eigenvalue decomposition of covariance matrices: $\boxed{\Sigma_k = L_k D_k A_k D_k^\top}$

  - *Volume:* $L_k = |\Sigma_k|^{1/d}$, *constant, or not.*
  - *Orientation:* D_k matrix of eigenvectors of Σ_k, *constant, or not.*
  - *Form:* A_k diagonal matrix of normalized eigenvectors of Σ_k, *spherical, diagonal or full.*

  ⤳ **14** type of covariance matrix:

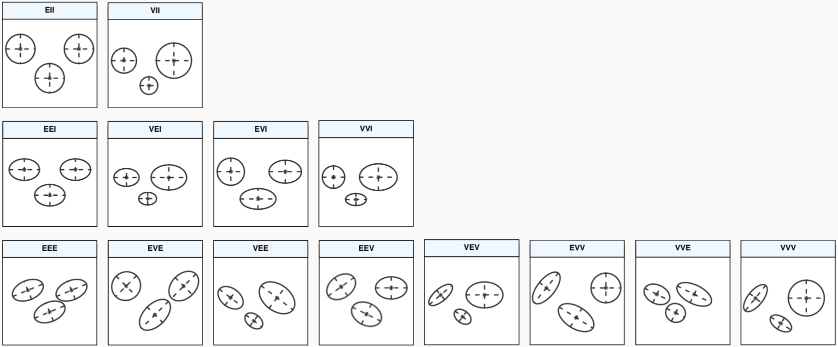
  - *Spherical:* Equal volume, or not (2),
  - *Diagonal:* Equal volume, or not ; Equal shape, or not (4),
  - *Full:* Equal volume, or not ; Equal shape, or not ; Equal orientation, or not (8).

  Proportions α_k assumed equal or free.

  ⤳ **28** possible forms of Gaussian mixtures.

Constraints on the Covariance Matrices

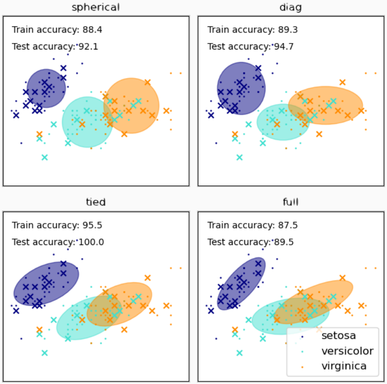
- Nomenclature in R: `modelNames = c("EEE","VEE","EVV","VVV")`.
  (Package `mclust`, page 38 for a description of all types.)

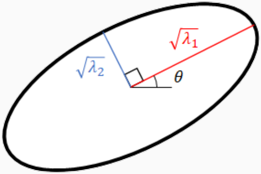


- In `Python`: “`full`”, “`tied`”, “`diag`”, “`spherical`”.
  (Function `sklearn.mixture.GaussianMixture`)

Covariance Matrices vs. Rotation Matrices in $\mathbb{R}^2$

- D: Rotation matrix defined by an angle α: $D = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}$,
- A: Diagonal matrix of diagonal terms b and $1/b$: $A = \begin{pmatrix} b & 0 \\ 0 & \frac{1}{b} \end{pmatrix}$,
- Ellipse of equidensity: $L = \lambda = volume$.



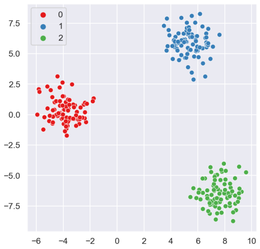


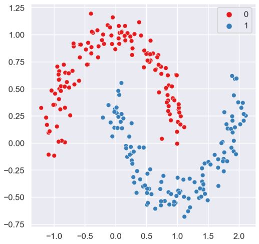
$$\begin{cases} \lambda_1 = \lambda b \\ \lambda_2 = \dfrac{\lambda}{b} \end{cases}$$

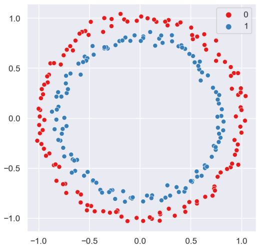
Strengths and Weaknesses

Pros:

- Give probabilistic cluster assignments,
- Have probabilistic interpretation,
- Can handle clusters with varying sizes, variance, *etc.*







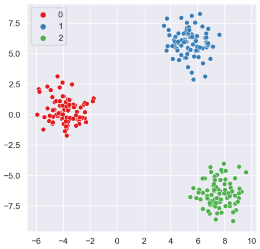
Strengths and Weaknesses

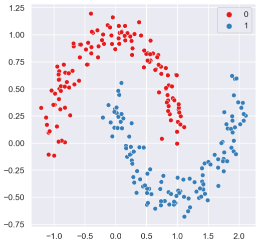
Pros:

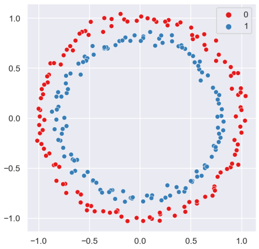
- Give probabilistic cluster assignments,
- Have probabilistic interpretation,
- Can handle clusters with varying sizes, variance, *etc.*

Cons:

- Initialization matters,
- Choose appropriate distributions,
- Overfitting issues.







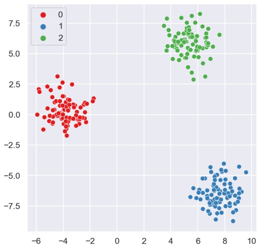
Strengths and Weaknesses

Pros:

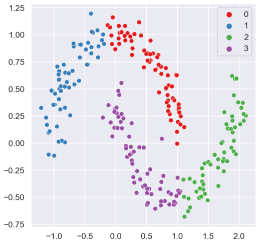
- Give probabilistic cluster assignments,
- Have probabilistic interpretation,
- Can handle clusters with varying sizes, variance, *etc.*

Cons:

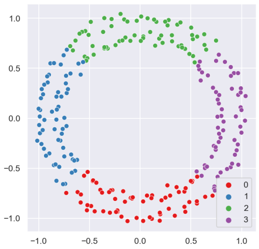
- Initialization matters,
- Choose appropriate distributions,
- Overfitting issues.



Tied, 3 components



Full, 4 components



Diagonal, 4 components

References i

Akaike, H., Petrov, B. N., and Csaki, F. (1973). **Second international symposium on information theory.**

Biernacki, C., Celeux, G., and Govaert, G. (2000). **Assessing a mixture model for clustering with the integrated completed likelihood.** IEEE transactions on pattern analysis and machine intelligence, 22(7):719–725.

Bisson, G. (2000). **La similarité: une notion symbolique/numérique.** Apprentissage symbolique-numérique, 2:169–201.

Celeux, G., Chauveau, D., and Diebolt, J. (1996). **Stochastic versions of the em algorithm: an experimental study in the mixture case.** Journal of statistical computation and simulation, 55(4):287–314.

Celeux, G. and Govaert, G. (1992). **A classification em algorithm for clustering and two stochastic versions.** Computational statistics & Data analysis, 14(3):315–332.

Delyon, B., Lavielle, M., and Moulines, E. (1999). **Convergence of a stochastic approximation version of the em algorithm.** Annals of statistics, pages 94–128.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). **Maximum likelihood from incomplete data via the em algorithm.** Journal of the royal statistical society: series B (methodological), 39(1):1–22.

Janssen, P. (2012). **Cluster analysis to understand socio-ecological systems: a guideline.**

Johnson, S. C. (1967). **Hierarchical clustering schemes.** Psychometrika, 32(3):241–254.

References ii

Lange, K. (1995). **A gradient algorithm locally equivalent to the em algorithm.** Journal of the Royal Statistical Society: Series B (Methodological), 57(2):425–437.

Schwarz, G. (1978). **Estimating the dimension of a model.** The annals of statistics, pages 461–464.

Ward, J. H. (1963). **Hierarchical grouping to optimize an objective function.** Journal of the American statistical association, 58(301):236–244.

Wei, G. C. and Tanner, M. A. (1990). **A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms.** Journal of the American statistical Association, 85(411):699–704.