Expectation Maximization (EM) Algorithm

lecturer: J. Matas, matas@cmp.felk.cvut.cz

authors: O. Drbohlav, J. Matas

Czech Technical University, Faculty of Electrical Engineering
Department of Cybernetics, Center for Machine Perception
121 35 Praha 2, Karlovo nám. 13, Czech Republic

http://cmp.felk.cvut.cz

5/Jan/2018

LECTURE PLAN

- Motivation: Observations with missing values
- Sketch of the algorithm, relation to K-means
- EM algorithm

EM Algorithm

- Used to find maximum likelihood parameters of a statistical model when the equations cannot be directly solved.
- Two typical cases of use:
  - **Missing data**: Some observations are incomplete. E.g. features are vectors in 5-dimensional space $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5) \in \mathbb{R}^D$ but observations have a component missing, e.g.: $(2, 5, \bullet, 1, 2)$ or $(\bullet, \bullet, 1, 4, 2)$, where '$\bullet$' are the unobserved components.
  - **Latent variables**: Observations are complete but the model can be formulated and solved more simply if further variables are introduced to it. A typical example are *mixture models* where for each observed point it is advantageous to introduce a random variable which specifies which component of the mixture generated that point.

Example: Complete Data? Easy (1)

We measure lengths of vehicles. The observation space is two-dimensional, with $x \in \{\text{car}, \text{truck}\}$ capturing vehicle type and $y \in \mathbb{R}$ capturing length.

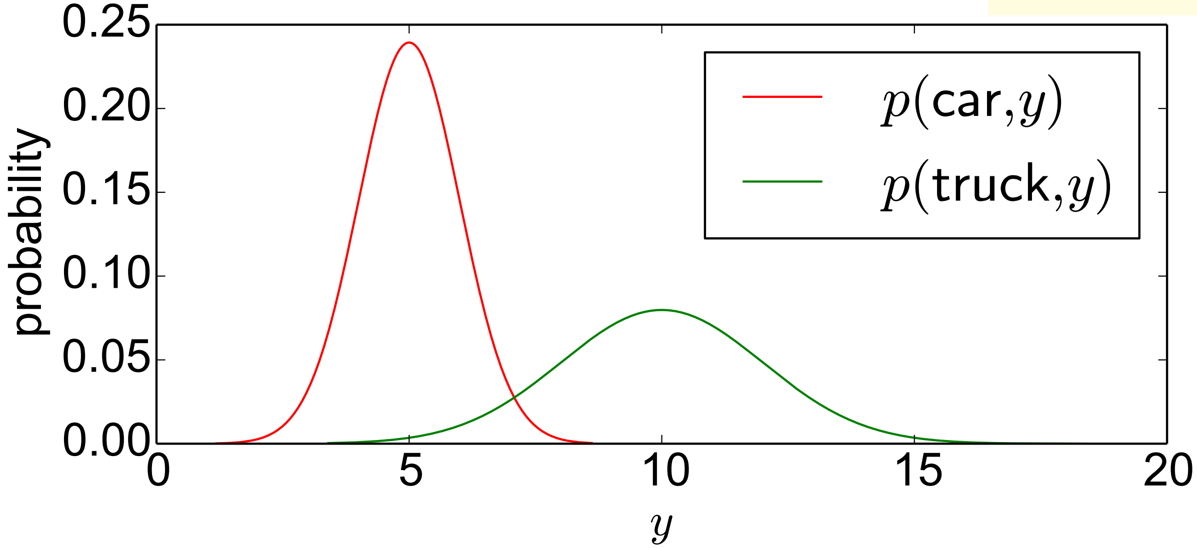
$$p(x, y) : \text{distribution}\,, \qquad x \in \{\text{car}, \text{truck}\}\,, \quad y \in \mathbb{R} \tag{1}$$

$$p(\text{car}, y) = \pi_\text{c} \mathcal{N}\left(y|\mu_\text{c}, \sigma_\text{c} = 1\right) = \kappa_\text{c} \exp\left\{-\frac{1}{2}\left(y - \mu_\text{c}\right)^2\right\}\,,\ (\kappa_\text{c} = \frac{\pi_\text{c}}{\sqrt{2\pi}}) \tag{2}$$

$$p(\text{truck}, y) = \pi_\text{t} \mathcal{N}\left(y|\mu_\text{t}, \sigma_\text{t} = 2\right) = \kappa_\text{t} \exp\left\{-\frac{1}{8}\left(y - \mu_\text{t}\right)^2\right\}\,,\ (\kappa_\text{t} = \frac{\pi_\text{c}}{\sqrt{8\pi}}) \tag{3}$$

Parameters κ_c, κ_t are considered to be known. The **only unknowns** are μ_c and μ_t. We want to recover μ_c and μ_t using Maximum Likelihood.

Example ($\pi_\text{c} = 0.6$, $\pi_\text{t} = 0.4$, $\sigma_\text{c} = 1$, $\sigma_\text{t} = 2$, $\mu_\text{c} = 5$, $\mu_\text{t} = 10$)



Example: Complete Data? Easy (2)

The observations are:

$$\mathcal{T} = \{(x_1, y_1), (x_2, y_2), , ..., (x_N, y_N)\} \tag{4}$$

$$= \{\underbrace{(\mathsf{car}, y_1^{(\mathsf{c})}), (\mathsf{car}, y_2^{(\mathsf{c})}), ..., (\mathsf{car}, y_C^{(\mathsf{c})})}_{C \text{ car observations}}, \underbrace{(\mathsf{truck}, y_1^{(\mathsf{t})}), (\mathsf{truck}, y_2^{(\mathsf{t})}), ..., (\mathsf{truck}, y_T^{(\mathsf{t})})}_{T \text{ truck observations}}\} \tag{5}$$

Log-likelihood $\ell(\mathcal{T}) = \ln p(\mathcal{T} | \mu_{\mathsf{c}}, \mu_{\mathsf{t}})$:

$$\ell(\mathcal{T}) = \sum_{i=1}^{N} \ln p(x_i, y_i | \mu_{\mathsf{c}}, \mu_{\mathsf{t}}) = C \ln \kappa_{\mathsf{c}} - \frac{1}{2} \sum_{i=1}^{C} (y_i^{(\mathsf{c})} - \mu_{\mathsf{c}})^2 + T \ln \kappa_{\mathsf{t}} - \frac{1}{8} \sum_{i=1}^{T} (y_i^{(\mathsf{t})} - \mu_{\mathsf{t}})^2 \tag{6}$$

Estimation of μ_1, μ_2 using ML is very easy:

$$\frac{\partial \ell(\mathcal{T})}{\partial \mu_{\mathsf{c}}} = \sum_{i=1}^{C} (y_i^{(\mathsf{c})} - \mu_{\mathsf{c}}) = 0 \qquad \Rightarrow \qquad \mu_{\mathsf{c}} = \frac{1}{C} \sum_{i=1}^{C} y_i^{(\mathsf{c})} \tag{7}$$

$$\frac{\partial \ell(\mathcal{T})}{\partial \mu_{\mathsf{t}}} = \frac{1}{4} \sum_{i=1}^{T} (y_i^{(\mathsf{t})} - \mu_{\mathsf{t}}) = 0 \qquad \Rightarrow \qquad \mu_{\mathsf{t}} = \frac{1}{T} \sum_{i=1}^{T} y_i^{(\mathsf{t})} \tag{8}$$

Example, Incomplete Data → Difficult (1)

Consider some observations to have the first coordinate **missing** (•):

$$\mathcal{T} = \{(\text{car}, y_1^{(\text{c})}), \ldots, (\text{car}, y_C^{(\text{c})}), (\text{truck}, y_1^{(\text{t})}), \ldots, (\text{truck}, y_T^{(\text{t})}), \underbrace{(\bullet, y_1^{\bullet}), \ldots, (\bullet, y_M^{\bullet})}_{\substack{\text{data with uknown} \\ \text{vehicle type}}}\} \tag{9}$$

What is the probability of observing $y^{\bullet}$?

$$p(y^{\bullet}) = p(\text{car}, y^{\bullet}) + p(\text{truck}, y^{\bullet})$$

Log-likelihood:

$$\ell(\mathcal{T}) = \sum_{i=1}^{N} \ln p(x_i, y_i | \mu_{\text{c}}, \mu_{\text{t}}) = \overbrace{C \ln \kappa_{\text{c}} - \frac{1}{2} \sum_{i=1}^{C} (y_i^{(\text{c})} - \mu_{\text{c}})^2 + T \ln \kappa_{\text{t}} - \frac{1}{8} \sum_{i=1}^{T} (y_i^{(\text{t})} - \mu_{\text{t}})^2}^{\text{same term as before}} \tag{10}$$

$$+ \sum_{i=1}^{M} \ln \left(\kappa_{\text{c}} \exp \left\{ -\frac{1}{2} (y_i^{\bullet} - \mu_{\text{c}})^2 \right\} + \kappa_{\text{t}} \exp \left\{ -\frac{1}{8} (y_i^{\bullet} - \mu_{\text{t}})^2 \right\} \right) \tag{11}$$

Example, Incomplete Data $\rightarrow$ Difficult (2)

Log-likelihood:

$$\ell(\mathcal{T}) = C \ln \kappa_{\mathsf{c}} - \frac{1}{2}\sum_{i=1}^{C}(y_i^{(\mathsf{c})} - \mu_{\mathsf{c}})^2 + T \ln \kappa_{\mathsf{t}} - \frac{1}{8}\sum_{i=1}^{T}(y_i^{(\mathsf{t})} - \mu_{\mathsf{t}})^2 \tag{12}$$

$$+ \sum_{i=1}^{M} \ln \left(\kappa_{\mathsf{c}} \exp\left\{ -\frac{1}{2}\left(y_i^{\bullet} - \mu_{\mathsf{c}}\right)^2 \right\} + \kappa_{\mathsf{t}} \exp\left\{ -\frac{1}{8}\left(y_i^{\bullet} - \mu_{\mathsf{t}}\right)^2 \right\} \right) \tag{13}$$

Optimality condition (shown for μ_{c} only):

$$0 = \frac{\partial \ell(\mathcal{T})}{\partial \mu_{\mathsf{c}}} = \sum_{i=1}^{C}(y_i^{(\mathsf{c})} - \mu_{\mathsf{c}}) \quad + \tag{14}$$

$$+ \quad \sum_{i=1}^{M} \frac{\kappa_{\mathsf{c}} \exp\left\{ -\frac{1}{2}\left(y_i^{\bullet} - \mu_{\mathsf{c}}\right)^2 \right\}}{\kappa_{\mathsf{c}} \exp\left\{ -\frac{1}{2}\left(y_i^{\bullet} - \mu_{\mathsf{c}}\right)^2 \right\} + \kappa_{\mathsf{t}} \exp\left\{ -\frac{1}{8}\left(y_i^{\bullet} - \mu_{\mathsf{t}}\right)^2 \right\}} (y_i^{\bullet} - \mu_{\mathsf{c}}) \tag{15}$$

Missing Values, Optimality Condition

Log-likelihood:

$$\ell(\mathcal{T}) = C \ln \kappa_{\mathsf{c}} - \frac{1}{2}\sum_{i=1}^{C}(y_i^{(\mathsf{c})} - \mu_{\mathsf{c}})^2 + T \ln \kappa_{\mathsf{t}} - \frac{1}{8}\sum_{i=1}^{T}(y_i^{(\mathsf{t})} - \mu_{\mathsf{t}})^2 \tag{16}$$

$$+ \sum_{i=1}^{M} \ln\left(\kappa_{\mathsf{c}} \exp\left\{-\frac{1}{2}\left(y_i^{\bullet} - \mu_{\mathsf{c}}\right)^2\right\} + \kappa_{\mathsf{t}} \exp\left\{-\frac{1}{8}\left(y_i^{\bullet} - \mu_{\mathsf{t}}\right)^2\right\}\right) \tag{17}$$

Optimality condition (shown for μ_{c} only):

$$0 = \frac{\partial \ell(\mathcal{T})}{\partial \mu_{\mathsf{c}}} = \sum_{i=1}^{C}(y_i^{(\mathsf{c})} - \mu_{\mathsf{c}}) \quad + \tag{18}$$

$$+ \quad \sum_{i=1}^{M} \frac{\overbrace{\kappa_{\mathsf{c}} \exp\left\{-\frac{1}{2}\left(y_i^{\bullet} - \mu_{\mathsf{c}}\right)^2\right\}}^{p(\mathsf{car},\, y_i^{\bullet} \,|\, \mu_{\mathsf{c}}, \mu_{\mathsf{t}})}}{\underbrace{\kappa_{\mathsf{c}} \exp\left\{-\frac{1}{2}\left(y_i^{\bullet} - \mu_{\mathsf{c}}\right)^2\right\}}_{p(\mathsf{car},\, y_i^{\bullet} \,|\, \mu_{\mathsf{c}}, \mu_{\mathsf{t}})} + \underbrace{\kappa_{\mathsf{t}} \exp\left\{-\frac{1}{8}\left(y_i^{\bullet} - \mu_{\mathsf{t}}\right)^2\right\}}_{p(\mathsf{truck},\, y_i^{\bullet} \,|\, \mu_{\mathsf{c}}, \mu_{\mathsf{t}})}} \left(y_i^{\bullet} - \mu_{\mathsf{c}}\right) \tag{19}$$

Missing Values, Optimality Condition

Log-likelihood:

$$\ell(\mathcal{T}) = C \ln \kappa_{\mathsf{c}} - \frac{1}{2} \sum_{i=1}^{C} (y_i^{(\mathsf{c})} - \mu_{\mathsf{c}})^2 + T \ln \kappa_{\mathsf{t}} - \frac{1}{8} \sum_{i=1}^{T} (y_i^{(\mathsf{t})} - \mu_{\mathsf{t}})^2 \tag{20}$$

$$+ \sum_{i=1}^{M} \ln \left(\kappa_{\mathsf{c}} \exp \left\{ -\frac{1}{2} \left(y_i^{\bullet} - \mu_{\mathsf{c}}\right)^2 \right\} + \kappa_{\mathsf{t}} \exp \left\{ -\frac{1}{8} \left(y_i^{\bullet} - \mu_{\mathsf{t}}\right)^2 \right\} \right) \tag{21}$$

Optimality condition (shown for μ_{c} only):

$$0 = \frac{\partial \ell(\mathcal{T})}{\partial \mu_{\mathsf{c}}} = \sum_{i=1}^{C} (y_i^{(\mathsf{c})} - \mu_{\mathsf{c}}) \quad + \tag{22}$$

$$+ \quad \sum_{i=1}^{M} \overbrace{\boxed{\frac{\kappa_{\mathsf{c}} \exp \left\{ -\frac{1}{2} \left(y_i^{\bullet} - \mu_{\mathsf{c}}\right)^2 \right\}}{\kappa_{\mathsf{c}} \exp \left\{ -\frac{1}{2} \left(y_i^{\bullet} - \mu_{\mathsf{c}}\right)^2 \right\} + \kappa_{\mathsf{t}} \exp \left\{ -\frac{1}{8} \left(y_i^{\bullet} - \mu_{\mathsf{t}}\right)^2 \right\}}}}^{p(\mathsf{car} | y_i^{\bullet}, \mu_{\mathsf{c}}, \mu_{\mathsf{t}})} (y_i^{\bullet} - \mu_{\mathsf{c}}) \tag{23}$$

Missing Values, Optimality Conditions

Optimality conditions (shown for both μ_{c} and μ_{t}):

$$0 = \frac{\partial \ell(\mathcal{T})}{\partial \mu_{\mathsf{c}}} = \sum_{i=1}^{C} (y_i^{(\mathsf{c})} - \mu_{\mathsf{c}}) \quad + \tag{24}$$

$$+ \quad \sum_{i=1}^{M} \overbrace{\boxed{\frac{\kappa_{\mathsf{c}} \exp\left\{-\frac{1}{2}\left(y_i^{\bullet} - \mu_{\mathsf{c}}\right)^2\right\}}{\kappa_{\mathsf{c}} \exp\left\{-\frac{1}{2}\left(y_i^{\bullet} - \mu_{\mathsf{c}}\right)^2\right\} + \kappa_{\mathsf{t}} \exp\left\{-\frac{1}{8}\left(y_i^{\bullet} - \mu_{\mathsf{t}}\right)^2\right\}}}}^{p(\mathsf{car}|y_i^{\bullet}, \mu_{\mathsf{c}}, \mu_{\mathsf{t}})} (y_i^{\bullet} - \mu_{\mathsf{c}}) \tag{25}$$

$$0 = 4\frac{\partial \ell(\mathcal{T})}{\partial \mu_{\mathsf{t}}} = \sum_{i=1}^{T} (y_i^{(\mathsf{t})} - \mu_{\mathsf{t}}) + \sum_{i=1}^{M} p(\mathsf{truck}|y_i^{\bullet}, \mu_{\mathsf{c}}, \mu_{\mathsf{t}}) \; (y_i^{\bullet} - \mu_{\mathsf{t}}) \tag{26}$$

Note:

- Complicated equations for the uknowns μ_{c}, μ_{t}
- Both equations contain μ_{c} and μ_{t} (cf. case with no missing variables)

Missing Values, EM Approach

Optimality conditions (shown for both μ_{c} and μ_{t}):

$$\sum_{i=1}^{C}(y_i^{(\mathsf{c})} - \mu_{\mathsf{c}}) + \sum_{i=1}^{M} p(\mathsf{car}|y_i^{\bullet}, \mu_{\mathsf{c}}, \mu_{\mathsf{t}})\ (y_i^{\bullet} - \mu_{\mathsf{c}}) = 0 \tag{27}$$

$$\sum_{i=1}^{T}(y_i^{(\mathsf{t})} - \mu_{\mathsf{t}}) + \sum_{i=1}^{M} p(\mathsf{truck}|y_i^{\bullet}, \mu_{\mathsf{c}}, \mu_{\mathsf{t}})\ (y_i^{\bullet} - \mu_{\mathsf{t}}) = 0 \tag{28}$$

If $p(\mathsf{car}|y_i^{\bullet}, \mu_{\mathsf{c}}, \mu_{\mathsf{t}})$ and $p(\mathsf{truck}|y_i^{\bullet}, \mu_{\mathsf{c}}, \mu_{\mathsf{t}})$ **were** known, the estimation would've been easy:

- Let z_i $(i = 1, 2, \ldots, M)$, $z_i \in \{\mathsf{car}, \mathsf{truck}\}$ denote the missing values. Define $q(z_i) = p(z_i|y_i^{\bullet}, \mu_{\mathsf{c}}, \mu_{\mathsf{t}})$
- The equations lead to

$$\sum_{i=1}^{C}(y_i^{(\mathsf{c})} - \mu_{\mathsf{c}}) + \sum_{i=1}^{M} q(z_i = \mathsf{car})\,(y_i^{\bullet} - \mu_{\mathsf{c}}) = 0 \tag{29}$$

$$\Rightarrow \qquad \mu_{\mathsf{c}} = \frac{\sum_{i=1}^{C} y_i^{(\mathsf{c})} + \sum_{i=1}^{M} q(z_i = \mathsf{car})\, y_i^{\bullet}}{C + \sum_{i=1}^{M} q(z_i = \mathsf{car})} \tag{30}$$

and similarly,

$$\mu_{\mathsf{t}} = \frac{\sum_{i=1}^{T} y_i^{(\mathsf{t})} + \sum_{i=1}^{M} q(z_i = \mathsf{truck})\, y_i^{\bullet}}{T + \sum_{i=1}^{M} q(z_i = \mathsf{truck})} \tag{31}$$

Missing Values, EM Approach

$$\mu_{\mathrm{c}} = \frac{\sum_{i=1}^{C} y_i^{(\mathrm{c})} + \sum_{i=1}^{M} q(z_i = \mathrm{car})\, y_i^{\bullet}}{C + \sum_{i=1}^{M} q(z_i = \mathrm{car})} \tag{32}$$

$$\mu_{\mathrm{t}} = \frac{\sum_{i=1}^{T} y_i^{(\mathrm{t})} + \sum_{i=1}^{M} q(z_i = \mathrm{truck})\, y_i^{\bullet}}{T + \sum_{i=1}^{M} q(z_i = \mathrm{truck})} \tag{33}$$

- These expressions are weighted averages of the observed y's. Data with non-missing x have weight 1, the data with missing x have weight $q(z_i)$. How about trying the following procedure for finding the ML estimate of μ_{c} and μ_{t}:

  1. Initialize μ_{c}, μ_{t}
  2. Compute $q(z_i) = p(z_i | y_i^{\bullet}, \mu_{\mathrm{c}}, \mu_{\mathrm{t}})$ for all $i = 1, 2, ..., M$
  3. Recompute μ_{c}, μ_{t} according to Eqs.(32, 33)
  4. If termination condition is met, finish. Otherwise goto 2.

- This is the essence of the **EM algorithm**, with Step 2 called the **Expectation** (E) step and Step 3 called the **Maximization** (M) step.

Clustering, Soft Assignment, Relation to K-means (1)

An extreme of the previous example is that **no** data have the x-coordinate value (car/truck vehicle type). Everything works just as well:

$$\mu_{\mathsf{c}} = \frac{\sum_{i=1}^{M} q(z_i = \mathsf{car})\, y_i^{\bullet}}{\sum_{i=1}^{M} q(z_i = \mathsf{car})} \tag{34}$$

$$\mu_{\mathsf{t}} = \frac{\sum_{i=1}^{M} q(z_i = \mathsf{truck})\, y_i^{\bullet}}{\sum_{i=1}^{M} q(z_i = \mathsf{truck})} \tag{35}$$

1. Initialize μ_{c}, μ_{t}
2. Compute $q(z_i) = p(z_i | y_i^{\bullet}, \mu_{\mathsf{c}}, \mu_{\mathsf{t}})$ for all $i = 1, 2, ..., M$
3. Recompute μ_{c}, μ_{t} according to Eqs.(36, 37)
4. If termination condition is met, finish. Otherwise goto 2.

Note: Can you imagine this algorithm to end up at a local maximum?

Clustering, Soft Assignment, Relation to K-means (2)

An extreme of the previous example is that **no** data have the x-coordinate (car/truck).

$$\mu_{\text{c}} = \frac{\sum_{i=1}^{M} q(z_i = \text{car})\, y_i}{\sum_{i=1}^{M} q(z_i = \text{car})} \tag{36}$$

$$\mu_{\text{t}} = \frac{\sum_{i=1}^{M} q(z_i = \text{truck})\, y_i}{\sum_{i=1}^{M} q(z_i = \text{truck})} \tag{37}$$

EM algorithm:

1. Initialize μ_{c}, μ_{t}
2. Compute $q(z_i) = p(z_i | y_i, \mu_{\text{c}}, \mu_{\text{t}})$ for all $i = 1, 2, ..., M$
3. Recompute μ_{c}, μ_{t} according to Eqs.(36, 37)
4. If termination condition is met, finish. Otherwise goto 2.

K-means:

1. ditto
2. $q(z_i = \text{car}) = [\![|y_i - \mu_{\text{c}}| < |y_i - \mu_{\text{t}}|]\!]$
   $q(z_i = \text{truck}) = [\![|y_i - \mu_{\text{t}}| \leq |y_i - \mu_{\text{c}}|]\!]$
   for all $i = 1, 2, ..., M$
3. ditto
4. ditto

EM-based clustering uses soft assignment. K-means can be interpreted as an EM-based clustering with hard assignment.

EM algorithm

- $\mathcal{T}$: training set
- $\mathbf{o}$: all observed values (no essential difference between $\mathcal{T}$ and $\mathbf{o}$, just notational convenience)
- $\mathbf{z}$: all unobserved values
- $\boldsymbol{\theta}$: model parameters to be estimated.

Goal: Find $\boldsymbol{\theta}^*$ using the Maximum Likelihood approach:

$$\boldsymbol{\theta}^* = \operatorname*{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \ln p(\mathbf{o}|\boldsymbol{\theta}) \tag{38}$$

Line of thought

Assume that solving this:

$$\operatorname*{argmax}_{\boldsymbol{\theta}} \ln p(\mathbf{o}, \mathbf{z}|\boldsymbol{\theta}) \tag{39}$$

is easy (optimal parameters had $\mathbf{z}$ been known.)

Our goal will be to rewrite Eq. (38) in a way which will involve optimization terms of kind as in Eq. (39).

Lower Bound on the Log Likelihood

$$\ln p(\mathbf{o}|\boldsymbol{\theta}) = \ln \sum_{\mathbf{z}} p(\mathbf{o},\mathbf{z}|\boldsymbol{\theta}) \tag{40}$$

Marginalizing over missing values

$$= \ln \sum_{\mathbf{z}} q(\mathbf{z}) \frac{p(\mathbf{o},\mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} \tag{41}$$

Introduction of distribution $q(\mathbf{z})$

As $\forall \mathbf{z} : 0 \le q(\mathbf{z}) \le 1$ and $\sum_{\mathbf{z}} q(\mathbf{z}) = 1$, the sum is now a convex combination of $p(\mathbf{o},\mathbf{z}|\boldsymbol{\theta})/q(\mathbf{z})$.

$$\ge \sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p(\mathbf{o},\mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} \tag{42}$$

Jensen's inequality. Here inequality holds because logarithm is a concave function.

Define

$$\mathcal{L}(q,\boldsymbol{\theta}) = \sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p(\mathbf{o},\mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} \,. \tag{43}$$

This $\mathcal{L}(q,\boldsymbol{\theta})$ is the lower bound for $\ln p(\mathbf{o}|\boldsymbol{\theta})$ due to Eq. (42), for any distribution q.

Maximizing $\mathcal{L}(q,\boldsymbol{\theta})$ will also push the log likelihood upwards.

How Tight Is This Bound? (1)

$$\ln p(\mathbf{o}|\boldsymbol{\theta}) - \mathcal{L}(q,\boldsymbol{\theta}) = \ln p(\mathbf{o}|\boldsymbol{\theta}) - \sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p(\mathbf{o},\mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} \quad (44)$$

$$= \ln p(\mathbf{o}|\boldsymbol{\theta}) - \sum_{\mathbf{z}} q(\mathbf{z}) \{\ln \underbrace{p(\mathbf{o},\mathbf{z}|\boldsymbol{\theta})}_{p(\mathbf{z}|\mathbf{o},\boldsymbol{\theta})p(\mathbf{o}|\boldsymbol{\theta})} - \ln q(\mathbf{z})\} \quad (45)$$

$$= \ln p(\mathbf{o}|\boldsymbol{\theta}) - \sum_{\mathbf{z}} q(\mathbf{z}) \{\ln p(\mathbf{z}|\mathbf{o},\boldsymbol{\theta}) + \ln p(\mathbf{o}|\boldsymbol{\theta}) - \ln q(\mathbf{z})\} \quad (46)$$

$$= \ln p(\mathbf{o}|\boldsymbol{\theta}) - \underbrace{\sum_{\mathbf{z}} q(\mathbf{z})}_{1} \ln p(\mathbf{o}|\boldsymbol{\theta}) - \sum_{\mathbf{z}} q(\mathbf{z}) \{\ln p(\mathbf{z}|\mathbf{o},\boldsymbol{\theta}) - \ln q(\mathbf{z})\} \quad (47)$$

$$= -\sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p(\mathbf{z}|\mathbf{o},\boldsymbol{\theta})}{q(\mathbf{z})} \quad (48)$$

This is the Kullback Leibler divergence between the two distributions $q(\mathbf{z})$ and $p(\mathbf{z}|\mathbf{o},\boldsymbol{\theta})$:

$$D_{\mathsf{KL}}(q||p) = \sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{o},\boldsymbol{\theta})} = -\sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p(\mathbf{z}|\mathbf{o},\boldsymbol{\theta})}{q(\mathbf{z})} \quad (49)$$

How Tight Is This Bound? (2)

$$\ln p(\mathbf{o}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + D_{\mathsf{KL}}(q||p) \tag{50}$$

$$\begin{array}{ccc} \uparrow & \uparrow & \uparrow \\ \text{log likelihood} & \text{lower bound} & \text{gap} \end{array}$$

We already know that due to Jensen's inequality, $\mathcal{L}(q, \boldsymbol{\theta})$ is indeed the lower bound. This is confirmed by the fact that $D_{\mathsf{KL}}(q||p) \geq 0$ for any q, p. Additionally,

$$D_{\mathsf{KL}}(q||p) = 0 \qquad \Leftrightarrow \qquad p = q. \tag{51}$$

When $q = p$, the bound is tight.

EM algorithm

$$\ln p(\mathbf{o}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + D_{\mathsf{KL}}(q||p) \tag{52}$$

$\uparrow$ log likelihood $\quad$ $\uparrow$ lower bound $\quad$ $\uparrow$ gap

EM algorithm attempts to maximize the log-likelihood by instead maximizing the lower bound (why 'attempts'? Because it may end up in local maximum).

1. Initialize $\boldsymbol{\theta} = \boldsymbol{\theta}^{(0)}$ $(t = 0)$
2. **E-step** (Expectation):
$$q^{(t+1)} = \operatorname*{argmax}_q \mathcal{L}(q, \boldsymbol{\theta}^{(t)}) \tag{53}$$
3. **M-step** (Maximization):
$$\boldsymbol{\theta}^{(t+1)} = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathcal{L}(q^{(t+1)}, \boldsymbol{\theta}) \tag{54}$$
4. If termination condition is not met, goto 2.

Expectation step

E-step: $\boldsymbol{\theta}^{(t)}$ is fixed

$$q^{(t+1)} = \operatorname*{argmax}_q \mathcal{L}(q, \boldsymbol{\theta}^{(t)}) \tag{55}$$

$$\mathcal{L}(q, \boldsymbol{\theta}^{(t)}) = \underbrace{\ln p(\mathbf{o}|\boldsymbol{\theta}^{(t)})}_{\text{const.}} - D_{\mathsf{KL}}(q||p) \tag{56}$$

Note: The distribution q maximizing this term is the one which minimizes the KL divergence. KL divergence is minimized when the two distributions are the same. Thus, the distribution maximizing Eq. (55) is

$$q^{(t+1)}(\mathbf{z}) = p(\mathbf{z}|\mathbf{o}, \boldsymbol{\theta}^{(t)})\,. \tag{57}$$

$$\left[\text{Recall: } D_{\mathsf{KL}}(q||p) = -\sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p(\mathbf{z}|\mathbf{o}, \boldsymbol{\theta})}{q(\mathbf{z})}\right] \tag{58}$$

Maximization step

M-step: $q^{(t+1)}$ is fixed

$$\boldsymbol{\theta}^{(t+1)} = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathcal{L}(q^{(t+1)}, \boldsymbol{\theta}) \tag{59}$$

$$\mathcal{L}(q^{(t+1)}, \boldsymbol{\theta}) = \sum_{\mathbf{z}} q^{(t+1)}(\mathbf{z}) \ln \frac{p(\mathbf{o}, \mathbf{z} | \boldsymbol{\theta})}{q^{(t+1)}(\mathbf{z})} \tag{60}$$

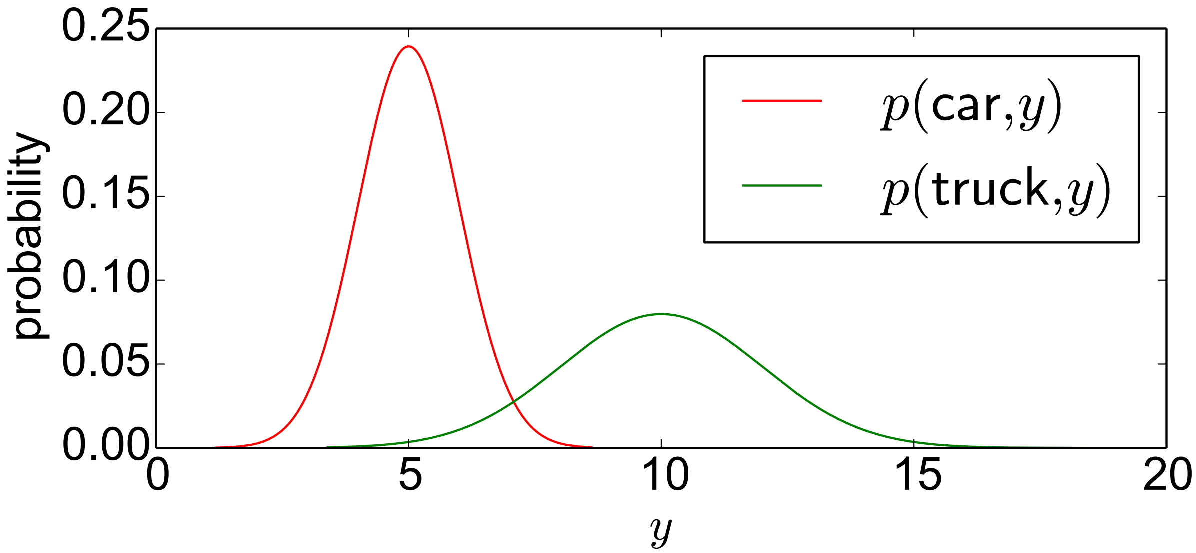
$$= \sum_{\mathbf{z}} q^{(t+1)}(\mathbf{z}) \ln p(\mathbf{o}, \mathbf{z} | \boldsymbol{\theta}) - \underbrace{\sum_{\mathbf{z}} q^{(t+1)}(\mathbf{z}) \ln q^{(t+1)}(\mathbf{z})}_{\text{const.}} \tag{61}$$

Result: The parameters $\boldsymbol{\theta}$ maximizing Eq. (59) are

$$\boldsymbol{\theta}^{(t+1)} = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{\mathbf{z}} q^{(t+1)}(\mathbf{z}) \ln p(\mathbf{o}, \mathbf{z} | \boldsymbol{\theta}) \,. \tag{62}$$

Example 1 - Setting

$\pi_{\text{c}} = 0.6,\ \pi_{\text{t}} = 0.4,\ \sigma_{\text{c}} = 1,\ \sigma_{\text{t}} = 2,\quad \mu_{\text{c}} = 5,\ \mu_{\text{t}} = 10$



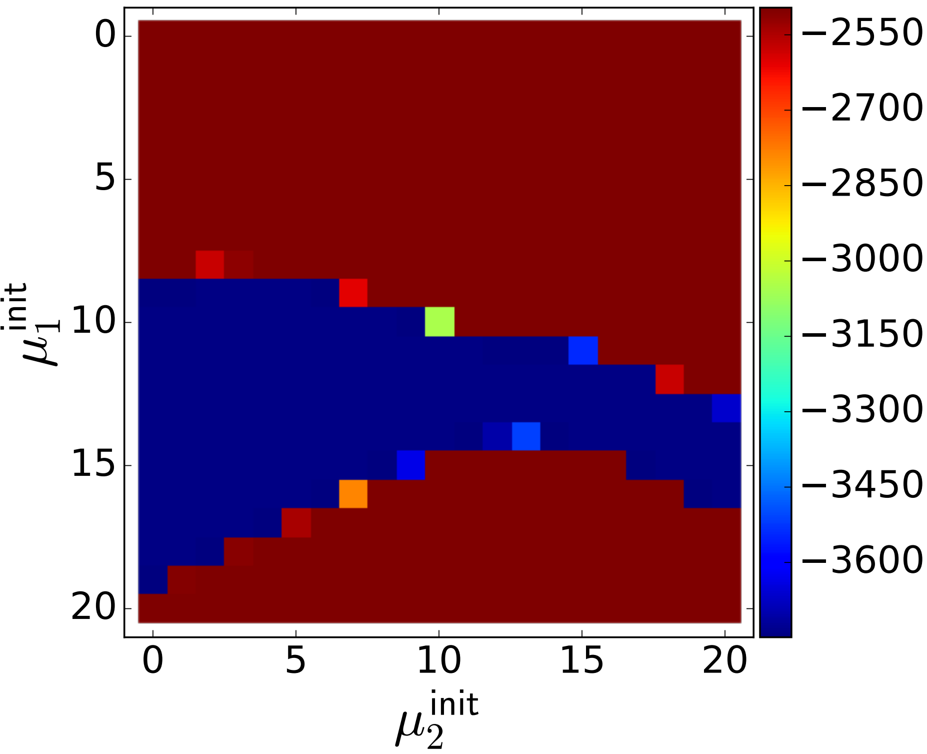
Data:

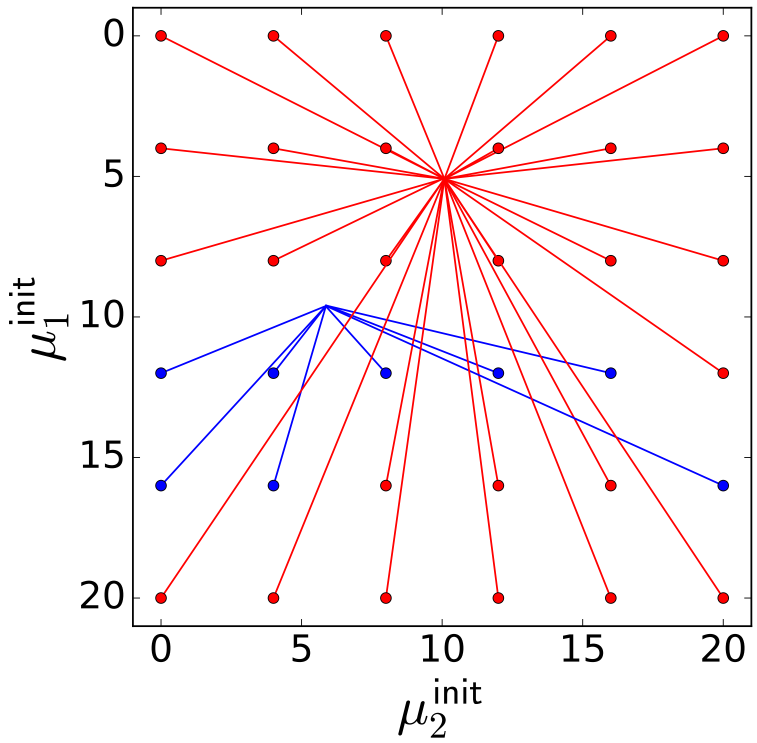
- 50 points from car distribution, 50 points from truck d., 1000 points from mixed distribution (car/truck coordinate unknown)

Experiment:

Employ EM algorithm for estimating μ_1, μ_2. Use different initializations.

Example 1 - Result





Log-likelihood ℓ after 10 iterations of EM, depending on initialization $(\mu_1^{\text{init}}, \mu_2^{\text{init}})$.

Convergence in this case is quite fast (3 iterations are enough for most of the initialization values.)

Value of (μ_1, μ_2) after 10 iterations, depending on initialization $(\mu_1^{\text{init}}, \mu_2^{\text{init}})$. The first point of convergence corresponds to the ground truth values $(\mu_1, \mu_2) = (5, 10)$. The second point is a only a local maximum of log-likelihood. It corresponds to car distribution approximating truck sample points, and vice versa.

Mixture Models

Generalization of the Motivation example with missing values.

$$\mu_{\mathsf{c}} = \frac{\sum_{i=1}^{M} q(z_i = \mathsf{car})\, y_i^{\bullet}}{\sum_{i=1}^{M} q(z_i = \mathsf{car})} \tag{63}$$

$$\sigma_{\mathsf{c}}^2 = \frac{\sum_{i=1}^{M} q(z_i = \mathsf{car})\, (y_i^{\bullet} - \mu_{\mathsf{c}})^2}{\sum_{i=1}^{M} q(z_i = \mathsf{car})} \tag{64}$$

$$\pi_{\mathsf{c}} = \frac{\sum_{i=1}^{M} q(z_i = \mathsf{car})}{M} \tag{65}$$

Example: Mixture of Gaussians

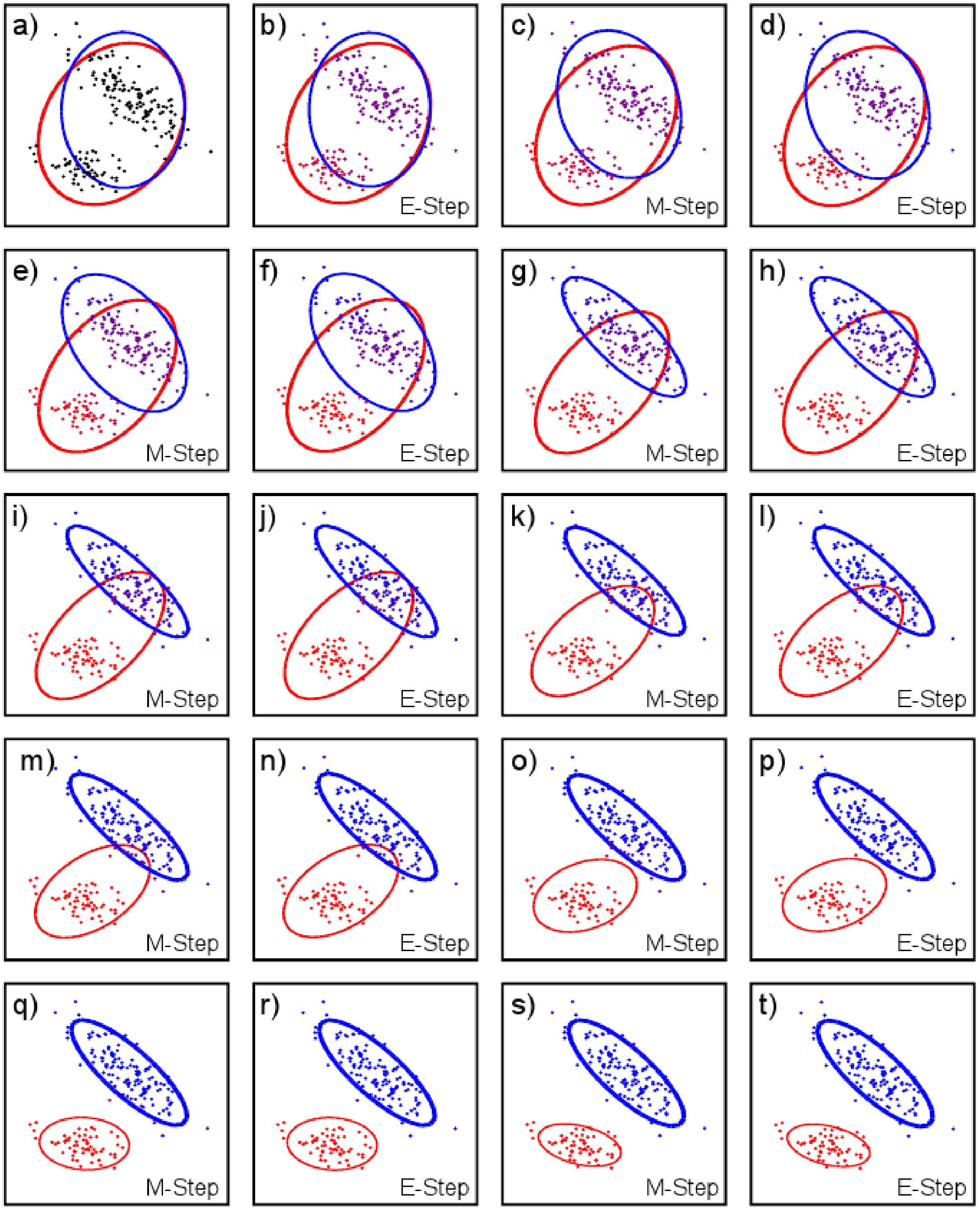


Figure 7.10 a) Initial model. b) E-step. For each data point the posterior probability that is was generated from each Gaussian is calculated (indicated by color of point). c) M-step. The mean, variance and weight of each Gaussian is updated based on these posterior probabilities. Ellipse shows Mahalanobis distance of two. Weight (thickness) of ellipse indicates weight of Gaussian. d-t) Further E-step and M-step iterations.

Image courtesy of Simon Prince. Computer Vision: Models, Learning and Inference, 2012