

Expectation Maximization (EM) Algorithm

lecturer: [J. Matas](mailto:matas@cmp.felk.cvut.cz), 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>

12/Jan/2018

LECTURE PLAN

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

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.

Toy Example 1: (Temperature × Snow) Model Estimation

You are measuring temperature and amount of snow in the mountains in the month of January. Both the temperature t and the snow s observations are binary:

$$t \in \{t_0=\text{low temperature}, t_1=\text{high temperature}\} \tag{1}$$

$$s \in \{s_0=\text{little snow}, s_1=\text{lot of snow}\} \tag{2}$$



Your own long-term research suggests that the model for the joint probability $p(t, s)$ can be parametrized by two scalars a and b and written as

$$p(t, s | a, b)$$

t_0	a	$5a$
t_1	$3b$	b
	s_0	s_1

(3)

At a big ski-center, you have N measurements in total, with counts for individual possibilities for t and s as follows:

observation counts

t_0	N_{00}	N_{01}
t_1	N_{10}	N_{11}
	s_0	s_1

(4)

What is the ML estimate for a and b ?

Toy Example 1: Model Estimation

$$p(t, s|a, b)$$

t_0	a	$5a$
t_1	$3b$	b
	s_0	s_1

observation counts

t_0	N_{00}	N_{01}
t_1	N_{10}	N_{11}
	s_0	s_1

Likelihood is $P(\mathcal{T}|a, b) = a^{N_{00}}(5a)^{N_{01}}(3b)^{N_{10}}(b)^{N_{11}}$.

Log-likelihood is $\ell(\mathcal{T}|a, b) = N_{00} \ln a + N_{01} \ln 5a + N_{10} \ln 3b + N_{11} \ln b$. Maximize this log-likelihood s.t. $6a + 4b = 1$. The Lagrangian is

$L(a, b, \lambda) = N_{00} \ln a + N_{01} \ln 5a + N_{10} \ln 3b + N_{11} \ln b + \lambda(6a + 4b - 1)$. Conditions of optimality are:

$$\frac{\partial L}{\partial a} = N_{00} \frac{1}{a} + N_{01} \frac{1}{a} + 6\lambda = 0 \tag{5}$$

$$\frac{\partial L}{\partial b} = N_{10} \frac{1}{b} + N_{11} \frac{1}{b} + 4\lambda = 0 \tag{6}$$

$$6a + 4b = 1 \tag{7}$$

and they have an easy solution:

$$a = \frac{N_{00} + N_{01}}{6N} \quad b = \frac{N_{10} + N_{11}}{4N} \tag{8}$$

Toy Example 1: Model Estimation (Incomplete Data)

Now imagine you have data from little village in the mountains. Unfortunately, there is *no* measurement for which both temperature and snow amount would be available. The data consist only of T_0 reports of low temperature, T_1 of high temperature, S_0 of little snow and S_1 of lots of snow.

$p(t, s a, b)$		
t_0	a	$5a$
t_1	$3b$	b
	s_0	s_1

 \Rightarrow

$p(t_0)$	$6a$
$p(t_1)$	$4b$
$p(s_0)$	$a + 3b$
$p(s_1)$	$5a + b$

observation counts	
t_0	T_0
t_1	T_1
s_0	S_0
s_1	S_1

Log-likelihood is $\ell(\mathcal{T}|a, b) = T_0 \ln 6a + T_1 \ln 4b + S_0 \ln(a + 3b) + S_1 \ln(5a + b)$.

Maximize this log-likelihood s.t. $6a + 4b = 1$. The Lagrangian is

$$L(a, b, \lambda) = T_0 \ln 6a + T_1 \ln 4b + S_0 \ln(a + 3b) + S_1 \ln(5a + b) + \lambda(6a + 4b - 1).$$

Conditions of optimality:

$$\frac{\partial L}{\partial a} = \frac{T_0}{a} + \frac{S_0}{a + 3b} + \frac{5S_1}{5a + b} + 6\lambda = 0 \tag{9}$$

$$\frac{\partial L}{\partial b} = \frac{T_1}{b} + \frac{3S_0}{a + 3b} + \frac{3S_1}{5a + b} + 4\lambda = 0 \tag{10}$$

$$6a + 4b = 1 \tag{11}$$

→ Not as easy to solve as in the previous case!

Toy Example 1: Model Estimation using EM algorithm

This is what EM algorithm would do to maximize likelihood for these incomplete data.

1. Make initial estimate of a and b
2. **E-step:** For each observation, compute the distribution over the missing value, given the observed value and current estimate of a and b .

Consider e.g. observation (\bullet, s_0) where ' \bullet ' is the unobserved temperature t and s_0 is the observed amount of snow. The distrib. $q(t) = p(t|s_0, a, b)$ is computed as follows:

$$p(t, s|a, b)$$

t_0	a	$5a$
t_1	$3b$	b
	s_0	s_1

$$q(t_0) = p(t_0|s_0, a, b) = \frac{a}{a + 3b} \quad (12)$$

$$q(t_1) = p(t_1|s_0, a, b) = \frac{3b}{a + 3b} \quad (13)$$

3. **M-step:** Recompute parameters a, b :

Use the distribution q computed in the previous step as weights of respective complete measurements. I.e. the considered incomplete observation (\bullet, s_0) produces two complete observations:

(t_0, s_0) with weight $q(t_0)$, and (t_1, s_0) with weight $q(t_1)$.

Let w_{ij} be the sum of weights for observations (t_i, s_j) across the entire dataset. Then a and b are computed simply using the result for complete data:

$$a = \frac{w_{00} + w_{01}}{6N}, \quad b = \frac{w_{10} + w_{11}}{4N} \quad (14)$$

4. Iterate (go to 2.)

Toy Example 2: Estimating Means of Two Normal Distributions

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

$$p(x, y) : \text{distribution}, \quad x \in \{\text{car, truck}\}, \quad y \in \mathbb{R} \quad (15)$$



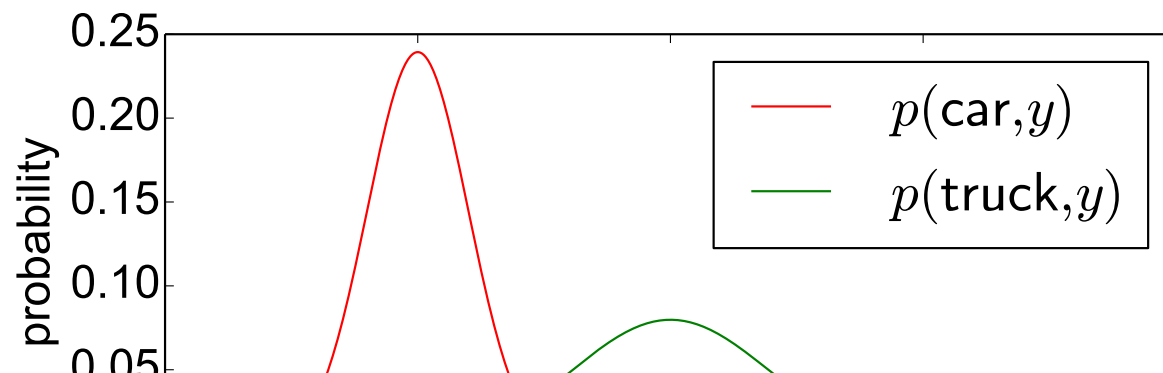
$$p(\text{car}, y) = \pi_c \mathcal{N}(y | \mu_c, \sigma_c = 1) = \kappa_c \exp \left\{ -\frac{1}{2} (y - \mu_c)^2 \right\}, \quad (\kappa_c = \frac{\pi_c}{\sqrt{2\pi}}) \quad (16)$$



$$p(\text{truck}, y) = \pi_t \mathcal{N}(y | \mu_t, \sigma_t = 2) = \kappa_t \exp \left\{ -\frac{1}{8} (y - \mu_t)^2 \right\}, \quad (\kappa_t = \frac{\pi_c}{\sqrt{8\pi}}) \quad (17)$$

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 ($\pi_c = 0.6, \pi_t = 0.4, \sigma_c = 1, \sigma_t = 2, \mu_c = 5, \mu_t = 10$)



Toy Example 2, Complete Data → Easy

The observations are:

$$\mathcal{T} = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\} \quad (18)$$

$$= \underbrace{\{(\text{car}, y_1^{(c)}), (\text{car}, y_2^{(c)}), \dots, (\text{car}, y_C^{(c)})\}}_{C \text{ car observations}}, \underbrace{\{(\text{truck}, y_1^{(t)}), (\text{truck}, y_2^{(t)}), \dots, (\text{truck}, y_T^{(t)})\}}_{T \text{ truck observations}} \quad (19)$$

Log-likelihood $\ell(\mathcal{T}) = \ln p(\mathcal{T} | \mu_c, \mu_t)$:

$$\ell(\mathcal{T}) = \sum_{i=1}^N \ln p(x_i, y_i | \mu_c, \mu_t) = C \ln \kappa_c - \frac{1}{2} \sum_{i=1}^C (y_i^{(c)} - \mu_c)^2 + T \ln \kappa_t - \frac{1}{8} \sum_{i=1}^T (y_i^{(t)} - \mu_t)^2 \quad (20)$$

Estimation of μ_1, μ_2 using ML is very easy:

$$\frac{\partial \ell(\mathcal{T})}{\partial \mu_c} = \sum_{i=1}^C (y_i^{(c)} - \mu_c) = 0 \quad \Rightarrow \quad \mu_c = \frac{1}{C} \sum_{i=1}^C y_i^{(c)} \quad (21)$$

$$\frac{\partial \ell(\mathcal{T})}{\partial \mu_t} = \frac{1}{4} \sum_{i=1}^T (y_i^{(t)} - \mu_t) = 0 \quad \Rightarrow \quad \mu_t = \frac{1}{T} \sum_{i=1}^T y_i^{(t)} \quad (22)$$

Toy Example 2, Incomplete Data → Difficult (1)

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

$$\mathcal{T} = \{(\text{car}, y_1^{(c)}), \dots, (\text{car}, y_C^{(c)}), (\text{truck}, y_1^{(t)}), \dots, (\text{truck}, y_T^{(t)}), \underbrace{(\bullet, y_1^\bullet), \dots, (\bullet, y_M^\bullet)}_{\substack{\text{data with unknown} \\ \text{vehicle type}}}\} \quad (23)$$

What is the probability of observing y^\bullet ?

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

Log-likelihood:

$$\ell(\mathcal{T}) = \sum_{i=1}^N \ln p(x_i, y_i | \mu_c, \mu_t) = \overbrace{C \ln \kappa_c - \frac{1}{2} \sum_{i=1}^C (y_i^{(c)} - \mu_c)^2 + T \ln \kappa_t - \frac{1}{8} \sum_{i=1}^T (y_i^{(t)} - \mu_t)^2}^{\text{same term as before}} \quad (24)$$

$$+ \sum_{i=1}^M \ln \left(\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\} + \kappa_t \exp \left\{ -\frac{1}{8} (y_i^\bullet - \mu_t)^2 \right\} \right) \quad (25)$$

Toy Example 2, Incomplete Data → Difficult (2)

Log-likelihood:

$$\ell(\mathcal{T}) = C \ln \kappa_c - \frac{1}{2} \sum_{i=1}^C (y_i^{(c)} - \mu_c)^2 + T \ln \kappa_t - \frac{1}{8} \sum_{i=1}^T (y_i^{(t)} - \mu_t)^2 \quad (26)$$

$$+ \sum_{i=1}^M \ln \left(\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\} + \kappa_t \exp \left\{ -\frac{1}{8} (y_i^\bullet - \mu_t)^2 \right\} \right) \quad (27)$$

Optimality condition (shown for μ_c only):

$$0 = \frac{\partial \ell(\mathcal{T})}{\partial \mu_c} = \sum_{i=1}^C (y_i^{(c)} - \mu_c) + \quad (28)$$

$$+ \sum_{i=1}^M \frac{\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\}}{\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\} + \kappa_t \exp \left\{ -\frac{1}{8} (y_i^\bullet - \mu_t)^2 \right\}} (y_i^\bullet - \mu_c) \quad (29)$$

Missing Values, Optimality Condition

Log-likelihood:

$$\ell(\mathcal{T}) = C \ln \kappa_c - \frac{1}{2} \sum_{i=1}^C (y_i^{(c)} - \mu_c)^2 + T \ln \kappa_t - \frac{1}{8} \sum_{i=1}^T (y_i^{(t)} - \mu_t)^2 \quad (30)$$

$$+ \sum_{i=1}^M \ln \left(\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\} + \kappa_t \exp \left\{ -\frac{1}{8} (y_i^\bullet - \mu_t)^2 \right\} \right) \quad (31)$$

Optimality condition (shown for μ_c only):

$$0 = \frac{\partial \ell(\mathcal{T})}{\partial \mu_c} = \sum_{i=1}^C (y_i^{(c)} - \mu_c) + \underbrace{p(\text{car}, y_i^\bullet | \mu_c, \mu_t)}_{\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\}} \quad (32)$$

$$+ \sum_{i=1}^M \frac{\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\}}{\underbrace{\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\} + \kappa_t \exp \left\{ -\frac{1}{8} (y_i^\bullet - \mu_t)^2 \right\}}_{\substack{p(\text{car}, y_i^\bullet | \mu_c, \mu_t) \\ p(\text{truck}, y_i^\bullet | \mu_c, \mu_t)}}} (y_i^\bullet - \mu_c) \quad (33)$$

Missing Values, Optimality Condition

Log-likelihood:

$$\ell(\mathcal{T}) = C \ln \kappa_c - \frac{1}{2} \sum_{i=1}^C (y_i^{(c)} - \mu_c)^2 + T \ln \kappa_t - \frac{1}{8} \sum_{i=1}^T (y_i^{(t)} - \mu_t)^2 \quad (34)$$

$$+ \sum_{i=1}^M \ln \left(\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\} + \kappa_t \exp \left\{ -\frac{1}{8} (y_i^\bullet - \mu_t)^2 \right\} \right) \quad (35)$$

Optimality condition (shown for μ_c only):

$$0 = \frac{\partial \ell(\mathcal{T})}{\partial \mu_c} = \sum_{i=1}^C (y_i^{(c)} - \mu_c) + \underbrace{p(\text{car} | y_i^\bullet, \mu_c, \mu_t)} \quad (36)$$

$$+ \sum_{i=1}^M \frac{\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\}}{\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\} + \kappa_t \exp \left\{ -\frac{1}{8} (y_i^\bullet - \mu_t)^2 \right\}} (y_i^\bullet - \mu_c) \quad (37)$$

Missing Values, Optimality Conditions

Optimality conditions (shown for both μ_c and μ_t):

$$0 = \frac{\partial \ell(\mathcal{T})}{\partial \mu_c} = \sum_{i=1}^C (y_i^{(c)} - \mu_c) + \underbrace{p(\text{car} | y_i^\bullet, \mu_c, \mu_t)}_{\text{from (39)}} \quad (38)$$

$$+ \sum_{i=1}^M \frac{\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\}}{\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\} + \kappa_t \exp \left\{ -\frac{1}{8} (y_i^\bullet - \mu_t)^2 \right\}} (y_i^\bullet - \mu_c) \quad (39)$$

$$0 = 4 \frac{\partial \ell(\mathcal{T})}{\partial \mu_t} = \sum_{i=1}^T (y_i^{(t)} - \mu_t) + \sum_{i=1}^M p(\text{truck} | y_i^\bullet, \mu_c, \mu_t) (y_i^\bullet - \mu_t) \quad (40)$$

Note:

- ◆ Complicated equations for the unknowns 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):

$$\sum_{i=1}^C (y_i^{(c)} - \mu_c) + \sum_{i=1}^M p(\text{car} | y_i^\bullet, \mu_c, \mu_t) (y_i^\bullet - \mu_c) = 0 \quad (41)$$

$$\sum_{i=1}^T (y_i^{(t)} - \mu_t) + \sum_{i=1}^M p(\text{truck} | y_i^\bullet, \mu_c, \mu_t) (y_i^\bullet - \mu_t) = 0 \quad (42)$$

If $p(\text{car} | y_i^\bullet, \mu_c, \mu_t)$ and $p(\text{truck} | y_i^\bullet, \mu_c, \mu_t)$ **were** known, the estimation would've been easy:

- ◆ Let z_i ($i = 1, 2, \dots, M$), $z_i \in \{\text{car}, \text{truck}\}$ denote the missing values. Define $q(z_i) = p(z_i | y_i^\bullet, \mu_c, \mu_t)$
- ◆ The equations lead to

$$\sum_{i=1}^C (y_i^{(c)} - \mu_c) + \sum_{i=1}^M q(z_i = \text{car}) (y_i^\bullet - \mu_c) = 0 \quad (43)$$

$$\Rightarrow \mu_c = \frac{\sum_{i=1}^C y_i^{(c)} + \sum_{i=1}^M q(z_i = \text{car}) y_i^\bullet}{C + \sum_{i=1}^M q(z_i = \text{car})} \quad (44)$$

and similarly,

$$\mu_t = \frac{\sum_{i=1}^T y_i^{(t)} + \sum_{i=1}^M q(z_i = \text{truck}) y_i^\bullet}{T + \sum_{i=1}^M q(z_i = \text{truck})} \quad (45)$$

Missing Values, EM Approach

$$\mu_c = \frac{\sum_{i=1}^C y_i^{(c)} + \sum_{i=1}^M q(z_i = \text{car}) y_i}{C + \sum_{i=1}^M q(z_i = \text{car})} \quad (46)$$

$$\mu_t = \frac{\sum_{i=1}^T y_i^{(t)} + \sum_{i=1}^M q(z_i = \text{truck}) y_i}{T + \sum_{i=1}^M q(z_i = \text{truck})} \quad (47)$$

- ◆ 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 $q(z_i) = p(z_i | y_i, \mu_c, \mu_t)$ for all $i = 1, 2, \dots, M$
 3. Recompute μ_c, μ_t according to Eqs.(46, 47)
 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:

$$\mu_c = \frac{\sum_{i=1}^M q(z_i = \text{car}) y_i}{\sum_{i=1}^M q(z_i = \text{car})} \quad (48)$$

$$\mu_t = \frac{\sum_{i=1}^M q(z_i = \text{truck}) y_i}{\sum_{i=1}^M q(z_i = \text{truck})} \quad (49)$$

1. Initialize μ_c, μ_t
2. Compute $q(z_i) = p(z_i | y_i, \mu_c, \mu_t)$ for all $i = 1, 2, \dots, M$
3. Recompute μ_c, μ_t according to Eqs.(50, 51)
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).

$$\mu_c = \frac{\sum_{i=1}^M q(z_i = \text{car}) y_i^\bullet}{\sum_{i=1}^M q(z_i = \text{car})} \quad (50)$$

$$\mu_t = \frac{\sum_{i=1}^M q(z_i = \text{truck}) y_i^\bullet}{\sum_{i=1}^M q(z_i = \text{truck})} \quad (51)$$

EM algorithm:

1. Initialize μ_c, μ_t
2. Compute $q(z_i) = p(z_i | y_i^\bullet, \mu_c, \mu_t)$
for all $i = 1, 2, \dots, M$
3. Recompute μ_c, μ_t according to Eqs.(50, 51)
4. If termination condition is met, finish.
Otherwise goto 2.

K-means:

1. ditto
2. $q(z_i = \text{car}) = \mathbb{I}[|y_i^\bullet - \mu_c| < |y_i^\bullet - \mu_t|]$
 $q(z_i = \text{truck}) = \mathbb{I}[|y_i^\bullet - \mu_t| \leq |y_i^\bullet - \mu_c|]$
for all $i = 1, 2, \dots, M$
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 - Derivation

- ◆ \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
- ◆ θ : model parameters to be estimated.

Goal: Find θ^* using the Maximum Likelihood approach:

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \ell(\theta) = \underset{\theta}{\operatorname{argmax}} \ln p(\mathbf{o}|\theta) \quad (52)$$

Line of thought

Assume that solving this:

$$\underset{\theta}{\operatorname{argmax}} \ln p(\mathbf{o}, \mathbf{z}|\theta) \quad (53)$$

is easy (that is, estimation of optimal parameters if data are complete.)

Our goal will be to rewrite Eq. (52) in a way which will involve optimization terms of kind as in Eq. (53).

Lower Bound on the Log Likelihood

$$\ln p(\mathbf{o}|\boldsymbol{\theta}) = \ln \sum_{\mathbf{z}} p(\mathbf{o}, \mathbf{z}|\boldsymbol{\theta}) \quad (54)$$

$$= \ln \sum_{\mathbf{z}} q(\mathbf{z}) \frac{p(\mathbf{o}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} \quad (55)$$

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

As $\forall \mathbf{z} : 0 \leq q(\mathbf{z}) \leq 1$ and $\sum_{\mathbf{z}} q(\mathbf{z}) = 1$, the sum is now a convex combination of

$p(\mathbf{o}, \mathbf{z}|\boldsymbol{\theta})/q(\mathbf{z})$.

$$\geq \sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p(\mathbf{o}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} \quad (56)$$

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

Define

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p(\mathbf{o}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})}. \quad (57)$$

This $\mathcal{L}(q, \boldsymbol{\theta})$ is the lower bound for $\ln p(\mathbf{o}|\boldsymbol{\theta})$ due to Eq. (56), for any distribution q .

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

How Tight Is This Bound? (1)

$$\ln p(\mathbf{o}|\boldsymbol{\theta}) - \mathcal{L}(q, \boldsymbol{\theta}) = \ln p(\mathbf{o}|\boldsymbol{\theta}) - \sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p(\mathbf{o}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} \quad (58)$$

$$= \ln p(\mathbf{o}|\boldsymbol{\theta}) - \sum_{\mathbf{z}} q(\mathbf{z}) \{ \underbrace{\ln p(\mathbf{o}, \mathbf{z}|\boldsymbol{\theta})}_{p(\mathbf{z}|\mathbf{o}, \boldsymbol{\theta})p(\mathbf{o}|\boldsymbol{\theta})} - \ln q(\mathbf{z}) \} \quad (59)$$

$$= \ln p(\mathbf{o}|\boldsymbol{\theta}) - \sum_{\mathbf{z}} q(\mathbf{z}) \{ \ln p(\mathbf{z}|\mathbf{o}, \boldsymbol{\theta}) + \ln p(\mathbf{o}|\boldsymbol{\theta}) - \ln q(\mathbf{z}) \} \quad (60)$$

$$= \ln p(\mathbf{o}|\boldsymbol{\theta}) - \underbrace{\sum_{\mathbf{z}} q(\mathbf{z}) \ln p(\mathbf{o}|\boldsymbol{\theta})}_1 - \sum_{\mathbf{z}} q(\mathbf{z}) \{ \ln p(\mathbf{z}|\mathbf{o}, \boldsymbol{\theta}) - \ln q(\mathbf{z}) \} \quad (61)$$

$$= - \sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p(\mathbf{z}|\mathbf{o}, \boldsymbol{\theta})}{q(\mathbf{z})} \quad (62)$$

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

$$D_{\text{KL}}(q||p) = \sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{o}, \boldsymbol{\theta})} = - \sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p(\mathbf{z}|\mathbf{o}, \boldsymbol{\theta})}{q(\mathbf{z})} \quad (63)$$

How Tight Is This Bound? (2)

$$\ln p(\mathbf{o}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + D_{\text{KL}}(q||p) \quad (64)$$

↑ ↑ ↑

log likelihood lower bound gap

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 $D_{\text{KL}}(q||p) \geq 0$ for any q, p . Additionally,

$$D_{\text{KL}}(q||p) = 0 \quad \Leftrightarrow \quad p = q. \quad (65)$$

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

EM algorithm

$$\ln p(\mathbf{o}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + D_{\text{KL}}(q||p) \quad (66)$$

↑ ↑ ↑
log likelihood lower bound 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):

$$q^{(t+1)} = \operatorname{argmax}_q \mathcal{L}(q, \boldsymbol{\theta}^{(t)}) \quad (67)$$

3. **M-step** (Maximization):

$$\boldsymbol{\theta}^{(t+1)} = \operatorname{argmax}_{\boldsymbol{\theta}} \mathcal{L}(q^{(t+1)}, \boldsymbol{\theta}) \quad (68)$$

4. If termination condition is not met, goto 2.

Expectation step

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

$$q^{(t+1)} = \operatorname{argmax}_q \mathcal{L}(q, \boldsymbol{\theta}^{(t)}) \quad (69)$$

$$\mathcal{L}(q, \boldsymbol{\theta}^{(t)}) = \underbrace{\ln p(\mathbf{o} | \boldsymbol{\theta}^{(t)})}_{\text{const.}} - D_{\text{KL}}(q || p) \quad (70)$$

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. (69) is

$$q^{(t+1)}(\mathbf{z}) = p(\mathbf{z} | \mathbf{o}, \boldsymbol{\theta}^{(t)}) . \quad (71)$$

$$\left[\text{Recall: } D_{\text{KL}}(q || p) = - \sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p(\mathbf{z} | \mathbf{o}, \boldsymbol{\theta})}{q(\mathbf{z})} \right] \quad (72)$$

Maximization step

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

$$\boldsymbol{\theta}^{(t+1)} = \operatorname{argmax}_{\boldsymbol{\theta}} \mathcal{L}(q^{(t+1)}, \boldsymbol{\theta}) \quad (73)$$

$$\mathcal{L}(q^{(t+1)}, \boldsymbol{\theta}) = \sum_{\mathbf{z}} q^{(t+1)}(\mathbf{z}) \ln \frac{p(\mathbf{o}, \mathbf{z} | \boldsymbol{\theta})}{q^{(t+1)}(\mathbf{z})} \quad (74)$$

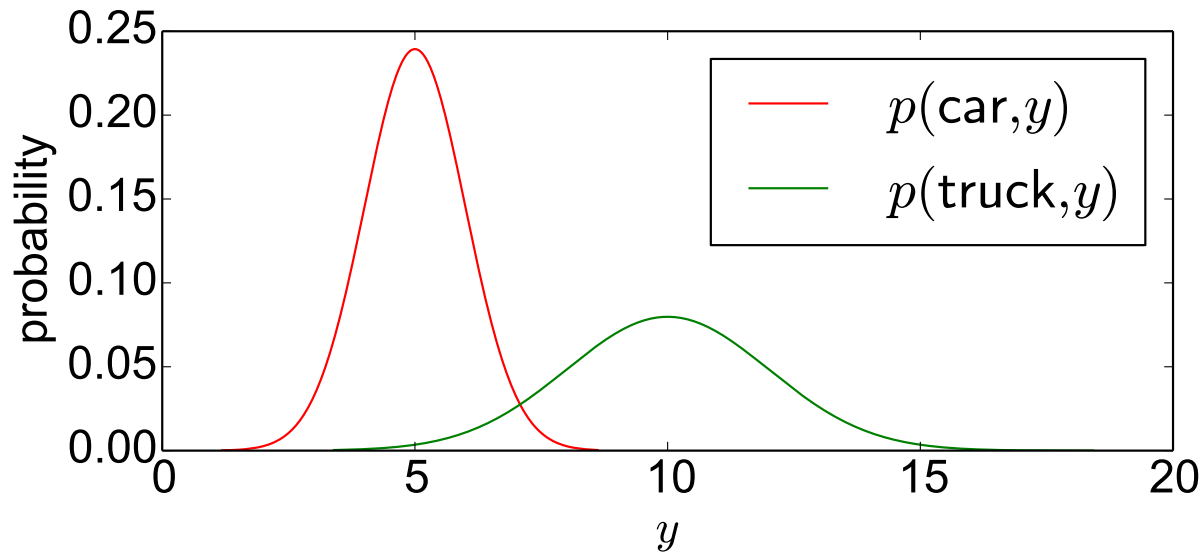
$$= \sum_{\mathbf{z}} q^{(t+1)}(\mathbf{z}) \ln p(\mathbf{o}, \mathbf{z} | \boldsymbol{\theta}) - \underbrace{\sum_{\mathbf{z}} q^{(t+1)}(\mathbf{z}) \ln q^{(t+1)}(\mathbf{z})}_{\text{const.}} \quad (75)$$

Result: The parameters $\boldsymbol{\theta}$ maximizing Eq. (73) are

$$\boldsymbol{\theta}^{(t+1)} = \operatorname{argmax}_{\boldsymbol{\theta}} \sum_{\mathbf{z}} q^{(t+1)}(\mathbf{z}) \ln p(\mathbf{o}, \mathbf{z} | \boldsymbol{\theta}). \quad (76)$$

Example 1 - Setting

$$\pi_c = 0.6, \pi_t = 0.4, \sigma_c = 1, \sigma_t = 2, \mu_c = 5, \mu_t = 10$$



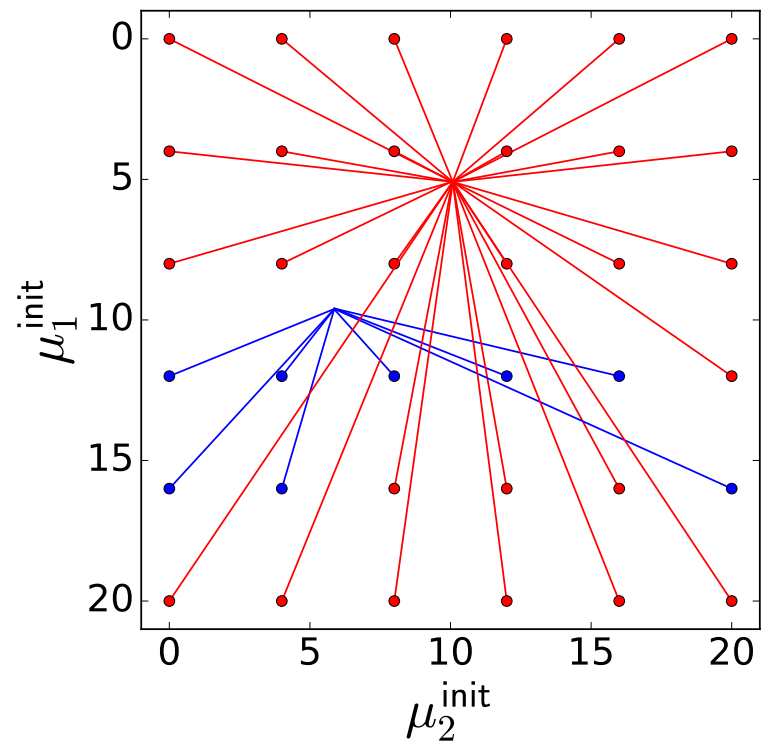
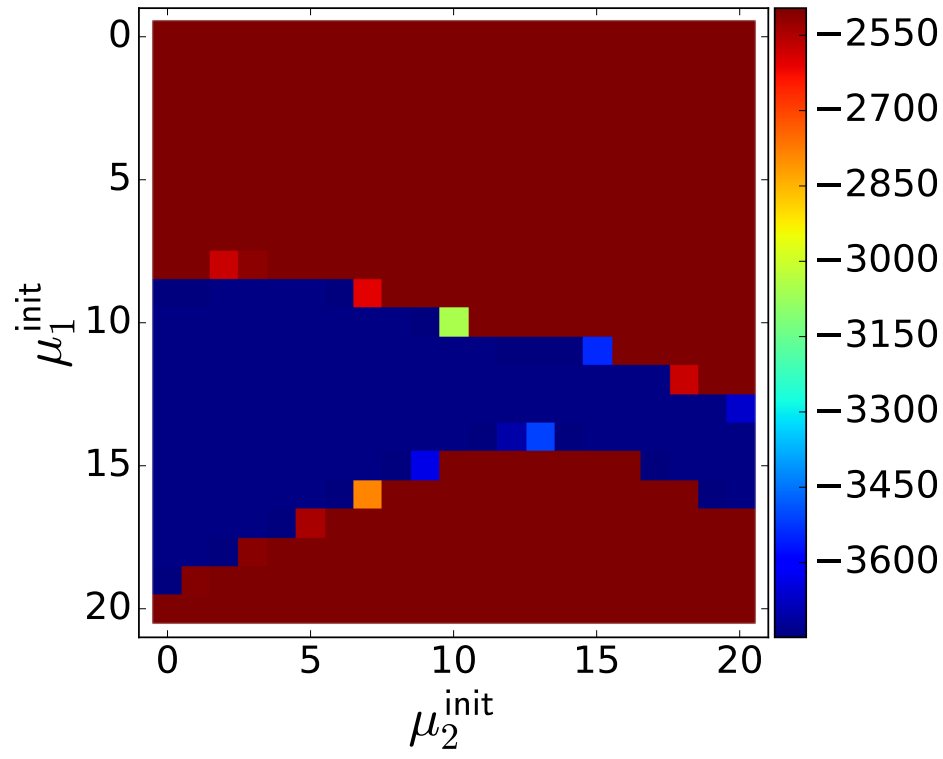
Data:

- ◆ 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.

$$\mu_c = \frac{\sum_{i=1}^M q(z_i = \text{car}) y_i^\bullet}{\sum_{i=1}^M q(z_i = \text{car})} \quad (77)$$

$$\sigma_c^2 = \frac{\sum_{i=1}^M q(z_i = \text{car}) (y_i^\bullet - \mu_c)^2}{\sum_{i=1}^M q(z_i = \text{car})} \quad (78)$$

$$\pi_c = \frac{\sum_{i=1}^M q(z_i = \text{car})}{M} \quad (79)$$

Example: Mixture of Gaussians

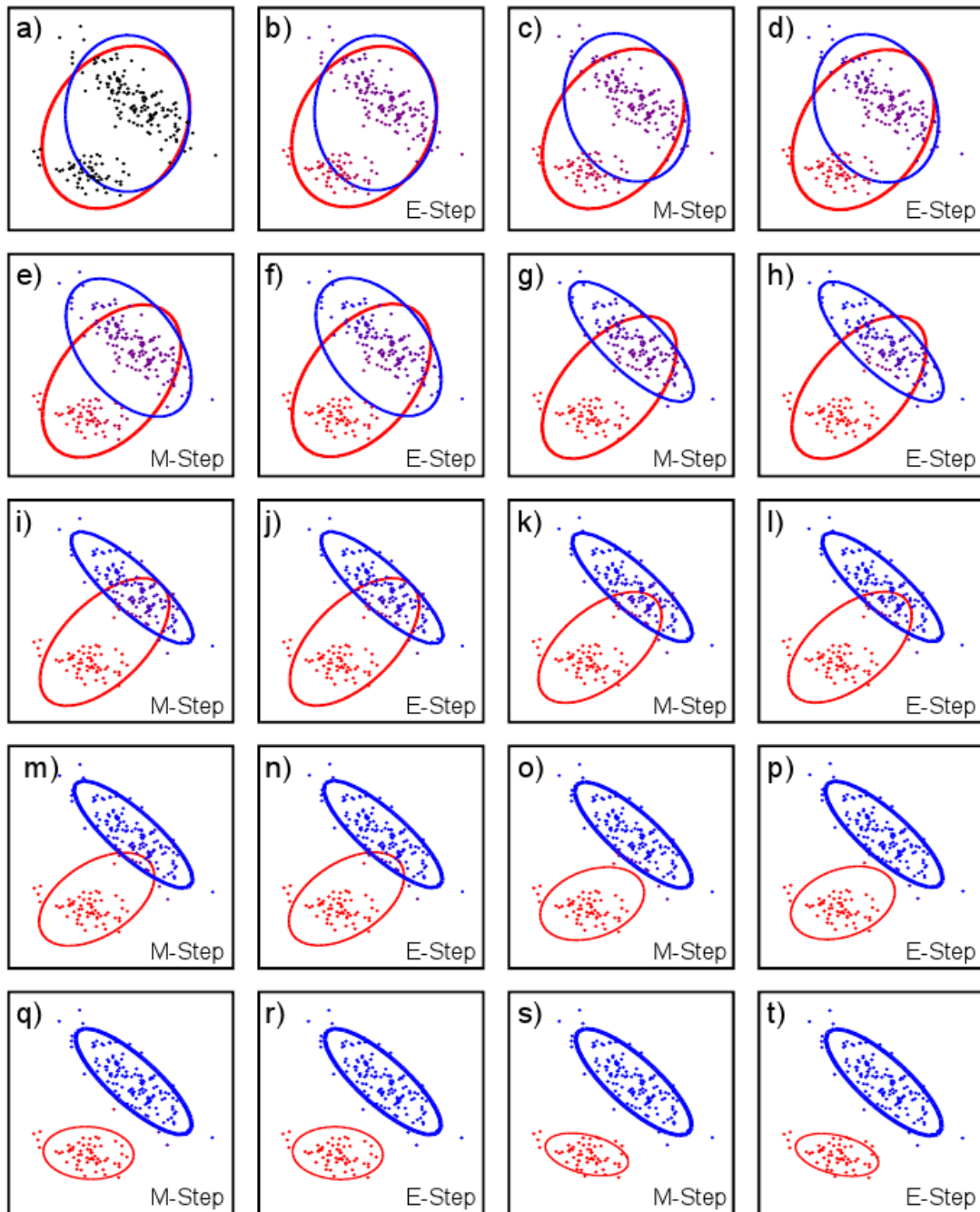


Figure 7.10 a) Initial model. b) E-step. For each data point the posterior probability that it was generated from each Gaussian is calculated (indicated by color of point). c) M-step. The mean, variance and weight of each Gaussian is updated based on these posterior probabilities. Ellipse shows Mahalanobis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