Mixture Densities (DHS 10.2)

- Assume the samples were obtained by selecting a state of nature S_j with probability $P(S_j)$ and then selecting an x according to the probability law $p(x|S_j,\theta_j)$.
- Thus we know the complete probability structures for the problem, except some parameters
- Look at the given assumptions in DHS 10.2:
  - The samples come from a known number *c* of classes
  - The prior probabilities $P(S_i)$ for each class was known
  - The forms for the class-conditional probability densities $p(x|S_i, \theta_j)$ are known
  - The values for the *c* parameter vectors are unknown
  - The category labels are unknown.
- Probability density function of the sample is given by

  $$p(x|\theta)=\sum_{j=1}^{J} p(x|S_j,\theta_j)P(S_j)$$

- This form is called mixture density
- $p(x|S_j,\theta_j)$ = component densities
- $P(S_j)$ = prior probabilities or mixing parameters
- What is unknown? Only the parameters θ
- Completely unidentifiable if we cannot recover a unique parameters θ
- Mixture densities of normal densities are usually identifiable

Maximum-Likelihood Estimates (DHS 10.3)

10.3 Maximum-Likelihood Estimates

Suppose now that we are given a set $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of n unlabeled samples drawn independently from the mixture density

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{j=1}^{c} p(\mathbf{x}|\omega_j, \boldsymbol{\theta}_j) P(\omega_j), \tag{1}$$

where the full parameter vector $\boldsymbol{\theta}$ is fixed but unknown. The likelihood of the observed samples is, by definition, the joint density

$$p(\mathcal{D}|\boldsymbol{\theta}) \equiv \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta}). \tag{3}$$

The maximum-likelihood estimate $\hat{\boldsymbol{\theta}}$ is that value of $\boldsymbol{\theta}$ that maximizes $p(\mathcal{D}|\boldsymbol{\theta})$.

If we assume that $p(\mathcal{D}|\boldsymbol{\theta})$ is a differentiable function of $\boldsymbol{\theta}$, then we can derive some interesting necessary conditions for $\hat{\boldsymbol{\theta}}$. Let l be the logarithm of the likelihood, and let $\nabla_{\boldsymbol{\theta}_i} l$ be the gradient of l with respect to $\boldsymbol{\theta}_i$. Then

$$l = \sum_{k=1}^{n} \ln\, p(\mathbf{x}_k|\boldsymbol{\theta}) \tag{4}$$

and

$$\nabla_{\boldsymbol{\theta}_i} l = \sum_{k=1}^{n} \frac{1}{p(\mathbf{x}_k|\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}_i} \left[\sum_{j=1}^{c} p(\mathbf{x}_k|\omega_j, \boldsymbol{\theta}_j) P(\omega_j) \right]. \tag{5}$$

If we assume that the elements of $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ are functionally independent if $i \neq j$, and if we introduce the posterior probability

$$P(\omega_i|\mathbf{x}_k, \boldsymbol{\theta}) = \frac{p(\mathbf{x}_k|\omega_i, \boldsymbol{\theta}_i) P(\omega_i)}{p(\mathbf{x}_k|\boldsymbol{\theta})}, \tag{6}$$

we see that the gradient of the log-likelihood can be written in the interesting form

$$\nabla_{\boldsymbol{\theta}_i} l = \sum_{k=1}^{n} P(\omega_i|\mathbf{x}_k, \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}_i} \ln\, p(\mathbf{x}_k|\omega_i, \boldsymbol{\theta}_i). \tag{7}$$

Since the gradient must vanish at the value of $\boldsymbol{\theta}_i$ that maximizes l, the maximum-likelihood estimate $\hat{\boldsymbol{\theta}}_i$ must satisfy the conditions

$$\sum_{k=1}^{n} P(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}) \nabla_{\boldsymbol{\theta}_i} \ln\, p(\mathbf{x}_k|\omega_i, \hat{\boldsymbol{\theta}}_i) = 0, \quad i = 1, \ldots, c. \tag{8}$$

Among the solutions to these equations for $\hat{\boldsymbol{\theta}}_i$ we may find the maximum-likelihood solution.

Application to Normal Mixtures (DHS 10.4)

- Three cases for Gaussian mixture

Case	$\boldsymbol{\mu}_i$	$\boldsymbol{\Sigma}_i$	$P(\omega_i)$	c
1	?	×	×	×
2	?	?	?	×
3	?	?	?	?

"x" indicates the parameters are known.

Case 1: The simplest, unknown mean vectors (given in **DHS 10.4.1**, solved in the class)

Case 2: More realistic (discussed in **DHS 10.4.2** but, not handled here)

Case 3: A completely unknown set of data. This cannot be solved by ML (not handled here)

10.4.1 Case 1: Unknown Mean Vectors

If the only unknown quantities are the mean vectors $\boldsymbol{\mu}_i$, then of course $\boldsymbol{\theta}_i$ consists of the components of $\boldsymbol{\mu}_i$. Equation 8 can then be used to obtain necessary conditions on the maximum-likelihood estimate for $\boldsymbol{\mu}_i$. Since the likelihood is

$$\ln p(\mathbf{x}|\omega_i, \boldsymbol{\mu}_i) = -\ln\left[(2\pi)^{d/2}|\boldsymbol{\Sigma}_i|^{1/2}\right] - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i), \tag{14}$$

its derivative is

$$\nabla_{\boldsymbol{\mu}_i} \ln p(\mathbf{x}|\omega_i, \boldsymbol{\mu}_i) = \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i). \tag{15}$$

Thus according to Eq. 8, the maximum-likelihood estimate $\hat{\boldsymbol{\mu}}_i$ must satisfy

$$\sum_{k=1}^{n} P(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\mu}})\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i) = 0, \quad \text{where } \hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_c). \tag{16}$$

After multiplying by $\boldsymbol{\Sigma}_i$ and rearranging terms, we obtain the solution:

$$\hat{\boldsymbol{\mu}}_i = \frac{\sum_{k=1}^{n} P(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\mu}})\mathbf{x}_k}{\sum_{k=1}^{n} P(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\mu}})}. \tag{17}$$

Case 1: Unknown Mean Vectors (DHS 10.4.1)

DHS 10.4.1 Eqs. (14)-(17).

$$\text{Log likelihood} = \log P(\text{data} \mid \mu_1 \cdots \mu_c)$$

Note: assume "c" classes, "n" datapoints. Also notation $P(\underset{\sim}{x} \mid w_j, \mu_1 \cdots \mu_c)$ means "prob of $\underset{\sim}{x}$ given $\underset{\sim}{\mu}_1 \cdots \underset{\sim}{\mu}_c$ and given we know $\underset{\sim}{x}$ is from class w_j"

$$= \log P(x_1 \cdots x_n \mid \mu_1 \cdots \mu_c)$$

$$= \log \prod_{k=1}^{n} P(x_k \mid \mu_1 \cdots \mu_c)$$

$$= \sum_{k=1}^{n} \log P(x_k \mid \mu_1 \cdots \mu_c)$$

$$= \sum_{k=1}^{n} \log \left[\sum_{j=1}^{c} P(x_k \mid w_j, \mu_1 \cdots \mu_c) P(w_j) \right]$$

NOTE $\frac{\partial}{\partial x} \log f(x) = \frac{1}{f(x)} \frac{\partial}{\partial x} f(x)$

$$\frac{\partial \log P}{\partial \mu_i} = \sum_{k=1}^{n} \frac{1}{P(x_k \mid \mu_1 \cdots \mu_c)} \frac{\partial}{\partial \mu_i} \sum_{j=1}^{c} P(x_k \mid w_j, \mu_1 \cdots \mu_c) P(w_j)$$

$$= \sum_{k=1}^{n} \frac{P(w_i)}{P(x_k \mid \mu_1 \cdots \mu_c)} \frac{\partial}{\partial \mu_i} P(x_k \mid w_i, \mu_1 \cdots \mu_c)$$

$$= \sum_{k=1}^{n} \frac{P(w_i)}{P(x_k \mid \mu_1 \cdots \mu_c)} P(x_k \mid w_i, \mu_1 \cdots \mu_c) \frac{\partial}{\partial \mu_i} \log P(x_k \mid w_i, \mu_1 \cdots \mu_c)$$

$$= \sum_{k=1}^{n} P(w_i \mid x_k, \mu_1 \cdots \mu_c) \frac{\partial}{\partial \mu_i} \log P(x_k \mid w_i, \mu_1 \cdots \mu_c) \quad \text{BY BAYES}$$

Eq. (7)

Case 1

$$= \sum_{k=1}^{n} P(w_i \mid x_k, \mu_1 \cdots \mu_c) \frac{\partial}{\partial \mu_i} \log \exp\left(-\frac{1}{2\sigma^2} (x_k - \mu_i)^2 \right)$$

$$= \sum_{k=1}^{n} P(w_i \mid x_k, \mu_1 \cdots \mu_c) \left(-\frac{1}{2\sigma^2} \frac{\partial}{\partial \mu_i} (x_k - \mu_i)^2 \right)$$

$$= \sum_{k=1}^{n} P(w_i \mid x_k, \mu_1 \cdots \mu_c) \frac{(x_k - \mu_i)}{\sigma^2} = 0 \quad \text{for max likelihood, so}$$

$$\underset{\sim}{\mu}_i = \frac{\sum_{k=1}^{n} P(w_i \mid \underset{\sim}{x}_k, \underset{\sim}{\mu}_1 \cdots \underset{\sim}{\mu}_c) \underset{\sim}{x}_k}{\sum_{k=1}^{n} P(w_i \mid \underset{\sim}{x}_k, \underset{\sim}{\mu}_1 \cdots \underset{\sim}{\mu}_c)}$$

Eq. (17)

- Example 1: Mixture of two 1D Gaussians

Example 1: Mixtures of two 1D Gaussians

To illustrate the kind of behavior that can occur, consider the simple two-component one-dimensional normal mixture:

$$p(x|\mu_1,\mu_2) = \underbrace{\frac{1}{3\sqrt{2\pi}}\exp\left[-\frac{1}{2}(x-\mu_1)^2\right]}_{\omega_1} + \underbrace{\frac{2}{3\sqrt{2\pi}}\exp\left[-\frac{1}{2}(x-\mu_2)^2\right]}_{\omega_2},$$

where ω_i denotes a Gaussian component. The 25 samples shown in the table were drawn sequentially from this mixture with $\mu_1 = -2$ and $\mu_2 = 2$. Let us use these samples to compute the log-likelihood function

$$l(\mu_1,\mu_2) = \sum_{k=1}^{n} \ln\, p(x_k|\mu_1,\mu_2)$$

for various values of μ_1 and μ_2. The bottom figure shows how l varies with μ_1 and μ_2. The maximum value of l occurs at $\hat{\mu}_1 = -2.130$ and $\hat{\mu}_2 = 1.668$, which is in the rough vicinity of the true values $\mu_1 = -2$ and $\mu_2 = 2$. However, l reaches another peak of comparable height at $\hat{\mu}_1 = 2.085$ and $\hat{\mu}_2 = -1.257$. Roughly speaking, this solution corresponds to interchanging μ_1 and μ_2. Note that had the prior probabilities been equal, interchanging μ_1 and μ_2 would have produced no change in the log-likelihood function. Thus, as we mentioned before, when the mixture density is not identifiable, the maximum-likelihood solution is not unique.

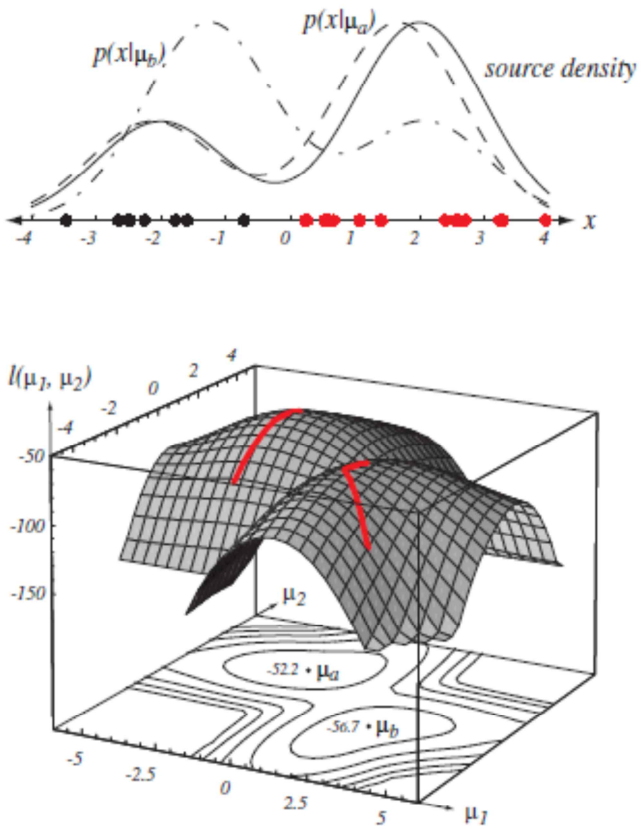
k	x_k	ω_1	ω_2
1	0.608		×
2	-1.590	×	
3	0.235		×
4	3.949		×
5	-2.249	×	
6	2.704		×
7	-2.473	×	
8	0.672		×

k	x_k	ω_1	ω_2
9	0.262		×
10	1.072		×
11	-1.773	×	
12	0.537		×
13	3.240		×
14	2.400		×
15	-2.499	×	
16	2.608		×

k	x_k	ω_1	ω_2
17	-3.458	×	
18	0.257		×
19	2.569		×
20	1.415		×
21	1.410		×
22	-2.653	×	
23	1.396		×
24	3.286		×
25	-0.712	×	

Additional insight into the nature of these multiple solutions can be obtained by examining the resulting estimates for the mixture density. The figure at the top shows the true (source) mixture density and the estimates obtained by using the two maximum-likelihood estimates as if they were the true parameter values. The 25 sample values are shown as a scatter of points along the abscissa — ω_1 points in

black, ω_2 points in red. Note that the peaks of both the true mixture density and the maximum-likelihood solutions are located so as to encompass two major groups of data points. The estimate corresponding to the smaller local maximum of the log-likelihood function has a mirror-image shape, but its peaks also encompass reasonable groups of data points. To the eye, neither of these solutions is clearly superior, and both are interesting.



(Above) The source mixture density used to generate sample data, and two maximum-likelihood estimates based on the data in the table. (Bottom) Log-likelihood of a mixture model consisting of two univariate Gaussians as a function of their means, for the data in the table. Trajectories for the iterative maximum-likelihood estimation of the means of a two-Gaussian mixture model based on the data are shown as red lines. Two local optima (with log-likelihoods -52.2 and -56.7) correspond to the two density estimates shown above.