Maximum-Likelihood Estimation for Mixture Models: the EM algorithm

1. Introduction

Thus far, we have looked at maximum-likelihood parameter estimation for simple exponential distributions, especially Gaussian models (i.e. Normal densities). We have seen that the classification capabilities of classifiers built on Gaussian modeling are limited to conic decision boundaries in d-dimensional space. As such, they may not have the requisite flexibility for modeling data distributions that are not well clustered. Therefore, we will now look at a more sophisticated modeling paradigm, namely, mixture-of-Gaussians modeling, or mixture modeling for short. In mixture models, a single statistical model is composed of the weighted sum of multiple Gaussians. As such, classifiers that use a mixture modeling representation are able to form more complex decision boundaries between classes. Unlike simple Gaussian models, where we were able to compute closed-form solutions for the maximum-likelihood parameter estimates, however, we will see that no such similar closed-form solution exists for mixture models. This will motivate our development of an iterative algorithm for estimating the maximum-likelihood parameters of mixture models called the Expectation-Maximization algorithm. The formal definition of this algorithm is nontrivial; however, before delving into the full theoretical details of Expectation-Maximization, we will gain some insight into the algorithm through a more intuitive, less rigorous formulation.

A. Mixture modeling problem formulation

Assume you are given a set of identically and independently distributed d-dimensional data $\mathbf{X} = \{\mathbf{x}_j\}$, $j \in \{1, 2, \ldots, n\}$, drawn from the probability density function,

$$p(\mathbf{x}|\Theta) = \sum_{i=1}^{k} p(\mathbf{x}|\phi_i)P(\omega_i) \tag{1}$$

where ϕ_i denotes a parameter vector fully specifying the ith component density $p(\mathbf{x}|\phi_i)$, $P(\omega_i)$ denotes the probability (i.e. weight) of the ith component density $p(\mathbf{x}|\phi_i)$, and,

$$\Theta = \{\Theta_i\}, \; i \in \{1, 2, \ldots, k\}, \text{ where,} \tag{2}$$

$$\Theta_i = \{\phi_i, P(\omega_i)\}. \tag{3}$$

Compute the maximum-likelihood parameter estimates for the parameters Θ. For the mixture-of-Gaussians model, each of the individual component densities is given by,

$$p(\mathbf{x}|\phi_i) = p(\mathbf{x}|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\mu_i)^T\Sigma_i^{-1}(\mathbf{x}-\mu_i)\right] = N(\mu_i, \Sigma_i) \tag{4}$$

such that,

$$\phi_i = \{\mu_i, \Sigma_i\} \tag{5}$$

where μ_i and Σ_i denote the mean vector and covariance matrix for the ith component density, respectively.

B. Maximum-likelihood estimation: a first look

For the maximum-likelihood solution, we want to maximize $p(\mathbf{X}|\Theta)$ with respect to Θ. That is, we want to find the set of parameters Θ^* such that,

$$p(\mathbf{X}|\Theta^*) \geq p(\mathbf{X}|\Theta), \; \forall\Theta. \tag{6}$$

For simple Gaussian modeling, we solved for Θ^* as the solution of the equation,

$$\nabla_\Theta \ln p(\mathbf{X}|\Theta) = 0 \tag{7}$$

$$\sum_{j=1}^{n} \nabla_{\Theta} \ln p(\mathbf{x}_j | \Theta) = 0 \tag{8}$$

which resulted in a closed-form solution for the Gaussian parameters $\{\mu, \Sigma\}$. For the mixture density in equation (1), however, equation (8) no longer results in a solvable set of equations. We illustrate this problem with a simple example in the following section.

C. Simple mixture modeling example

Problem statement: Let us consider a very simple mixture modeling problem. Assume you are given a set of identically and independently distributed one-dimensional data $\mathbf{X} = \{x_j\}$, $j \in \{1, 2, \ldots, n\}$, drawn from the probability density function,

$$p(x|\Theta) = \sum_{i=1}^{2} p(x|\phi_i)P(\omega_i) = p(x|\phi_1)P(\omega_1) + p(x|\phi_2)P(\omega_2) \tag{9}$$

where,

$$p(x|\phi_i) = N(\mu_i, \sigma_i^2) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[\frac{-(x-\mu_i)^2}{2\sigma_i^2}\right] \tag{10}$$

such that,

$$\phi_i = \{\mu_i, \sigma_i^2\}, \tag{11}$$

$$\Theta_i = \{\phi_i, P(\omega_i)\}, \text{ and,} \tag{12}$$

$$\Theta = \{\Theta_i\}, i \in \{1, 2\}. \tag{13}$$

Furthermore, assume that you know the following parameter values:

$$\sigma_1 = \sigma_2 = 1 \tag{14}$$

$$P(\omega_1) = 1/3, P(\omega_2) = 2/3. \tag{15}$$

Compute the maximum-likelihood estimate of the remaining unknown parameters in Θ, namely $\{\mu_1, \mu_2\}$.

Attempted solution:

$$p(x|\Theta) \equiv p(x|\mu_1, \mu_2) \text{ (notational change to emphasize } \mu_1 \text{ and } \mu_2\text{)} \tag{16}$$

$$\ln p(\mathbf{X}|\Theta) = \sum_{j=1}^{n} \ln[p(x|\mu_1, \mu_2)] \tag{17}$$

$$\ln p(\mathbf{X}|\Theta) = \sum_{j=1}^{n} \ln[p(x_j|\mu_1)P(\omega_1) + p(x_j|\mu_2)P(\omega_2)] \tag{18}$$

$$\ln p(\mathbf{X}|\Theta) = \sum_{j=1}^{n} \ln\left\{\exp\left[-\frac{1}{2}(x_j-\mu_1)^2\right](1/3) + \exp\left[-\frac{1}{2}(x_j-\mu_2)^2\right](2/3)\right\} - n\ln\sqrt{2\pi} \tag{19}$$

Equation (19) is very nonlinear in μ_1 and μ_2; consequently, the previous solution equation for finding the maximum-likelihood parameters will not yield a useful set of equations. To show this, let’s start from equations (7) and (18):

$$\nabla_{\Theta} \ln p(\mathbf{X}|\Theta) = 0 \tag{20}$$

$$\nabla_{(\mu_1,\,\mu_2)}\ln p(\mathbf{X}|\Theta) = 0 \tag{21}$$

$$\frac{\partial}{\partial\mu_i}\left(\sum_{j=1}^{n}\ln[p(x_j|\mu_1)P(\omega_1)+p(x_j|\mu_2)P(\omega_2)]\right) = 0 \tag{22}$$

$$\sum_{j=1}^{n}\frac{\frac{\partial}{\partial\mu_i}p(x_j|\mu_i)P(\omega_i)}{p(x_j|\mu_1)P(\omega_1)+p(x_j|\mu_2)P(\omega_2)} = 0 \tag{23}$$

Now, note that for the definitions in the problem statement,

$$\frac{\partial}{\partial\mu_i}p(x_j|\mu_i) = \frac{\partial}{\partial\mu_i}\exp\left[-\frac{1}{2}(x_j-\mu_i)^2\right] = \exp\left[-\frac{1}{2}(x_j-\mu_i)^2\right](x_j-\mu_i) \tag{24}$$

$$\frac{\partial}{\partial\mu_i}p(x_j|\mu_i) = p(x_j|\mu_i)(x_j-\mu_i) \tag{25}$$

so that,

$$\sum_{j=1}^{n}\frac{p(x_j|\mu_i)P(\omega_i)(x_j-\mu_i)}{p(x_j|\mu_1)P(\omega_1)+p(x_j|\mu_2)P(\omega_2)} = 0\,,\ i\in\{1,2\} \tag{26}$$

$$\sum_{j=1}^{n}\left(\frac{\exp\left[-\frac{1}{2}(x_j-\mu_i)^2\right]P(\omega_i)(x_j-\mu_i)}{\exp\left[-\frac{1}{2}(x_j-\mu_1)^2\right]P(\omega_1)+\exp\left[-\frac{1}{2}(x_j-\mu_2)^2\right]P(\omega_2)}\right) = 0\,,\ i\in\{1,2\} \tag{27}$$

The two equations defined by (27) cannot be solved readily for μ_1 and μ_2. In fact, it is not entirely clear that there is only one maxima on the log-likelihood function, as was the case for the simple Gaussian modeling problem.

D. Numeric example

Let us explore the previous example with some experimental data. We first generate 25 points from the mixture density in (9) for means $\{\mu_1, \mu_2\} = \{-2, 2\}$. The generating mixture density function and the resulting data are plotted in Figure 1 below; red data points indicate points that were generated from the first component density (with probability $P(\omega_1) = 1/3$), while blue data points indicate data points generated from the second component density (with probability $P(\omega_2) = 2/3$).

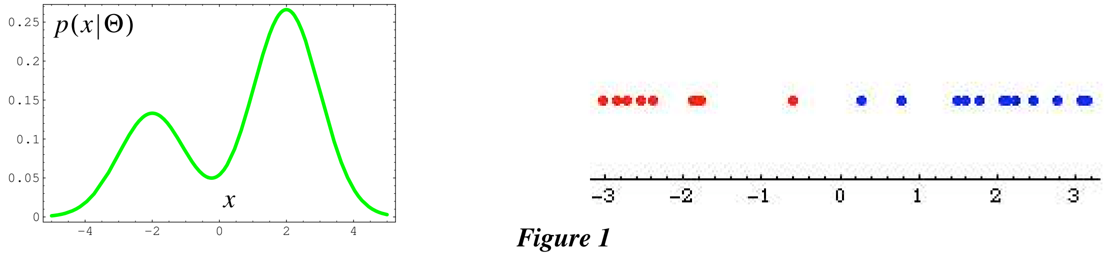


Figure 1

For the data in Figure 1, we now compute $\ln p(\mathbf{X}|\Theta)$ as a function of the "unknown" parameters μ_1 and μ_2 [see equation (19)], and plot the result as a contour plot in Figure 2. Note that the log-likelihood function $\ln p(\mathbf{X}|\Theta)$ over the whole data set has two local maxima, as indicated in Table 1 below. From Table 1, we note that the global maximum corresponds very closely to the generating mixture density means, even with

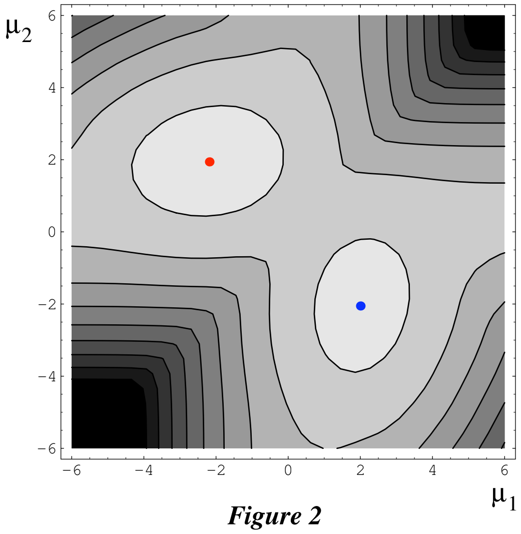


Figure 2

very limited data (25 points). Also note that the secondary peak essentially switches the two peaks, but evaluates to a lower log-likelihood over the data.

Table 1: Local maxima solutions

solution type	(μ_1, μ_2)	$\ln p(\mathbf{X}\vert\Theta)$
global maximum	$(-2.18, 1.94)$	-45.5
secondary local maximum	$(2.01, -2.05)$	-50.2
generating function	$(-2.00, 2.00)$	N/A

Finally, Figure 3 plots the estimated mixture densities for the global maximum-likelihood solution and the secondary local maximum-likelihood solution, superimposed over the original generating mixture density.

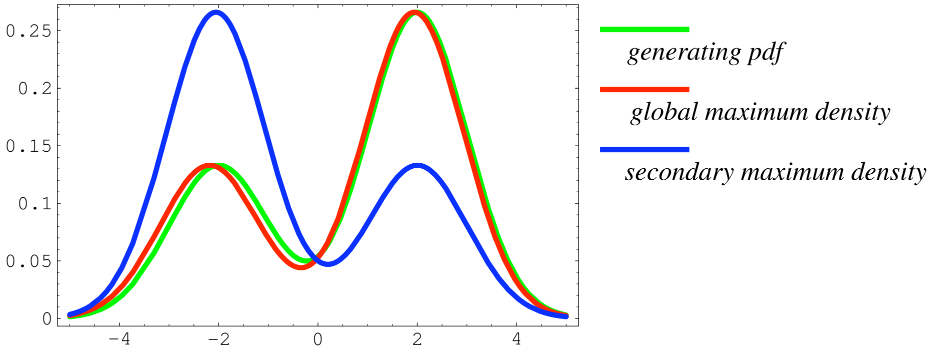


Figure 3

While these results are encouraging in one respect — namely, that we were able to recover the unknown parameters of the generating mixture density with very limited data — they are discouraging in another respect — namely, that no closed-form solution exists over a highly nonlinear log-likelihood function. What happens when we have more than two parameters over which we want to optimize, and we can't generate a nice contour plot like the one in Figure 2? While standard numerical optimization techniques are applicable, a better approach in this statistical framework is the *Expectation-Maximization algorithm*, introduced in the next section.

2. Gentle Formulation of Expectation-Maximization (EM) algorithm

A. Introduction

The principal difficulty in estimating the maximum-likelihood parameters of a mixture model is that we do not know which of the component densities $p(\mathbf{x}|\phi_i)$ generated datum $\mathbf{x}_j$; that is, we do not know the labeling of each point; For example, in Figure 1 we color coded each of the data points based on which component density generated that data point; such labeling is, unfortunately, not available in a realistic mixture modeling problem.

For the moment, however, let us assume that we do know the labeling for each datum in $\mathbf{X}$. Now, the parameter estimation problem for mixture models reduces to the simple Gaussian parameter estimation problem, since the parameters $\Theta_i = \{\mu_i, \Sigma_i, P(\omega_i)\}$ for i th component density can be computed based solely on the aggregate statistics of those data in $\mathbf{X}$ that were generated from the i th mixture density. Let us annotate the data $\mathbf{X} = \{\mathbf{x}_j\}$, $j \in \{1, 2, \ldots, n\}$ to indicate to which class each vector belongs:

$$\mathbf{X} = \{\mathbf{x}_j^{(i)}\}, j \in \{1, 2, \ldots, n_i\}, i \in \{1, 2, \ldots, k\}, \tag{28}$$

where n_i = number of data points belonging to class i. Note that,

$$\sum_{i=1}^{k} n_i = n . \tag{29}$$

Using the notation in (28), we can now write the maximum-likelihood estimates for each component density:

$$\mu_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{x}_j^{(i)} \tag{30}$$

$$\Sigma_i = \frac{1}{n_i}\sum_{j=1}^{n_i} (\mathbf{x}_j^{(i)} - \mu_i)(\mathbf{x}_j^{(i)} - \mu_i)^T \tag{31}$$

$$P(\omega_i) = n_i/n \tag{32}$$

Note that the estimates for μ_i and Σ_i are identical to the single Gaussian case, and that $P(\omega_i)$ simply computes the fraction of the data $\mathbf{X}$ belonging to the i th component density.

B. Hidden variables

We will now formalize the discussion in the previous section somewhat. We begin by introducing the notion of *hidden variables*. The basic idea behind this concept is that each observed datum $\mathbf{x}_j$ is, in fact, *incomplete*, and that the corresponding complete datum is given by,

$$\mathbf{z}_j = \{\mathbf{x}_j, \mathbf{y}_j\} \tag{33}$$

where $\mathbf{y}_j$ represents an unobserved, or *hidden* component of the j th datum, and $\mathbf{z}_j$ denotes the *complete* j th data vector. For the whole data set, $\mathbf{X} = \{\mathbf{x}_j\}$, $\mathbf{Y} = \{\mathbf{y}_j\}$ and $\mathbf{Z} = \{\mathbf{z}_j\}$, $j \in \{1, 2, \ldots, n\}$, so that,

$$\mathbf{Z} = \mathbf{X} \cup \mathbf{Y} . \tag{34}$$

In the mixture modeling problem, it should be clear what the hidden variable vector $\mathbf{y}_j$ ought to encode — namely, the labeling of each datum $\mathbf{x}_j$. One way to do this is through vectors of length k composed of simple binary random variables y_{ij}:

$$y_{ij} \equiv \begin{cases} 1 & \mathbf{x}_j \text{ belongs to class } \omega_i \\ 0 & \text{otherwise} \end{cases} \tag{35}$$

so that,

$$\mathbf{y}_j = \{y_{1j}, y_{2j}, \ldots, y_{kj}\} \tag{36}$$

$$\mathbf{y}_j = \{y_{ij}\}, \; i \in \{1, 2, \ldots, k\}. \tag{37}$$

Note that the vector $\mathbf{y}_j$ can only take on k distinct values:

$$\{1, 0, \ldots, 0\}, \{0, 1, \ldots, 0\}, \ldots, \{0, 0, \ldots, 1\}. \tag{38}$$

We can now rewrite equations (30) through (32) in terms of the complete data (observed and hidden) defined above:

$$\mu_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{x}_j^{(i)} = \left(\sum_{j=1}^{n} y_{ij}\mathbf{x}_j\right) \Big/ \left(\sum_{j=1}^{n} y_{ij}\right) \tag{39}$$

$$\Sigma_i = \frac{1}{n_i}\sum_{j=1}^{n_i} (\mathbf{x}_j^{(i)} - \mu_i)(\mathbf{x}_j^{(i)} - \mu_i)^T = \left(\sum_{j=1}^{n} y_{ij}(\mathbf{x}_j - \mu_i)(\mathbf{x}_j - \mu_i)^T\right) \Big/ \left(\sum_{j=1}^{n} y_{ij}\right) \tag{40}$$

$$P(\omega_i) = \frac{n_i}{n} = \left(\sum_{j=1}^{n} y_{ij}\right) \Big/ n \tag{41}$$

Note that in the above equations,

$$n_i = \sum_{j=1}^{n} y_{ij}. \tag{42}$$

The big question now is how we can actually use the formulation of the maximum-likelihood estimates in terms of the hidden variables in equations (39) through (41), especially since we do not know the actual value of the hidden variables. It is here where the Expectation-Maximization (EM) algorithm helps us.

C. Informal formulation of the EM algorithm

In the EM algorithm, rather than use the actual values for the hidden variables, we instead use the *expected values* of the hidden variables (y_{ij}) to compute our current maximum-likelihood estimate of the parameters (μ_i). Thus the iterative EM algorithm can be divided into two steps:

Expectation step: Calculate the expected value $E[y_{ij}|\Theta]$ for the hidden variables y_{ij}, given the current estimate for the parameters Θ.

Maximization step: Calculate a new maximum-likelihood estimate $\bar{\Theta}$ for the parameters, assuming that the value taken on by each hidden variable y_{ij} is its expected value $E[y_{ij}|\Theta]$ (calculated in the Expectation step). Then replace the old estimate Θ with the new estimate $\bar{\Theta}$ and iterate.

Since this is an iterative algorithm, the parameters Θ have to be initialized to some values prior to the first Expectation step.

D. Application to mixture modeling

Before we write the actual update equations for the parameters of the mixture-of-Gaussians model, let us derive an expression for $E[y_{ij}|\Theta]$. Applying the definition of the expectation operator, we can write,

$$E[y_{ij}|\Theta] = 0 \cdot P(y_{ij} = 0|\Theta) + 1 \cdot P(y_{ij} = 1|\Theta) \tag{43}$$

$$E[y_{ij}|\Theta] = P(y_{ij} = 1|\Theta) \tag{44}$$

Let us rewrite the right-hand-side of equation (44):

$$P(y_{ij} = 1|\Theta) = P(\omega_i|\mathbf{x}_j, \Theta) \tag{45}$$

so that:

$$E[y_{ij}|\Theta] = P(\omega_i|\mathbf{x}_j, \Theta) \tag{46}$$

Now, let us write $P(\omega_i|\mathbf{x}_j, \Theta)$ in a form that we can compute. Applying Bayes Theorem,

$$P(\omega_i|\mathbf{x}_j, \Theta) = \frac{p(\mathbf{x}_j|\Theta, \omega_i)P(\omega_i)}{p(\mathbf{x}_j|\Theta)} = \frac{p(\mathbf{x}_j|\phi_i)P(\omega_i)}{p(\mathbf{x}_j|\Theta)} \tag{47}$$

$$E[y_{ij}|\Theta] = \frac{p(\mathbf{x}_j|\phi_i)P(\omega_i)}{p(\mathbf{x}_j|\Theta)} \tag{48}$$

Note that equation (48) completely defines $E[y_{ij}|\Theta]$ in terms of computable functions [see equations (1) and (4)]. Thus, we can now write the iterative EM equations by combining equation (46) with equations (39) through (41):

$$\bar{\mu}_i = \frac{\sum_{j=1}^{n} P(\omega_i|\mathbf{x}_j, \Theta)\mathbf{x}_j}{\sum_{j=1}^{n} P(\omega_i|\mathbf{x}_j, \Theta)}, \; i \in \{1, 2, \ldots, k\} \tag{49}$$

$$\bar{\Sigma}_i = \frac{\sum_{j=1}^{n} P(\omega_i|\mathbf{x}_j, \Theta)(\mathbf{x}_j - \bar{\mu}_i)(\mathbf{x}_j - \bar{\mu}_i)^T}{\sum_{j=1}^{n} P(\omega_i|\mathbf{x}_j, \Theta)}, \; i \in \{1, 2, \ldots, k\} \tag{50}$$

$$\overline{P(\omega_i)} = \frac{1}{n}\sum_{j=1}^{n} P(\omega_i|\mathbf{x}_j, \Theta), \; i \in \{1, 2, \ldots, k\} \tag{51}$$

It is important to remember that $P(\omega_i|\mathbf{x}_j, \Theta)$ is computed in terms of the current parameter estimates Θ, while the left-hand sides of equations (49) through (51) give a new set of parameter estimates. These equations should be intuitively appealing, since the contribution of each $\mathbf{x}_j$ to the ith component density is weighted by the probability that $\mathbf{x}_j$ belongs to the ith component density.[1]

E. Another look at simple mixture modeling example

Let us take another look at the example from Section 1(C)-(D). There we attempted to solve for the maximum-likelihood estimates of the means $\{\mu_1, \mu_2\}$ starting from,

$$\nabla_{(\mu_1, \mu_2)}\ln p(\mathbf{X}|\Theta) = 0 \text{ [equation (21)].} \tag{52}$$

This formulation resulted in the following two equations:

[1] *Note that in computing new estimates for Σ_i, we use the new estimates of the means $\bar{\mu}_i$; while it may not be immediately obvious why we do this, in Section 5, we derive this update rule directly from the formal statement of the EM algorithm.*

$$\sum_{j=1}^{n} \frac{p(x_j|\mu_i)P(\omega_i)(x_j-\mu_i)}{p(x_j|\mu_1)P(\omega_1)+p(x_j|\mu_2)P(\omega_2)} = 0\,,\ i \in \{1,2\}\ \text{[equation (26)]}. \tag{53}$$

Equation (53) can be simplified using (47) above (Bayes Theorem),

$$\frac{p(x_j|\mu_i)P(\omega_i)}{p(x_j|\mu_1)P(\omega_1)+p(x_j|\mu_2)P(\omega_2)} = P(\omega_i|x_j,\mu_1,\mu_2) = P(\omega_i|x_j,\Theta) \tag{54}$$

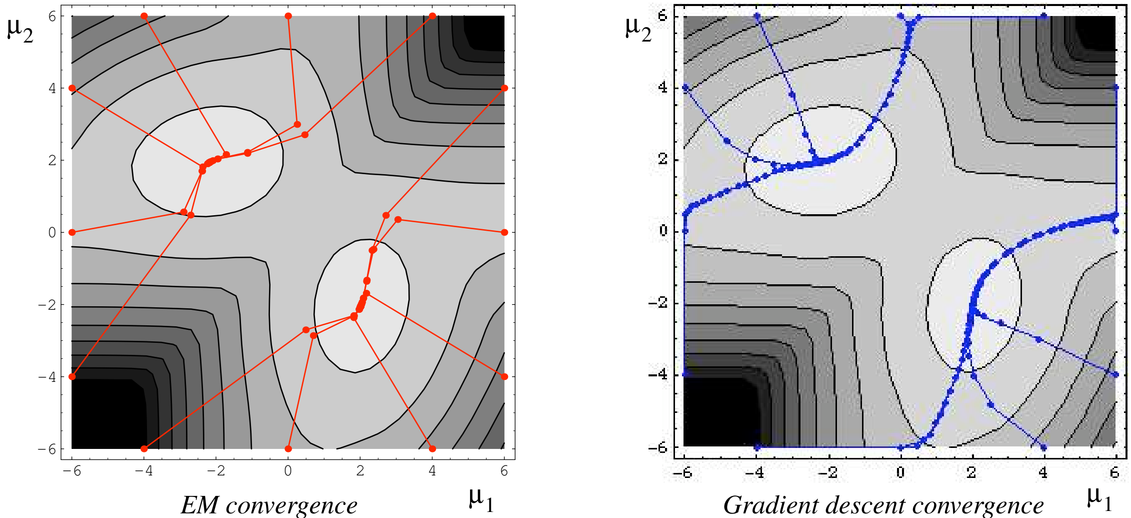
$$\sum_{j=1}^{n} P(\omega_i|x_j,\Theta)(x_j-\mu_i) = 0\,,\ i \in \{1,2\}\,. \tag{55}$$

Solving equation (55) for μ_i,

$$\mu_i = \frac{\sum_{j=1}^{n} P(\omega_i|x_j,\Theta)x_j}{\sum_{j=1}^{n} P(\omega_i|x_j,\Theta)} \tag{56}$$

Note that if we use equation (56) in an iterative manner, it is identical to the EM update rule in equation (49), which should be a little surprising, since the two results were arrived at from completely different formulations.

Below, we plot EM trajectories for the same data as in Section 1(D), and some different initial values for μ_1 and μ_2 (Figure 4, left plot, red lines). For these trajectories, the EM algorithm converges on average in about 13 steps, which compares quite favorably to the gradient descent-algorithm, which converges on average in about 50 steps for the same initial parameter values and a hand-optimized learning rate of $\eta = 0.04$ (Figure 4, right plot, blue lines).



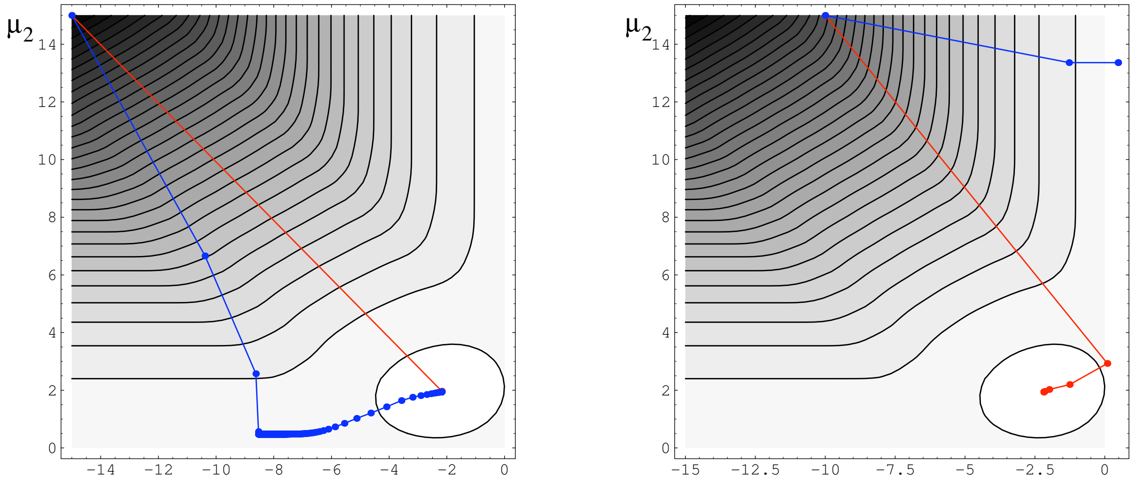
EM convergence

Gradient descent convergence

Figure 4

The gradient-descent algorithm for statistical modeling suffers from two main deficiencies: (1) it converges slowly, despite careful hand selection of the learning rate; and (2) it requires explicit computation of the gradient of the log-likelihood function $\nabla_{(\mu_1,\mu_2)}\ln p(\mathbf{X}|\Theta)$, which the EM algorithm does not. Note that both the EM and gradient-descent algorithms are *not* guaranteed to converge to a *global* maximum of the log-likelihood function, only a *local* maximum, depending on initial parameter values. In fact, the gradient-descent algorithm can fail to converge altogether, as shown in Figure 5. In Figure 5 we plot two EM and gradient descent trajectories in parameter space with initial parameter values of $\{\mu_1,\mu_2\} = \{-15, 15\}$ and

$\{\mu_1, \mu_2\} = \{-10, 15\}$, respectively. Note that in the first instance, gradient descent requires 11,890 steps to converge (compared to 9 for EM), while in the second instance, gradient descent fails to converge entirely, getting stuck in a part of the log-likelihood function where $\nabla_{(\mu_1, \mu_2)} \ln p(\mathbf{X}|\Theta) \approx 0$.



EM (red line, 9 steps)
Gradient descent (blue line, 11,890 steps)

Figure 5

EM (red line, 12 steps)
Gradient descent (blue line, fails)

3. Generalized Formulation of Expectation-Maximization (EM) algorithm

The previous section gives an intuitive, informal formulation of the EM algorithm, and shows how the EM algorithm can be used to solve the maximum-likelihood parameter estimation problem in mixture-of-Gaussians modeling. In this section, we formalize and generalize our previous discussion on the maximum-likelihood estimation problem for data with hidden components. We start with a formal definition of the problem and the corresponding EM algorithm, as stated in Figure 6 (next page).

This generalized formulation of the EM algorithm is not easy to understand. It does, however, describe a powerful algorithm for maximum-likelihood estimation that is applicable not only for mixture-of-Gaussians modeling, but for vector quantization, general mixture modeling and hidden Markov models as well. The EM algorithm is particularly applicable when the Q-function is easy to compute and is easier to maximize than maximizing the log-likelihood $L(\mathbf{X}|\Theta) = \ln p(\mathbf{X}|\Theta)$ directly. In the following two sections, we will first prove that maximizing the Q-function also maximizes the log-likelihood function; that is, we will show that,

$$Q(\Theta, \bar{\Theta}) \geq Q(\Theta, \Theta) \text{ implies } L(\mathbf{X}|\bar{\Theta}) \geq L(\mathbf{X}|\Theta). \tag{57}$$

Second, we will derive the EM solution for mixture-of-Gaussians modeling, starting with the general formulation in Figure 6; not surprisingly, we will see that this exercise will lead us to the same parameter update equations as the informal statement of the EM algorithm did [see equations (49) through (51)].

4. Convergence proof for the EM algorithm

In this section, we will prove that maximizing the Q-function also maximizes the log-likelihood function; that is, we will show that,

$$Q(\Theta, \bar{\Theta}) \geq Q(\Theta, \Theta) \text{ implies } L(\mathbf{X}|\bar{\Theta}) \geq L(\mathbf{X}|\Theta). \tag{58}$$

Below, we will break the proof into two steps. First, we will show that property (58) holds for a *single* datum $\mathbf{x}$. Then will generalize this result to multiple observations $\mathbf{X} = \{\mathbf{x}_j\}$, $j \in \{1, 2, \ldots, n\}$.

A. Relating log-likelihood and Q-function for one observation

Let us denote,

$$q(\Theta, \bar{\Theta}) = E[l(\mathbf{z}|\bar{\Theta})|\mathbf{x}, \Theta]. \tag{59}$$

Expectation-Maximization (EM) algorithm

Problem statement: Let $\mathbf{X} = \{\mathbf{x}_j\}$, $j \in \{1, 2, \ldots, n\}$, denote n independently and identically distributed *observed* (incomplete) data vectors; let $\mathbf{Y} = \{\mathbf{y}_j\}$, $j \in \{1, 2, \ldots, n\}$, denote the corresponding *unobserved* (hidden) data vectors for the n observed vectors $\mathbf{X}$; and let $\mathbf{Z} = \{\mathbf{z}_j\}$, $j \in \{1, 2, \ldots, n\}$, denote the *complete* data, where,

$$\mathbf{z}_j = \{\mathbf{x}_j, \mathbf{y}_j\}, j \in \{1, 2, \ldots, n\}, \text{ and,} \tag{EM-1}$$

$$\mathbf{Z} = \mathbf{X} \cup \mathbf{Y}. \tag{EM-2}$$

Furthermore, let $p(\mathbf{x}|\Theta)$ and $p(\mathbf{y}|\Theta)$ be members of a parametric family of probability density functions (pdfs) defined for sufficient parameters Θ. Find the parameters Θ^* that maximize the log-likelihood $L(\mathbf{X}|\Theta)$,

$$L(\mathbf{X}|\Theta) = \sum_{j=1}^{n} l(\mathbf{x}_j|\Theta) = \sum_{j=1}^{n} \ln p(\mathbf{x}_j|\Theta) \tag{EM-3}$$

such that,

$$L(\mathbf{X}|\Theta) \le L(\mathbf{X}|\Theta^*), \ \forall\Theta. \tag{EM-4}$$

Solution:

1. Choose an initial estimate for Θ.
2. **(E)xpectation step**: Compute $Q(\Theta, \bar{\Theta})$,

$$Q(\Theta, \bar{\Theta}) = E[L(\mathbf{Z}|\bar{\Theta})|\mathbf{X}, \Theta] \text{, where,} \tag{EM-5}$$

$$L(\mathbf{Z}|\bar{\Theta}) = \sum_{j=1}^{n} l(\mathbf{z}_j|\bar{\Theta}) = \sum_{j=1}^{n} \ln p(\mathbf{x}_j, \mathbf{y}_j|\bar{\Theta}). \tag{EM-6}$$

3. **(M)aximization step**: Replace the current estimate Θ with the new estimate $\bar{\Theta}$ where,

$$\bar{\Theta} = \underset{\bar{\Theta}}{\operatorname{argmax}} \ Q(\Theta, \bar{\Theta}) \tag{EM-7}$$

4. Iterate steps 2 and 3 until convergence.

Figure 6

We will now show that:

$$q(\Theta, \bar{\Theta}) \ge q(\Theta, \Theta) \text{ implies } l(\mathbf{x}|\bar{\Theta}) \ge l(\mathbf{x}|\Theta). \tag{60}$$

for a single datum $\mathbf{x}$. Let us first find an expression for $l(\mathbf{x}|\Theta)$ in terms of both the observed data $\mathbf{x}$ and the corresponding hidden data $\mathbf{y}$. Throughout our derivation, we will assume discrete hidden variables $\mathbf{y}$; results for continuous hidden variables follow trivially. From the basic laws of conditional probabilities, we can write,

$$p(\mathbf{x}, \mathbf{y}|\Theta) = p(\mathbf{y}|\mathbf{x}, \Theta)p(\mathbf{x}|\Theta) \tag{61}$$

so that,

$$p(\mathbf{x}|\Theta) = \frac{p(\mathbf{x}, \mathbf{y}|\Theta)}{p(\mathbf{y}|\mathbf{x}, \Theta)} \tag{62}$$

Thus, $l(\mathbf{x}, \Theta)$ can be expressed as,

$$l(\mathbf{x}, \Theta) \equiv \ln p(\mathbf{x}|\Theta) = \ln p(\mathbf{x}, \mathbf{y}|\Theta) - \ln p(\mathbf{y}|\mathbf{x}, \Theta). \tag{63}$$

Now, consider two parameter vectors Θ and $\bar{\Theta}$. The expectation of the incomplete log-likelihood $l(\mathbf{x}, \bar{\Theta})$ over the complete data $\mathbf{z} = \{\mathbf{x}, \mathbf{y}\}$ conditioned on $\mathbf{x}$ and Θ is given by,

$$E[l(\mathbf{x}, \bar{\Theta})|\mathbf{x}, \Theta] = E[\{\ln p(\mathbf{x}, \mathbf{y}|\bar{\Theta}) - \ln p(\mathbf{y}|\mathbf{x}, \bar{\Theta})\}|\mathbf{x}, \Theta] \tag{64}$$

$$\underset{(a)}{E[l(\mathbf{x}, \bar{\Theta})|\mathbf{x}, \Theta]} = \underset{(b)}{E[\ln p(\mathbf{x}, \mathbf{y}|\bar{\Theta})|\mathbf{x}, \Theta]} - \underset{(c)}{E[\ln p(\mathbf{y}|\mathbf{x}, \bar{\Theta})|\mathbf{x}, \Theta]} \tag{65}$$

Let us pause and think about what is meant by the conditional expectation operator $E[\ \bullet\ |\mathbf{x}, \Theta]$ used in equations (65). First, consider some function $f(\mathbf{x}, \mathbf{y})$ for which $\mathbf{x}$ is a known constant and $\mathbf{y}$ is a vector of discrete random variables. For such a function,

$$E[f(\mathbf{x}, \mathbf{y})] \equiv \sum_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) p(\mathbf{y}) \tag{66}$$

Note that we do not sum up over all possible values of $\mathbf{x}$ because $\mathbf{x}$ is known and constant. The conditional expected value $E[f(\mathbf{x}, \mathbf{y})|\mathbf{x}, \Theta]$ is similarly given by,

$$E[f(\mathbf{x}, \mathbf{y})|\mathbf{x}, \Theta] \equiv \sum_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) p(\mathbf{y}|\mathbf{x}, \Theta) \text{ (discrete variables } \mathbf{y}\text{)} \tag{67}$$

Let us now derive expressions for each of the terms in equation (65). First consider term (a):

$$\begin{aligned} E[l(\mathbf{x}, \bar{\Theta})|\mathbf{x}, \Theta] &= E[\ln p(\mathbf{x}|\bar{\Theta})|\mathbf{x}, \Theta] \\ &= \sum_{\mathbf{y}} [\ln p(\mathbf{x}|\bar{\Theta})] p(\mathbf{y}|\mathbf{x}, \Theta) \\ &= \ln p(\mathbf{x}|\bar{\Theta}) \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}, \Theta) \end{aligned} \tag{68}$$

Note that in equation (68),

$$\sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}, \Theta) = 1 \tag{69}$$

since the probability that $\mathbf{y}$ assumes any of its possible values is equal to one. Therefore, equation (68) reduces to:

$$E[l(\mathbf{x}, \bar{\Theta})|\mathbf{x}, \Theta] = \ln p(\mathbf{x}|\bar{\Theta}) = l(\mathbf{x}, \bar{\Theta}) \tag{70}$$

Next, consider term (b):

$$E[\ln p(\mathbf{x}, \mathbf{y}|\bar{\Theta})|\mathbf{x}, \Theta] = E[l(\mathbf{z}|\bar{\Theta})|\mathbf{x}, \Theta] \equiv q(\Theta, \bar{\Theta}) \tag{71}$$

Finally, consider term (c):

$$E[\ln p(\mathbf{y}|\mathbf{x}, \bar{\Theta})|\mathbf{x}, \Theta] = \sum_{\mathbf{y}} \ln p(\mathbf{y}|\mathbf{x}, \bar{\Theta}) p(\mathbf{y}|\mathbf{x}, \Theta) \tag{72}$$

Since we cannot simplify equation (72), let us denote it as $h(\Theta, \bar{\Theta})$:

$$h(\Theta, \bar{\Theta}) = \sum_{\mathbf{y}} \ln p(\mathbf{y}|\mathbf{x}, \bar{\Theta}) p(\mathbf{y}|\mathbf{x}, \Theta). \tag{73}$$

Let us now substitute equations (70), (71) and (73) into (65):

$$l(\mathbf{x}, \bar{\Theta}) = q(\Theta, \bar{\Theta}) - h(\Theta, \bar{\Theta}) \tag{74}$$

Equation (74) gives us a relationship between the log-likelihood of $\mathbf{x}$ with respect to the new parameter vector $\bar{\Theta}$ and the Q-function $q(\Theta, \bar{\Theta})$ for a single observation. In the next section, we will exploit that relationship to show that $q(\Theta, \bar{\Theta}) \geq q(\Theta, \Theta)$ implies $l(\mathbf{x}|\bar{\Theta}) \geq l(\mathbf{x}|\Theta)$.

B. Jensen's inequality

We will now show that,

$$h(\Theta, \bar{\Theta}) \leq h(\Theta, \Theta) . \tag{75}$$

Equation (75) is known as *Jensen's inequality*. Let us begin with the definition of $h(\Theta, \bar{\Theta})$:

$$h(\Theta, \bar{\Theta}) = \sum_{\mathbf{y}} \ln p(\mathbf{y}|\mathbf{x}, \bar{\Theta}) p(\mathbf{y}|\mathbf{x}, \Theta) \text{ [same as (73)]} \tag{76}$$

For $h(\Theta, \Theta)$:

$$h(\Theta, \Theta) = \sum_{\mathbf{y}} \ln p(\mathbf{y}|\mathbf{x}, \Theta) p(\mathbf{y}|\mathbf{x}, \Theta) \tag{77}$$

Let us now subtract (77) from (76):

$$h(\Theta, \bar{\Theta}) - h(\Theta, \Theta) = \sum_{\mathbf{y}} \ln p(\mathbf{y}|\mathbf{x}, \bar{\Theta}) p(\mathbf{y}|\mathbf{x}, \Theta) - \sum_{\mathbf{y}} \ln p(\mathbf{y}|\mathbf{x}, \Theta) p(\mathbf{y}|\mathbf{x}, \Theta) \tag{78}$$

$$h(\Theta, \bar{\Theta}) - h(\Theta, \Theta) = \sum_{\mathbf{y}} [\ln p(\mathbf{y}|\mathbf{x}, \bar{\Theta}) p(\mathbf{y}|\mathbf{x}, \Theta) - \ln p(\mathbf{y}|\mathbf{x}, \Theta) p(\mathbf{y}|\mathbf{x}, \Theta)] \tag{79}$$

$$h(\Theta, \bar{\Theta}) - h(\Theta, \Theta) = \sum_{\mathbf{y}} \left[\ln \frac{p(\mathbf{y}|\mathbf{x}, \bar{\Theta})}{p(\mathbf{y}|\mathbf{x}, \Theta)} \right] p(\mathbf{y}|\mathbf{x}, \Theta) \tag{80}$$

Now, we observe the following inequality (as depicted in Figure 7):

$$\ln x \leq (x - 1) , \forall x . \tag{81}$$

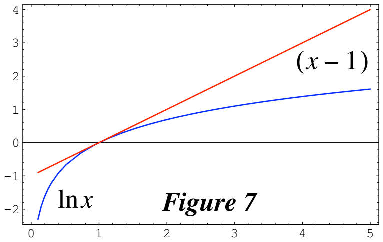


Figure 7

Since $p(\mathbf{y}|\mathbf{x}, \Theta) \geq 0$ we can combine (80) and (81):

$$h(\Theta, \bar{\Theta}) - h(\Theta, \Theta) = \sum_{\mathbf{y}} \left[\ln \frac{p(\mathbf{y}|\mathbf{x}, \bar{\Theta})}{p(\mathbf{y}|\mathbf{x}, \Theta)} \right] p(\mathbf{y}|\mathbf{x}, \Theta) \leq \sum_{\mathbf{y}} \left[\frac{p(\mathbf{y}|\mathbf{x}, \bar{\Theta})}{p(\mathbf{y}|\mathbf{x}, \Theta)} - 1 \right] p(\mathbf{y}|\mathbf{x}, \Theta) \tag{82}$$

$$h(\Theta, \bar{\Theta}) - h(\Theta, \Theta) \leq \sum_{\mathbf{y}} \left[\frac{p(\mathbf{y}|\mathbf{x}, \bar{\Theta})}{p(\mathbf{y}|\mathbf{x}, \Theta)} - 1 \right] p(\mathbf{y}|\mathbf{x}, \Theta) \tag{83}$$

$$h(\Theta, \bar{\Theta}) - h(\Theta, \Theta) \leq \sum_{\mathbf{y}} [p(\mathbf{y}|\mathbf{x}, \bar{\Theta}) - p(\mathbf{y}|\mathbf{x}, \Theta)] \tag{84}$$

$$h(\Theta, \bar{\Theta}) - h(\Theta, \Theta) \leq \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}, \bar{\Theta}) - \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}, \Theta) \tag{85}$$

Note that in equation (85),

$$\sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}, \Theta) = \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}, \bar{\Theta}) = 1 \tag{86}$$

so that (85) reduces to:

$$h(\Theta, \bar{\Theta}) - h(\Theta, \Theta) \le 0 \tag{87}$$

$$h(\Theta, \bar{\Theta}) \le h(\Theta, \Theta) \tag{88}$$

C. Corollary to Jensen's inequality

Recalling equation (74),

$$l(\mathbf{x}, \bar{\Theta}) = q(\Theta, \bar{\Theta}) - h(\Theta, \bar{\Theta}) \tag{89}$$

note that Jensen's inequality directly leads to the following corollary:

$$q(\Theta, \bar{\Theta}) \ge q(\Theta, \Theta) \text{ implies } l(\mathbf{x}|\bar{\Theta}) \ge l(\mathbf{x}|\Theta)\,. \tag{90}$$

In other words, maximizing the Q-function for a single data point implies an increase in the log-likelihood function $l(\mathbf{x}|\Theta)$.

D. Multiple observations

We now want to generalize the result of the previous section for a single data $\mathbf{x}$ to multiple data $\mathbf{X} = \{\mathbf{x}_j\}$, $j \in \{1, 2, \ldots, n\}$. From equations (EM-3) and (63),

$$L(\mathbf{X}|\Theta) = \sum_{j=1}^{n} \ln p(\mathbf{x}_j|\Theta) = \sum_{j=1}^{n} \ln p(\mathbf{x}_j, \mathbf{y}_j|\Theta) - \sum_{j=1}^{n} \ln p(\mathbf{y}_j|\mathbf{x}_j, \Theta) \tag{91}$$

$$L(\mathbf{X}|\Theta) = L(\mathbf{Z}|\Theta) - L(\mathbf{Y}|\mathbf{X}, \Theta) \tag{92}$$

Changing Θ to $\bar{\Theta}$, and applying the conditional expectation operator $E[\,\bullet\,|\mathbf{X}, \Theta]$ to equation (92), we get,

$$E[L(\mathbf{X}|\bar{\Theta})|\mathbf{X}, \Theta] = E[L(\mathbf{Z}|\bar{\Theta})|\mathbf{X}, \Theta] - E[L(\mathbf{Y}|\mathbf{X}, \bar{\Theta})|\mathbf{X}, \Theta] \tag{93}$$

Let us now expand each of the terms in equation (93). First, we compute $E[L(\mathbf{X}|\bar{\Theta})|\mathbf{X}, \Theta]$:

$$\begin{aligned} E[L(\mathbf{X}|\bar{\Theta})|\mathbf{X}, \Theta] &= E\left[\sum_{j=1}^{n} \ln p(\mathbf{x}_j|\bar{\Theta}) \middle| \mathbf{X}, \Theta\right] \\ &= \sum_{\mathbf{y}}\left[\sum_{j=1}^{n} \ln p(\mathbf{x}_j|\bar{\Theta}) \prod_{l=1}^{n} p(\mathbf{y}_l|\mathbf{x}_l, \Theta)\right] \\ &= \sum_{j=1}^{n} \ln p(\mathbf{x}_j|\bar{\Theta}) \prod_{l=1}^{n}\left[\sum_{\mathbf{y}_l} p(\mathbf{y}_l|\mathbf{x}_l, \Theta)\right] \\ &= \sum_{j=1}^{n} \ln p(\mathbf{x}_j|\bar{\Theta}) \\ &= L(\mathbf{X}|\bar{\Theta}) \end{aligned} \tag{94}$$

The second term is, by definition (EM-5):

$$Q(\Theta, \bar{\Theta}) \equiv E[L(\mathbf{Z}|\bar{\Theta})|\mathbf{X}, \Theta] \tag{95}$$

In terms of functions $q_j(\Theta, \bar{\Theta})$,

$$q_j(\Theta, \bar{\Theta}) \equiv E[l(\mathbf{z}_j|\bar{\Theta})|\mathbf{x}_j, \Theta]\,, \tag{96}$$

$$\begin{aligned}
Q(\Theta, \bar{\Theta}) &= E\left[\sum_{j=1}^{n} \ln p(\mathbf{x}_j, \mathbf{y}_j|\bar{\Theta})\middle|\mathbf{X}, \Theta\right] \\
&= \sum_{\mathbf{y}}\left[\sum_{j=1}^{n} \ln p(\mathbf{x}_j, \mathbf{y}_j|\bar{\Theta})\prod_{l=1}^{n} p(\mathbf{y}_l|\mathbf{x}_l, \Theta)\right] \\
&= \sum_{j=1}^{n}\left(\sum_{\mathbf{y}_j} \ln p(\mathbf{x}_j, \mathbf{y}_j|\bar{\Theta}) p(\mathbf{y}_j|\mathbf{x}_j, \Theta)\right)\prod_{l \neq j}\left[\sum_{\mathbf{y}_l} p(\mathbf{y}_l|\mathbf{x}_l, \Theta)\right] \\
&= \sum_{j=1}^{n}\left(\sum_{\mathbf{y}_j} \ln p(\mathbf{x}_j, \mathbf{y}_j|\bar{\Theta}) p(\mathbf{y}_j|\mathbf{x}_j, \Theta)\right) \\
&= \sum_{j=1}^{n}\left(\sum_{\mathbf{y}_j} \ln p(\mathbf{z}_j|\bar{\Theta}) p(\mathbf{y}_j|\mathbf{x}_j, \Theta)\right) \\
&\equiv \sum_{j=1}^{n} q_j(\Theta, \bar{\Theta})
\end{aligned} \tag{97}$$

Finally, we compute $H(\Theta, \bar{\Theta}) = E[L(\mathbf{Y}|\mathbf{X}, \bar{\Theta})|\mathbf{X}, \Theta]$. Similar to equation (97),

$$\begin{aligned}
H(\Theta, \bar{\Theta}) &= E\left[\sum_{j=1}^{n} \ln p(\mathbf{y}_j|\mathbf{x}_j, \bar{\Theta})\middle|\mathbf{X}, \Theta\right] \\
&= \sum_{\mathbf{y}}\left[\sum_{j=1}^{n} \ln p(\mathbf{y}_j|\mathbf{x}_j, \bar{\Theta})\prod_{l=1}^{n} p(\mathbf{y}_l|\mathbf{x}_l, \Theta)\right] \\
&= \sum_{j=1}^{n}\left(\sum_{\mathbf{y}_j} \ln p(\mathbf{y}_j|\mathbf{x}_j, \bar{\Theta}) p(\mathbf{y}_j|\mathbf{x}_j, \Theta)\right)\prod_{l \neq j}\left[\sum_{\mathbf{y}_l} p(\mathbf{y}_l|\mathbf{x}_l, \Theta)\right] \\
&= \sum_{j=1}^{n}\left(\sum_{\mathbf{y}_j} \ln p(\mathbf{y}_j|\mathbf{x}_j, \bar{\Theta}) p(\mathbf{y}_j|\mathbf{x}_j, \Theta)\right) \\
&\equiv \sum_{j=1}^{n} h_j(\Theta, \bar{\Theta})
\end{aligned} \tag{98}$$

$$h_j(\Theta, \bar{\Theta}) = E[\ln p(\mathbf{y}_j|\mathbf{x}_j, \bar{\Theta})|\mathbf{x}_j, \Theta]\,. \tag{99}$$

Combining the above results, we get,

$$L(\mathbf{X}, \bar{\Theta}) = Q(\Theta, \bar{\Theta}) - H(\Theta, \bar{\Theta}) \tag{100}$$

$$\sum_{j=1}^{n} \ln p(\mathbf{x}_j|\bar{\Theta}) = \sum_{j=1}^{n} q_j(\Theta, \bar{\Theta}) - \sum_{j=1}^{n} h_j(\Theta, \bar{\Theta}) \tag{101}$$

Since equation (101) simply represents a summation over individual data $\{\mathbf{x}_j\}$, we observe that the same result holds as before, namely:

$$Q(\Theta, \bar{\Theta}) \geq Q(\Theta, \Theta) \text{ implies } L(\mathbf{X}|\bar{\Theta}) \geq L(\mathbf{X}|\Theta)\,. \tag{102}$$

E. Concluding thoughts

We have now shown that improving the Q-function at each step of the EM algorithm will also improve the log-likelihood of the data given the new parameters. This is an important result, since it tells us that every step of an EM update for a specific problem will always improve the current log-likelihood $L(\mathbf{X}|\Theta)$ until we reach a local maximum on the log-likelihood function. In the next section, we will derive the EM update equations for mixture-of-Gaussians modeling.

5. Maximum-likelihood solution for mixture-of-Gaussians modeling

A. Problem statement

Assume you are given a set of identically and independently distributed d-dimensional data $\mathbf{X} = \{\mathbf{x}_j\}$, $j \in \{1, 2, \ldots, n\}$, drawn from the probability density function,

$$p(\mathbf{x}|\Theta) = \sum_{i=1}^{k} p(\mathbf{x}|\phi_i)P(\omega_i) \tag{103}$$

where ϕ_i denotes a parameter vector fully specifying the ith component density $p(\mathbf{x}|\phi_i)$, $P(\omega_i)$ denotes the probability (i.e. weight) of the ith component density $p(\mathbf{x}|\phi_i)$,

$$p(\mathbf{x}|\phi_i) = p(\mathbf{x}|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\mu_i)^T\Sigma_i^{-1}(\mathbf{x}-\mu_i)\right] = N[\mathbf{x}, \mu_i, \Sigma_i] \tag{104}$$

$$\phi_i = \{\mu_i, \Sigma_i\}\text{,} \tag{105}$$

$$\Theta_i = \{\phi_i, P(\omega_i)\}\text{, and,} \tag{106}$$

$$\Theta = \{\Theta_i\}\text{, } i \in \{1, 2, \ldots, k\}\text{.} \tag{107}$$

Compute the maximum-likelihood parameter estimates for the parameters Θ.

B. Definition of hidden variables

Solution: Let us begin with the definition of the EM algorithm. First, we need to define our hidden variables. Second, we need to perform the *expectation* step of the EM algorithm for this problem; that is, we need to compute $Q(\Theta, \overline{\Theta})$ as defined in equation (EM-5). Third, we need to perform the *maximization* step of the EM algorithm, as indicated in equation (EM-7).

Let us define the *hidden* data vector $\mathbf{y}_j$ corresponding to the *observed* datum $\mathbf{x}_j$ as,

$$\mathbf{y}_j = \{y_{1j}, y_{2j}, \ldots, y_{kj}\} \tag{108}$$

where the y_{ij} are discrete random variables with the following possible values,

$$y_{ij} \equiv \begin{cases} 1 & \mathbf{x}_j \text{ belongs to class } \omega_i \\ 0 & \text{otherwise} \end{cases} \tag{109}$$

Note that the vector $\mathbf{y}_j$ can only take on k distinct values:

$$\{1, 0, \ldots, 0\}, \{0, 1, \ldots, 0\}, \ldots, \{0, 0, \ldots, 1\}\text{.} \tag{110}$$

Finally, let $\mathbf{y}_j^{(i)}$ denote the vector $\mathbf{y}_j$ where $y_{ij} = 1$ and all other $y_{lj} = 0$, $\forall l \neq i$, and let $\mathbf{z}_j = \{\mathbf{x}_j, \mathbf{y}_j\}$ denote the jth complete data vector.

C. Expectation step

Previously, we shown that the Q-function defined in (EM-5),

$$Q(\Theta,\bar{\Theta}) = E[L(\mathbf{Z}|\bar{\Theta})|\mathbf{X},\Theta] \tag{111}$$

can be written as,

$$Q(\Theta,\bar{\Theta}) = \sum_{j=1}^{n}\left(\sum_{\mathbf{y}_j}\ln p(\mathbf{x}_j,\mathbf{y}_j|\bar{\Theta})p(\mathbf{y}_j|\mathbf{x}_j,\Theta)\right). \tag{112}$$

From basic probability theory,

$$p(\mathbf{y}_j|\mathbf{x}_j,\Theta) = \frac{p(\mathbf{x}_j,\mathbf{y}_j|\Theta)}{p(\mathbf{x}_j|\Theta)} \tag{113}$$

so that,

$$Q(\Theta,\bar{\Theta}) = \sum_{j=1}^{n}\sum_{\mathbf{y}_j}\ln p(\mathbf{x}_j,\mathbf{y}_j|\bar{\Theta})\frac{p(\mathbf{x}_j,\mathbf{y}_j|\Theta)}{p(\mathbf{x}_j|\Theta)} \tag{114}$$

$$Q(\Theta,\bar{\Theta}) = \sum_{j=1}^{n}\sum_{i=1}^{k}\ln p(\mathbf{x}_j,\mathbf{y}_j^{(i)}|\bar{\Theta})\frac{p(\mathbf{x}_j,\mathbf{y}_j^{(i)}|\Theta)}{p(\mathbf{x}_j|\Theta)}. \tag{115}$$

Let us derive some of the sub-expressions in the above Q-function.

$$p(\mathbf{x}_j,\mathbf{y}_j^{(i)}|\Theta) = p(\mathbf{x}_j|\phi_i)P(\omega_i) \tag{116}$$

$$p(\mathbf{x}_j|\Theta) = \sum_{l=1}^{k}p(\mathbf{x}_j|\phi_l)P(\omega_l) \tag{117}$$

$$\begin{aligned}\ln p(\mathbf{x}_j,\mathbf{y}_j^{(i)}|\bar{\Theta}) &= \ln p(\mathbf{x}_j|\bar{\phi}_i)\overline{P(\omega_i)}\\ &= \ln\overline{P(\omega_i)} + \ln p(\mathbf{x}_j|\bar{\phi}_i)\end{aligned} \tag{118}$$

From equations (115) through (118) and switching the order of summation,

$$Q(\Theta,\bar{\Theta}) = \sum_{i=1}^{k}c_i\ln\overline{P(\omega_i)} + \sum_{i=1}^{k}\sum_{j=1}^{n}\frac{p(\mathbf{x}_j|\phi_i)P(\omega_i)}{p(\mathbf{x}_j|\Theta)}\ln p(\mathbf{x}_j|\bar{\phi}_i) \tag{119}$$

where,

$$c_i = \sum_{j=1}^{n}\frac{p(\mathbf{x}_j|\phi_i)P(\omega_i)}{p(\mathbf{x}_j|\Theta)} \tag{120}$$

This completes the *Expectation* step of the EM algorithm. What remains is to maximize $Q(\Theta,\bar{\Theta})$ with respect to $\bar{\Theta}$. This will be the *Maximization* step of the EM algorithm. We will perform this maximization by maximizing each of the two terms in equation (120) separately.

D. Maximization step

First, we want to maximize,

$$\sum_{i=1}^{k}c_i\ln\overline{P(\omega_i)} \tag{121}$$

with respect to the $\overline{P(\omega_i)}$. Note that the $\overline{P(\omega_i)}$ are constrained by,

$$\sum_{i=1}^{k} \overline{P(\omega_i)} = 1 \tag{122}$$

Remember from your basic calculus, that this type of *constrained optimization* can be performed through the method of *Lagrange multipliers*.

[Note: In general, a function,

$$f(\mathbf{x}) \tag{123}$$

with constraint,

$$g(\mathbf{x}) = a \tag{124}$$

is maximized by maximizing the augmented objective function $h(\mathbf{x})$,

$$h(\mathbf{x}) = f(\mathbf{x}) + \lambda g(\mathbf{x}), \tag{125}$$

where λ is called the Lagrange multiplier.]

In our case,

$$h = \sum_{i=1}^{k} c_i \ln \overline{P(\omega_i)} + \lambda \sum_{i=1}^{k} \overline{P(\omega_i)} \tag{126}$$

and, switching the index of summation from i to l,

$$h = \sum_{l=1}^{k} [c_l \ln \overline{P(\omega_l)} + \lambda \overline{P(\omega_l)}] \tag{127}$$

Taking the derivative of (127) with respect to $\overline{P(\omega_i)}$,

$$\frac{\partial h}{\partial \overline{P(\omega_i)}} = \frac{c_i}{\overline{P(\omega_i)}} + \lambda = 0 \tag{128}$$

$$c_i + \lambda \overline{P(\omega_i)} = 0 \tag{129}$$

Let us now sum equation (129) over all i ,and solve for λ:

$$\sum_{l=1}^{k} [c_l + \lambda \overline{P(\omega_l)}] = 0 \tag{130}$$

$$\lambda = -\sum_{l=1}^{k} c_l \Big/ \sum_{l=1}^{k} \overline{P(\omega_l)} \tag{131}$$

Note, however that,

$$\sum_{l=1}^{k} \overline{P(\omega_l)} = 1 \tag{132}$$

Therefore, equation (131) simplifies to,

$$\lambda = -\sum_{l=1}^{k} c_l \tag{133}$$

Combining equations (129) and (133),

$$c_i - \sum_{l=1}^{k} c_l \overline{P(\omega_i)} = 0 \tag{134}$$

$$\overline{P(\omega_i)} = \frac{c_i}{\sum_{l=1}^{k} c_l} \tag{135}$$

Inserting equation (120) into (135), and switching the order of summation in the denominator,

$$\overline{P(\omega_i)} = \frac{\sum_{j=1}^{n} \frac{p(\mathbf{x}_j|\phi_i)P(\omega_i)}{p(\mathbf{x}_j|\Theta)}}{\sum_{j=1}^{n}\sum_{l=1}^{k} \frac{p(\mathbf{x}_j|\phi_l)P(\omega_l)}{p(\mathbf{x}_j|\Theta)}} \tag{136}$$

$$\overline{P(\omega_i)} = \frac{\sum_{j=1}^{n} \frac{p(\mathbf{x}_j|\phi_i)P(\omega_i)}{p(\mathbf{x}_j|\Theta)}}{\sum_{j=1}^{n} \frac{\sum_{l=1}^{k} p(\mathbf{x}_j|\phi_l)P(\omega_l)}{p(\mathbf{x}_j|\Theta)}} \tag{137}$$

Note that,

$$\frac{\sum_{l=1}^{k} p(\mathbf{x}_j|\phi_l)P(\omega_l)}{p(\mathbf{x}_j|\Theta)} = \frac{p(\mathbf{x}_j|\Theta)}{p(\mathbf{x}_j|\Theta)} = 1 \tag{138}$$

Therefore, equation (137) simplifies to,

$$\overline{P(\omega_i)} = \frac{1}{n}\sum_{j=1}^{n} \frac{p(\mathbf{x}_j|\phi_i)P(\omega_i)}{p(\mathbf{x}_j|\Theta)} = \frac{1}{n}\sum_{j=1}^{n} P(\omega_i|\mathbf{x}_j, \Theta)\,,\ i \in \{1, 2, \ldots, k\} \tag{139}$$

Thus, equation (139) gives us the EM iterative update rule for estimating the priors $P(\omega_i)$. The maximization of the second term in equation (119) proceeds by setting the gradient with respect to $\bar{\phi}_i$ of,

$$\sum_{l=1}^{k}\sum_{j=1}^{n} \frac{p(\mathbf{x}_j|\phi_l)P(\omega_l)}{p(\mathbf{x}_j|\Theta)} \ln p(\mathbf{x}_j|\bar{\phi}_l) \tag{140}$$

equal to zero,

$$\nabla_{\bar{\phi}_i} \sum_{l=1}^{k}\sum_{j=1}^{n} \frac{p(\mathbf{x}_j|\phi_l)P(\omega_l)}{p(\mathbf{x}_j|\Theta)} \ln p(\mathbf{x}_j|\bar{\phi}_l) = \sum_{j=1}^{n} \frac{p(\mathbf{x}_j|\phi_i)P(\omega_i)}{p(\mathbf{x}_j|\Theta)} \frac{\nabla_{\bar{\phi}_i} p(\mathbf{x}_j|\bar{\phi}_i)}{p(\mathbf{x}_j|\bar{\phi}_i)} = 0 \tag{141}$$

and solving for $\bar{\phi}_i$. In order to compute the gradients,

$$\nabla_{\bar{\phi}_i} p(\mathbf{x}_j|\bar{\phi}_i) \tag{142}$$

for the Gaussian distribution,

$$p(\mathbf{x}_j|\bar{\phi}_i) = p(\mathbf{x}_j|\bar{\mu}_i, \bar{\Sigma}_i) = \frac{1}{(2\pi)^{d/2}|\bar{\Sigma}_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}_j - \bar{\mu}_i)^T \bar{\Sigma}_i^{-1} (\mathbf{x}_j - \bar{\mu}_i)\right] \equiv N[\mathbf{x}_j, \bar{\mu}_i, \bar{\Sigma}_i] \tag{143}$$

we will need to compute,

$$\nabla_{\bar{\mu}_i} p(\mathbf{x}_j|\bar{\phi}_i) \text{ and } \nabla_{\bar{\Sigma}_i} p(\mathbf{x}_j|\bar{\phi}_i) \tag{144}$$

separately. In order to undertake this, we will make use of the following results from linear algebra:

$$\nabla_{\mathbf{b}}(\mathbf{b}^T\mathbf{A}\mathbf{b}) = \mathbf{A}\mathbf{b} + \mathbf{A}^T\mathbf{b} \tag{145}$$

$$\nabla_{\mathbf{A}}(\mathbf{b}^T\mathbf{A}\mathbf{b}) = \mathbf{b}\mathbf{b}^T \tag{146}$$

$$\nabla_{\mathbf{A}}|\mathbf{A}| = (\mathbf{A}^T)^{-1}|\mathbf{A}| \tag{147}$$

$$|\mathbf{A}^{-1}| = |\mathbf{A}|^{-1} \tag{148}$$

where $\mathbf{b}$ is a d-dimensional vector, and $\mathbf{A}$ is a $d \times d$ matrix. Note that if $\mathbf{A}$ is symmetric, then equation (145) reduces to,

$$\nabla_{\mathbf{b}}(\mathbf{b}^T\mathbf{A}\mathbf{b}) = 2\mathbf{A}\mathbf{b} \tag{149}$$

and equation (147) reduces to,

$$\nabla_{\mathbf{A}}|\mathbf{A}| = \mathbf{A}^{-1}|\mathbf{A}| . \tag{150}$$

Given these identities from linear algebra, we can now compute the gradients in (144). First, for the mean vector $\bar{\mu}_i$,

$$\nabla_{\bar{\mu}_i} p(\mathbf{x}_j|\bar{\phi}_i) = \nabla_{\bar{\mu}_i} N[\mathbf{x}_j, \bar{\mu}_i, \bar{\Sigma}_i] \tag{151}$$

From equation (149),

$$\nabla_{\mu} N[\mathbf{x}, \mu, \Sigma] = N[\mathbf{x}, \mu, \Sigma]\nabla_{\mu}\left[-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right] \tag{152}$$

(Here we drop the indices i, j and the over bar for notational simplicity.)

$$\nabla_{\mu} N[\mathbf{x}, \mu, \Sigma] = N[\mathbf{x}, \mu, \Sigma]\Sigma^{-1}(\mathbf{x}-\mu) \tag{153}$$

Reinserting the problem-specific notation,

$$\nabla_{\bar{\mu}_i} N[\mathbf{x}_j, \bar{\mu}_i, \bar{\Sigma}_i] = N[\mathbf{x}_j, \bar{\mu}_i, \bar{\Sigma}_i]\bar{\Sigma}_i^{-1}(\mathbf{x}_j - \bar{\mu}_i) \tag{154}$$

We can now solve for $\bar{\mu}_i$. Combining equation (141), (143) and (154),

$$\sum_{j=1}^{n} \frac{p(\mathbf{x}_j|\phi_i)P(\omega_i)}{p(\mathbf{x}_j|\Theta)} \frac{\nabla_{\bar{\mu}_i} p(\mathbf{x}_j|\bar{\phi}_i)}{p(\mathbf{x}_j|\bar{\phi}_i)} = 0 \tag{155}$$

$$\sum_{j=1}^{n} \frac{p(\mathbf{x}_j|\phi_i)P(\omega_i)}{p(\mathbf{x}_j|\Theta)}[\bar{\Sigma}_i^{-1}(\mathbf{x}_j - \bar{\mu}_i)] = 0 \tag{156}$$

$$\bar{\mu}_i = \frac{\displaystyle\sum_{j=1}^{n} \frac{p(\mathbf{x}_j|\phi_i)P(\omega_i)}{p(\mathbf{x}_j|\Theta)}\mathbf{x}_j}{\displaystyle\sum_{j=1}^{n} \frac{p(\mathbf{x}_j|\phi_i)P(\omega_i)}{p(\mathbf{x}_j|\Theta)}} = \frac{\displaystyle\sum_{j=1}^{n} P(\omega_i|\mathbf{x}_j, \Theta)\mathbf{x}_j}{\displaystyle\sum_{j=1}^{n} P(\omega_i|\mathbf{x}_j, \Theta)}, \; i \in \{1, 2, \ldots, k\} . \tag{157}$$

For the covariance matrices $\bar{\Sigma}_i$, rather than derive,

$$\nabla_{\bar{\Sigma}_i} p(\mathbf{x}_j|\bar{\phi}_i) \tag{158}$$

we will compute,

$$\nabla_{\bar{\Sigma}_i^{-1}} p(\mathbf{x}_j|\bar{\phi}_i) = \nabla_{\bar{\Sigma}_i^{-1}} N[\mathbf{x}_j, \bar{\mu}_i, \bar{\Sigma}_i] \tag{159}$$

instead, because it will be a little easier. (During the derivation, we will again drop the problem specific notation.) Using the product and chain rules for differentiation,

$$\begin{aligned}\nabla_{\Sigma^{-1}} N[\mathbf{x}, \mu, \Sigma] = {} & N[\mathbf{x}, \mu, \Sigma]\nabla_{\Sigma^{-1}}\left[-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right] + \\ & \frac{1}{(2\pi)^{d/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right]\nabla_{\Sigma^{-1}}[|\Sigma|^{-1/2}]\end{aligned} \tag{160}$$

Note that we can write,

$$\frac{1}{(2\pi)^{d/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right] = N[\mathbf{x}, \mu, \Sigma]|\Sigma|^{1/2} \tag{161}$$

so that equation (160) reduces to,

$$\nabla_{\Sigma^{-1}} N[\mathbf{x}, \mu, \Sigma] = N[\mathbf{x}, \mu, \Sigma]\nabla_{\Sigma^{-1}}\left[-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right] + N[\mathbf{x}, \mu, \Sigma]|\Sigma|^{1/2}\nabla_{\Sigma^{-1}}[|\Sigma|^{-1/2}] \tag{162}$$

where,

$$\nabla_{\Sigma^{-1}}\left[-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right] = -\frac{1}{2}(\mathbf{x}-\mu)^T(\mathbf{x}-\mu) \text{ [see equation (146)]} \tag{163}$$

and [see equations (148) and (150)],

$$\begin{aligned}\nabla_{\Sigma^{-1}}[|\Sigma|^{-1/2}] &= \nabla_{\Sigma^{-1}}[|\Sigma^{-1}|^{1/2}] \\ &= \frac{1}{2}|\Sigma^{-1}|^{-1/2}\nabla_{\Sigma^{-1}}|\Sigma^{-1}| \\ &= \frac{1}{2}|\Sigma^{-1}|^{-1/2}\Sigma|\Sigma^{-1}| \\ &= \frac{1}{2}|\Sigma|^{1/2}\Sigma|\Sigma|^{-1} \\ &= \frac{1}{2}|\Sigma|^{-1/2}\Sigma\end{aligned} \tag{164}$$

Combining equations (162), (163) and (164),

$$\nabla_{\Sigma^{-1}} N[\mathbf{x}, \mu, \Sigma] = -\frac{1}{2}N[\mathbf{x}, \mu, \Sigma](\mathbf{x}-\mu)(\mathbf{x}-\mu)^T + \frac{1}{2}N[\mathbf{x}, \mu, \Sigma]\Sigma \tag{165}$$

$$\nabla_{\Sigma^{-1}} N[\mathbf{x}, \mu, \Sigma] = \frac{1}{2}N[\mathbf{x}, \mu, \Sigma]\{\Sigma - (\mathbf{x}-\mu)(\mathbf{x}-\mu)^T\} \tag{166}$$

Reinserting the problem-specific notation,

$$\nabla_{\bar{\Sigma}_i^{-1}} N[\mathbf{x}_j, \bar{\mu}_i, \bar{\Sigma}_i] = \frac{1}{2}N[\mathbf{x}_j, \bar{\mu}_i, \bar{\Sigma}_i]\{\bar{\Sigma}_i - (\mathbf{x}_j-\bar{\mu}_i)(\mathbf{x}_j-\bar{\mu}_i)^T\} \tag{167}$$

we can now solve for $\bar{\Sigma}_i$. Combining equation (141), (143) and (167),

$$\sum_{j=1}^{n} \frac{p(\mathbf{x}_j|\phi_i)P(\omega_i)}{p(\mathbf{x}_j|\Theta)} \frac{\nabla_{\bar{\Sigma}_i^{-1}} p(\mathbf{x}_j|\bar{\phi}_i)}{p(\mathbf{x}_j|\bar{\phi}_i)} = 0 \tag{168}$$

$$\frac{1}{2}\sum_{j=1}^{n} \frac{p(\mathbf{x}_j|\phi_i)P(\omega_i)}{p(\mathbf{x}_j|\Theta)}[\bar{\Sigma}_i - (\mathbf{x}_j - \bar{\mu}_i)(\mathbf{x}_j - \bar{\mu}_i)^T] = 0 \tag{169}$$

$$\bar{\Sigma}_i = \frac{\displaystyle\sum_{j=1}^{n} \frac{p(\mathbf{x}_j|\phi_i)P(\omega_i)}{p(\mathbf{x}_j|\Theta)}(\mathbf{x}_j - \bar{\mu}_i)(\mathbf{x}_j - \bar{\mu}_i)^T}{\displaystyle\sum_{j=1}^{n} \frac{p(\mathbf{x}_j|\phi_i)P(\omega_i)}{p(\mathbf{x}_j|\Theta)}} = \frac{\displaystyle\sum_{j=1}^{n} P(\omega_i|\mathbf{x}_j, \Theta)(\mathbf{x}_j - \bar{\mu}_i)(\mathbf{x}_j - \bar{\mu}_i)^T}{\displaystyle\sum_{j=1}^{n} P(\omega_i|\mathbf{x}_j, \Theta)} \tag{170}$$

Therefore, the EM iterative update equations for the mixture-of-Gaussians problem are given by equations (139), (154) and (170):

$$\overline{P(\omega_i)} = \frac{1}{n}\sum_{j=1}^{n} P(\omega_i|\mathbf{x}_j, \Theta)\,,\ i \in \{1, 2, \ldots, k\} \tag{171}$$

$$\bar{\mu}_i = \frac{\displaystyle\sum_{j=1}^{n} P(\omega_i|\mathbf{x}_j, \Theta)\mathbf{x}_j}{\displaystyle\sum_{j=1}^{n} P(\omega_i|\mathbf{x}_j, \Theta)}\,,\ i \in \{1, 2, \ldots, k\} \tag{172}$$

$$\bar{\Sigma}_i = \frac{\displaystyle\sum_{j=1}^{n} P(\omega_i|\mathbf{x}_j, \Theta)(\mathbf{x}_j - \bar{\mu}_i)(\mathbf{x}_j - \bar{\mu}_i)^T}{\displaystyle\sum_{j=1}^{n} P(\omega_i|\mathbf{x}_j, \Theta)}\,,\ i \in \{1, 2, \ldots, k\} \tag{173}$$

Note that these equations are identical to the ones previously derived using the intuitive formulation of the EM algorithm [equations (49) through (51)]. Second, note that what ultimately made the derivation of equations (171) through (173) possible was the exponential nature of the component densities $p(\mathbf{x}_j|\phi_i)$, which led to the cancellation of numerator and denominator in both equations (155) and (168).

6. Examples of mixture modeling

To see mixture-modeling examples for both synthetic and real data, check out the *Mathematica* notebooks, and quicktime animations on the course web page: http://mil.ufl.edu/~nechyba/eel6825.