# EXPECTATION MAXIMIZATION: FROM THE HORSE'S MOUTH

MAKSIM LEVENTAL

Expectation[1] maximization (EM) is a way to iteratively approximate the maximum likelihood estimators (MLEs) for a parametric family when solving the MLE equations analytically is intractable. Recall that finding the MLEs for a parametric family

$$\mathbf{Y} \sim f_{\mathbf{Y}}(\cdot;\boldsymbol{\theta})\ \boldsymbol{\theta} \in \Theta$$

is tantamount to maximizing the likelihood function

$$L(\boldsymbol{\theta};\mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})$$

as a function of $\boldsymbol{\theta}$. The maximum likelihood estimators are estimators of/for $\boldsymbol{\theta}$, usually denoted $\hat{\boldsymbol{\theta}}$. The actual maximization is effected by finding critical points of $L$ with respect to $\boldsymbol{\theta}$ and testing concavity, i.e. solving

$$\nabla L(\boldsymbol{\theta};\mathbf{y}) = 0$$

and checking the negative definiteness of Hessian of $L$. In general $\partial_{\theta_i} L(\boldsymbol{\theta};\mathbf{y})$ might be highly non-linear in $\theta_i$ and hence finding each might be very difficult. What to do?

## 1. Expectation Maximization Beta

Suppose[2] there exists another random, unobserved, vector $\mathbf{X}$ such that were $\mathbf{X}$ in fact observed, maximizing the joint likelihood for $\mathbf{X}$ and $\mathbf{Y}$

$$L(\boldsymbol{\theta};\mathbf{x},\mathbf{y}) = f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y};\boldsymbol{\theta})$$

is easier than the original problem. $\mathbf{X}$ is called a latent or hidden variable. Now a priori there's absolutely no reason to assume that having more information in hand and complicating the model should make things easier. On the contrary, for example, a good physical model makes simplifying assumptions and thereby becomes tractable. But indeed for some very useful models such as Hidden Markov Models and Gaussian Mixture Models this ansatz does simplify computing the MLEs.

All is going swimmingly except $\mathbf{X}$ is unobserved - the only observed data are $\mathbf{Y}$. What to do? Estimate $\mathbf{X}$ of course. How? Using the best estimator $E[\mathbf{X}|\mathbf{y};\boldsymbol{\theta}]$ of $\mathbf{X}$, based on observed $\mathbf{y}$, that minimizes the risk for the quadratic loss function[3], i.e. minimizes the mean square error. Note two things. Firstly, it's important that you can actually compute *this* in closed form, otherwise you've traded one intractable problem for another. Secondly, since $E[\mathbf{X}|\mathbf{y};\boldsymbol{\theta}]$ is implicitly a function of $\boldsymbol{\theta}$, which is unknown to begin with[4] you need some initial guess for it too, otherwise you can't

[1] These notes are basically a transcription of notes (taken dutifully by Chris Gianelli) from lectures given by Dr. Kshitij Khare at UF in Spring 15. He's not a horse - it's just an expression.

[2] Suppose!

[3] $l\left(\mathbf{X},\hat{\mathbf{X}}\right) = \left(\mathbf{X}-\hat{\mathbf{X}}\right)^2$

[4] Don't forget the point of all this is actually to estimate $\boldsymbol{\theta}$.

compute the expectation. Hence the expectation computed is actually $E\left[\mathbf{X}|\mathbf{y};\boldsymbol{\theta}^{(r)}\right]$ where $\boldsymbol{\theta}^{(r)}$ is the current guess for $\boldsymbol{\theta}$. Then once you have this estimate for $\mathbf{X}$ just maximize

$$L\left(\boldsymbol{\theta};E\left[\mathbf{X}|\mathbf{y};\boldsymbol{\theta}^{(r)}\right],\mathbf{y}\right)=f_{\mathbf{X},\mathbf{Y}}\left(E\left[\mathbf{X}|\mathbf{y};\boldsymbol{\theta}^{(r)}\right],\mathbf{y};\boldsymbol{\theta}\right)$$

The procedure alternates between estimating $\mathbf{x}$ using $E\left[\mathbf{X}|\mathbf{y};\boldsymbol{\theta}^{(r)}\right]$ and maximizing $L\left(\boldsymbol{\theta};E\left[\mathbf{X}|\mathbf{y};\boldsymbol{\theta}^{(r)}\right],\mathbf{y}\right)$. Compute expectation, then perform maxization, compute expectation, then perform maximization, compute expectation,... hence **Expectation Maximization** algorithm. Just to be clear

**Definition 1. Expectation Algorithm Beta.** Given some observed data $\mathbf{y}$, in order to perform the intractable maximization of $L(\boldsymbol{\theta};\mathbf{y})$, posit the existence of some latent $\mathbf{x}$ such that $L(\boldsymbol{\theta};\mathbf{x},\mathbf{y})$ is easier. Set $\boldsymbol{\theta}^{(0)}$ to be some initial guess for $\boldsymbol{\theta}$ then

(1) E-step: Set $\mathbf{x}^{(r)}=E\left[\mathbf{X}|\mathbf{y};\boldsymbol{\theta}^{(r)}\right]$
(2) M-step: Set $\boldsymbol{\theta}^{(r+1)}=\arg\max_{\boldsymbol{\theta}}L\left(\boldsymbol{\theta};\mathbf{x}^{(r)},\mathbf{y}\right)$
(3) Go to step 1 unless $\left|\boldsymbol{\theta}^{(r+1)}-\boldsymbol{\theta}^{(r)}\right|<\varepsilon$ for some $\varepsilon$ of your choosing.

The equations in step 1 and 2 are called update equations because specify how to update the current estimate for $\mathbf{x}^{(i)}$ and $\boldsymbol{\theta}^{(i)}$.

**Example 2.** Here's an example: let $Y_i$ be iid such that

$$f_{Y_i}(y_i;\theta)=\theta g_1(y_i)+(1-\theta)g_0(y_i)$$

where $\theta\in(0,1)$ and $g_0$ and $g_1$ are Gaussians with known means $(\mu_0,\mu_1)$and variances $\left(\sigma_0^2,\sigma_1^2\right)$. We want the MLE for $\theta$. Note that

$$L(\theta;\mathbf{y})=\prod_{i=1}^{n}\left(\theta g_1(y_i)+(1-\theta)g_0(y_i)\right)$$

Quite messy[5]. What to do? EM to the rescue! Suppose the $Y_i$ are drawn by a process where a $\theta$-biased coin is flipped and either $g_0$ or $g_1$ generates the $y_i$ depending on whether the coin lands heads up or down. The latent variable here then is which Gaussian generated $y_i$. Hence let $X_i$ be Bernoulli random variables where

$$\begin{aligned}P(X_i=x_i) &= \begin{cases}\theta & \text{if } x_i=1\\ 1-\theta & \text{if } x_i=0\end{cases}\\ &= \theta^{x_i}(1-\theta)^{1-x_i}\end{aligned}$$

and $f_{Y_i|X_i}(y_i|x_i=1;\theta)=g_1(y_i)$ and $f_{Y_i|X_i}(y_i|x_i=0;\theta)=g_0(y_i)$. Then

$$f_{X_i,Y_i}(x_i,y_i;\theta)=f_{Y_i|X_i}(y_i|x_i;\theta)P(X_i=x_i)$$

[5]Even if you think you're clever and try to maximize log-likelihood you're still going to have a rough go at it.

and so

$$\begin{aligned} L(\theta;\mathbf{x},\mathbf{y}) &= \prod_{i=1}^{n} f_{Y_i|X_i}(y_i|x_i;\theta)\, P(X_i = x_i) \\ &= \prod_{i=1}^{n} (g_1(y_i))^{x_i} (g_0(y_i))^{1-x_i} \theta^{x_i} (1-\theta)^{1-x_i} \\ &= \theta^{\sum_i x_i} (1-\theta)^{n-\sum_i x_i} \prod_{i=1}^{n} (g_1(y_i))^{x_i} (g_0(y_i))^{1-x_i} \end{aligned}$$

Note the trick in writing $f_{Y_i|X_i}(y_i|x_i;\theta) = (g_1)^{x_i}(g_0)^{1-x_i}$ - it comes up a lot for a class of models called mixture models. For the E-step, to compute $E[\mathbf{X}|\mathbf{y};\theta']$ we use i.i.d-ness and compute $E[X_i|y_i;\theta']$ instead, which will hold for each $i$. To compute the conditional expectation $E[X_i|y_i;\theta]$ we need the conditional distribution $f_{X_i|Y_i}(x_i|y_i)$. By Bayes' Theorem

$$\begin{aligned} f_{X_i|Y_i}(x_i|y_i;\theta) &= \frac{f_{X_i,Y_i}(x_i,y_i;\theta)}{f_{Y_i}(y_i;\theta)} \\ &= \frac{f_{X_i,Y_i}(x_i,y_i;\theta)}{\sum_{x_i} f_{X_i,Y_i}(x_i,y_i;\theta)} \\ &= \frac{(g_1(y_i))^{x_i}(g_0(y_i))^{1-x_i}\theta^{x_i}(1-\theta)^{1-x_i}}{\sum_{x_i}(g_1(y_i))^{x_i}(g_0(y_i))^{1-x_i}\theta^{x_i}(1-\theta)^{1-x_i}} \\ &= \frac{(g_1(y_i))^{x_i}(g_0(y_i))^{1-x_i}\theta^{x_i}(1-\theta)^{1-x_i}}{g_1(y_i)\theta + g_0(y_i)(1-\theta)} \\ &= \left(\frac{g_1(y_i)\theta}{g_1(y_i)\theta + g_0(y_i)(1-\theta)}\right)^{x_i} \left(\frac{g_0(y_i)(1-\theta)}{g_1(y_i)\theta + g_0(y_i)(1-\theta)}\right)^{1-x_i} \end{aligned}$$

So $X_i|Y_i$ is still Bernoulli just renormalized. Hence

$$E[X_i|y_i;\theta] = \frac{g_1(y_i)\theta}{g_1(y_i)\theta + g_0(y_i)(1-\theta)}$$

For the M-step, since

$$L(\theta;\mathbf{x},\mathbf{y}) = \left[\theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}\right]\left[\prod_{i=1}^{n}(g_1(y_i))^{x_i}(g_0(y_i))^{1-x_i}\right]$$

and the second term is independent of $\theta$ we can just maximize the first term. But this is just the the joint distribution for $n$ i.i.d Bernoulli random variables and the MLE $\hat{\theta}$ is

$$\hat{\theta} = \frac{\sum_i x_i}{n}$$

Therefore, finally, the update equations are

$$\begin{aligned} x_i^{(r)} &= \frac{g_1(y_i)\theta^{(r)}}{g_1(y_i)\theta^{(r)} + g_0(y_i)\left(1-\theta^{(r)}\right)} \\ \theta^{(r+1)} &= \frac{1}{n}\sum_{i=1}^{n} x_i^{(r)} \end{aligned}$$

■

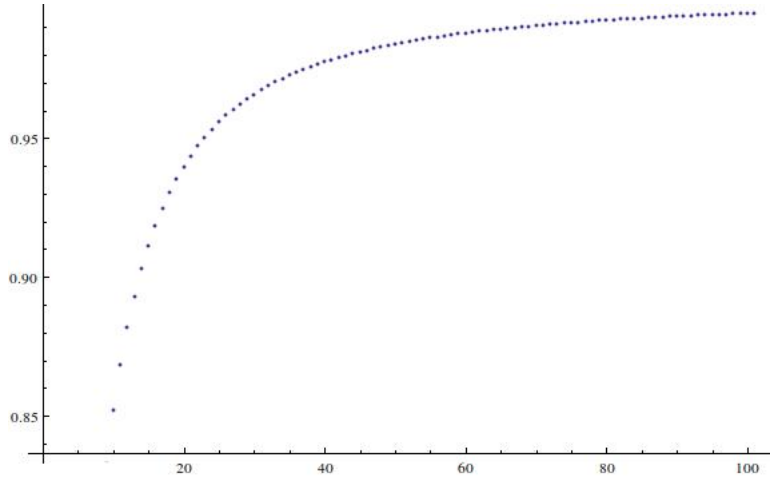


FIGURE 1.1. EM Beta

So why did I dub it EM **Beta** instead of just EM? Because this isn't the standard EM algorithm. But why alter this algorithm at all? What's wrong with it as is? Well there are no convergence guarantees. Indeed the iterates in Example 1 don't converge to the right answer: figure **??** shows the first 100 iterates for $g_1 \sim \mathrm{n}(1,2), g_0 \sim \mathrm{n}(3,4), \theta = 2/3$. So they converge but to $1 \neq 2/3$. Why did I present this algorithm first? Standard EM is slightly more complicated and much less intuitive but legend has it that it was in fact conceived in this way first and then manipulated to get convergence.

## 2. Expectation Maximization for Real

Recall that the whole point of this procedure is to actually maximize the likelihood for $\mathbf{Y}$. This is equivalent to maximizing the log-likelihood for $\mathbf{Y}$

$$l(\boldsymbol{\theta};\mathbf{y}) = \log L(\boldsymbol{\theta};\mathbf{y}) = \log f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})$$

The ansatz here is the same: suppose there exists another random, unobserved, vector $\mathbf{X}$ such that were $\mathbf{X}$ in fact observed, maximizing the joint log-likelihood for $\mathbf{X}$ and $\mathbf{Y}$

$$l(\boldsymbol{\theta};\mathbf{x},\mathbf{y}) = \log L(\boldsymbol{\theta};\mathbf{x},\mathbf{y}) = \log f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y};\boldsymbol{\theta})$$

is easier than the original problem. The ease of maximizing $l(\boldsymbol{\theta};\mathbf{x},\mathbf{y})$ over $l(\boldsymbol{\theta};\mathbf{y})$ isn't immediately apparent but first using Bayes' theorem we see that

$$l(\boldsymbol{\theta};\mathbf{y}) = \log f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta}) \tag{2.1}$$

$$= \log\left(\frac{f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y};\boldsymbol{\theta})}{f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y};\boldsymbol{\theta})}\right) \tag{2.2}$$

$$= \log\left(f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y};\boldsymbol{\theta})\right) - \log\left(f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y};\boldsymbol{\theta})\right) \tag{2.3}$$

If we can easily maximize the first term in **??**, with respect to $\boldsymbol{\theta}$ and the second term in **??** doesn't spoil things somehow, then we'll indeed maximize $l(\boldsymbol{\theta};\mathbf{y})$. This seems like a workflow different from that of the beta algorithm[6] but the analogy follows. Since $\mathbf{X}$ is unobserved we face the same

[6]Here we're explicitly maximizing a function of the likelihood for $\mathbf{Y}$ by maximizing a lowerbound.

difficulty as in the beta algorithm: the terms in **??** can't be computed and must be estimated. We use the same estimator as before $E\left[\cdot|\mathbf{y};\boldsymbol{\theta}^{(r)}\right]$

$$l(\boldsymbol{\theta};\mathbf{y}) = E\left[\log\left(f_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{y};\boldsymbol{\theta})\right)|\mathbf{y};\boldsymbol{\theta}^{(r)}\right] - E\left[\log\left(f_{\mathbf{X}|\mathbf{Y}}(\mathbf{X}|\mathbf{y};\boldsymbol{\theta})\right)|\mathbf{y};\boldsymbol{\theta}^{(r)}\right] \tag{2.4}$$

Note some subtle things:

- $\mathbf{x} \to \mathbf{X}$ because each of the terms on the right-hand side of **??** are estimates of the random variable $\mathbf{X}$ as a function of observed data $\mathbf{y}$.
- The expectations are again computed with respect to the conditional distribution of $\mathbf{X}$ given $\mathbf{Y}$ for fixed (iterate) values of $\boldsymbol{\theta}^{(r)}$, i.e. $f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}^{(r)}\right)$.
- The estimators, for which the expectations are computed, are functions of $\boldsymbol{\theta}$.
- Equality is maintained because since $l(\boldsymbol{\theta};\mathbf{y})$ is independent of $\mathbf{X}$

$$E\left[l(\boldsymbol{\theta};\mathbf{y})|\mathbf{y};\boldsymbol{\theta}^{(r)}\right] = l(\boldsymbol{\theta};\mathbf{y})$$

The algorithm then is

**Definition 3. Expectation Algorithm.** Given some observed data $\mathbf{y}$, in order to perform the intractable maximization of $L(\boldsymbol{\theta};\mathbf{y})$, posit the existence of some latent $\mathbf{X}$ such that $L(\boldsymbol{\theta};\mathbf{x},\mathbf{y})$ is easier. Set $\boldsymbol{\theta}^{(0)}$ to be some initial guess for $\boldsymbol{\theta}$ then

(1) E-step: Compute $E\left[\log\left(f_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{y};\boldsymbol{\theta})\right)|\mathbf{y};\boldsymbol{\theta}^{(r)}\right]$
(2) M-step: Set $\boldsymbol{\theta}^{(r+1)} = \arg\max_{\boldsymbol{\theta}} E\left[\log\left(f_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{y};\boldsymbol{\theta})\right)|\mathbf{y};\boldsymbol{\theta}^{(r)}\right]$
(3) Go to step 1 unless $\left|\boldsymbol{\theta}^{(r+1)} - \boldsymbol{\theta}^{(r)}\right| < \varepsilon$ for some $\varepsilon$ of your choosing.

The only thing to remains is to prove that maximizing $E\left[\log\left(f_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{y};\boldsymbol{\theta})\right)|\mathbf{y};\boldsymbol{\theta}^{(r)}\right]$ is sufficient to maximize $l(\boldsymbol{\theta};\mathbf{y})$. I won't but I'll prove a thing on the way to that result, namely that $l\left(\boldsymbol{\theta}^{(r)};\mathbf{y}\right) \geq l\left(\boldsymbol{\theta}^{(r-1)};\mathbf{y}\right)$ and since $l(\boldsymbol{\theta};\mathbf{y})$ is bounded above (it's a density) the sequence of $l\left(\boldsymbol{\theta}^{(r)};\mathbf{y}\right)$ must converge. The last hurdle would be proving that convergence of $l\left(\boldsymbol{\theta}^{(r)};\mathbf{y}\right)$ implies the convergence of the iterates $\boldsymbol{\theta}^{(r)}$ themselves[**?**].

**Theorem 4.** *(Monotonic EM Sequence) The sequence* $\left\{\boldsymbol{\theta}^{(r)}\right\}$ *satisfies* $l\left(\boldsymbol{\theta}^{(r+1)};\mathbf{y}\right) \geq l\left(\boldsymbol{\theta}^{(r)};\mathbf{y}\right)$.

*Proof.* Start with

$$l(\boldsymbol{\theta};\mathbf{y}) = \log\left(f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y};\boldsymbol{\theta})\right) - \log\left(f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y};\boldsymbol{\theta})\right)$$

Take conditional expectation with respect to $\mathbf{X}|\mathbf{Y};\boldsymbol{\theta}^{(r)}$ of both sides

$$\begin{aligned}\int l(\boldsymbol{\theta};\mathbf{y}) f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}^{(r)}\right) d\mathbf{x} &= \int \log\left(f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y};\boldsymbol{\theta})\right) f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}^{(r)}\right) d\mathbf{x} \\ &\quad - \int \log\left(f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y};\boldsymbol{\theta})\right) f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}^{(r)}\right) d\mathbf{x}\end{aligned}$$

Since $l(\boldsymbol{\theta};\mathbf{y})$ is independent of $\mathbf{x}$

$$\begin{aligned}\int l(\boldsymbol{\theta};\mathbf{y}) f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}^{(r)}\right) d\mathbf{x} &= l(\boldsymbol{\theta};\mathbf{y}) \int f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}^{(r)}\right) d\mathbf{x} \\ &= l(\boldsymbol{\theta};\mathbf{y})\end{aligned}$$

and so

$$\begin{aligned} l(\boldsymbol{\theta};\mathbf{y}) &= \int \log\left(f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y};\boldsymbol{\theta})\right) f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}^{(r)}\right) d\mathbf{x} \\ &\quad - \int \log\left(f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y};\boldsymbol{\theta})\right) f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}^{(r)}\right) d\mathbf{x} \\ &= E\left[\log\left(f_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{y};\boldsymbol{\theta})\right)|\mathbf{y};\boldsymbol{\theta}^{(r)}\right] - E\left[\log\left(f_{\mathbf{X}|\mathbf{Y}}(\mathbf{X}|\mathbf{y};\boldsymbol{\theta})\right)|\mathbf{y};\boldsymbol{\theta}^{(r)}\right] \end{aligned}$$

Let

$$Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(r)}\right) = E\left[\log\left(f_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{y};\boldsymbol{\theta})\right)|\mathbf{y};\boldsymbol{\theta}^{(r)}\right]$$

and $K\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(r)}\right) = E\left[\log\left(f_{\mathbf{X}|\mathbf{Y}}(\mathbf{X}|\mathbf{y};\boldsymbol{\theta})\right)|\mathbf{y};\boldsymbol{\theta}^{(r)}\right]$. Then

$$l(\boldsymbol{\theta};\mathbf{y}) = Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(r)}\right) - K\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(r)}\right)$$

If we can show that

$$\begin{aligned} Q\left(\boldsymbol{\theta}^{(r)},\boldsymbol{\theta}^{(r)}\right) &\leq Q\left(\boldsymbol{\theta}^{(r+1)},\boldsymbol{\theta}^{(r)}\right) \\ &\text{and} \\ K\left(\boldsymbol{\theta}^{(r)},\boldsymbol{\theta}^{(r)}\right) &\geq K\left(\boldsymbol{\theta}^{(r+1)},\boldsymbol{\theta}^{(r)}\right) \end{aligned}$$

then it will follow that $l\left(\boldsymbol{\theta}^{(r+1)};\mathbf{y}\right) \geq l\left(\boldsymbol{\theta}^{(r)};\mathbf{y}\right)$. Well by definition of

$$\boldsymbol{\theta}^{(r+1)} = \arg\max_{\boldsymbol{\theta}} E\left[\log\left(f_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{y};\boldsymbol{\theta})\right)|\mathbf{y};\boldsymbol{\theta}^{(r)}\right] = \arg\max_{\boldsymbol{\theta}} Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(r)}\right)$$

so

$$Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(r)}\right) \leq Q\left(\boldsymbol{\theta},\arg\max_{\boldsymbol{\theta}} Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(r)}\right)\right) = Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(r+1)}\right)$$

To see that $K\left(\boldsymbol{\theta}^{(r)},\boldsymbol{\theta}^{(r)}\right) \geq K\left(\boldsymbol{\theta}^{(r+1)},\boldsymbol{\theta}^{(r)}\right)$ inspect

$$\int \log\left(\frac{f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}^{(r+1)}\right)}{f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}^{(r)}\right)}\right) f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}^{(r)}\right) d\mathbf{x}$$

On the one hand

$$\begin{aligned} \int \log\left(\frac{f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}^{(r+1)}\right)}{f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}^{(r)}\right)}\right) f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}^{(r)}\right) d\mathbf{x} &= \int \log\left(f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}^{(r+1)}\right)\right) f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}^{(r)}\right) d\mathbf{x} \\ &\quad - \int \log\left(f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}^{(r)}\right)\right) f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}^{(r)}\right) d\mathbf{x} \\ &= K\left(\boldsymbol{\theta}^{(r+1)},\boldsymbol{\theta}^{(r)}\right) - K\left(\boldsymbol{\theta}^{(r)},\boldsymbol{\theta}^{(r)}\right) \end{aligned}$$

on the other hand, by Jensen's[7] inequality[8]

$$\begin{aligned}\int \log\left(\frac{f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}^{(r+1)}\right)}{f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}^{(r)}\right)}\right) f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}^{(r)}\right) d\mathbf{x} &\leq \log\left[\int\left(\frac{f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}^{(r+1)}\right)}{f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}^{(r)}\right)}\right) f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}^{(r)}\right) d\mathbf{x}\right] \\ &= \log\left[\int f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}^{(r+1)}\right) d\mathbf{x}\right] \\ &= 0\end{aligned}$$

Hence

$$K\left(\boldsymbol{\theta}^{(r+1)},\boldsymbol{\theta}^{(r)}\right) - K\left(\boldsymbol{\theta}^{(r)},\boldsymbol{\theta}^{(r)}\right) \leq 0$$

which completes the proof. □

## 3. Applications

3.1. **Gaussian Mixture Models.** EM works really well for mixture models, e.g. $\mathbf{Y}_i$ is distributed iid such that

$$f_{\mathbf{Y}_1}(\mathbf{y}_1) = \sum_{j=1}^{m} \alpha_j g_j(\mathbf{y}_1;\boldsymbol{\theta}_j)$$

where $g_j$ are (in general multivariate) densities with parameter vectors $\boldsymbol{\theta}_j$ (in general, distinct for distinct $j$) and $\sum_j \alpha_j = 1$. If $g_j \sim \mathrm{n}\left(\mu_j,\sigma_j^2\right)$ then the mixture model is called a Gaussian mixture model (GMM). Naively if you wanted to find the MLEs you would maximize

$$L\left((\mathbf{y}_1,\ldots,\mathbf{y}_n);\boldsymbol{\theta}\right) = \prod_{i=1}^{n} f_{\mathbf{Y}_i}(\mathbf{y}_i;\boldsymbol{\theta}) = \prod_{i=1}^{n}\sum_{j=1}^{m} \alpha_j g_j(\mathbf{y}_i;\boldsymbol{\theta}_j)$$

Even taking the log won't help you because of the interior sum. EM to the rescue! Let $\mathbf{X} = (X_1,\ldots,X_n)$ be the mixture components, i.e. each $X_i$ is a categorical random variable that indicates which component of the mixture density produced the correspondent $\mathbf{Y}_i$

$$P(X_i = j) = \begin{cases}\alpha_0 & \text{if } x_i = 0 \\ \alpha_1 & \text{if } x_i = 1 \\ \vdots & \\ \alpha_m & \text{if } x_i = m\end{cases}$$

Hence

$$P(X_i = j) = a_j$$

And the $\mathbf{Y}_i|X_i$ takes the form

$$f_{\mathbf{Y}_i|X_i}(\mathbf{y}_i|x_i = j;\boldsymbol{\theta}) = g_j(\mathbf{y}_i;\boldsymbol{\theta}_j)$$

[7] Pronounce Yen-sen you uncultured boor!

[8] For convex $g(X)$ it's the case that $g(E(X)) \leq Eg(X)$ and for concave $g(X)$ the inequality is reversed.

The $x_i$ in $g_{x_i}(\mathbf{y}_i;\boldsymbol{\theta}_{x_i})$ "picks" which mixture component the $\mathbf{y}_i$ is generated by. Then the log-likelihood becomes

$$\begin{aligned} l((x_1,\ldots,x_n),(\mathbf{y}_1,\ldots,\mathbf{y}_n);\boldsymbol{\theta}) &= \log\left[\prod_{i=1}^{n} f_{X_i,\mathbf{Y}_i}(x_i,\mathbf{y}_i;\boldsymbol{\theta})\right] \\ &= \log\left[\prod_{i=1}^{n}\left[f_{\mathbf{Y}_i|X_i}(\mathbf{y}_i|x_i;\boldsymbol{\theta})\,P(X_i=j)\right]\right] \\ &= \log\left[\prod_{i=1}^{n}\left[g_j(\mathbf{y}_i;\boldsymbol{\theta}_j)\,\alpha_j\right]\right] \\ &= \sum_{i=1}^{n}\log\left(g_j(\mathbf{y}_i;\boldsymbol{\theta}_j)\,\alpha_j\right) \\ &= \sum_{i=1}^{n}\left[\log\alpha_j+\log\left(g_j(\mathbf{y}_i;\boldsymbol{\theta}_j)\right)\right] \end{aligned}$$

There are a lot of indices and subscripts and superscripts to keep track of: $\boldsymbol{\theta}$ is all of the parameters of all of the mixture components, $\mathbf{y}_i$ are observed samples ($n$ of them), the $X_i$ are the unobserved data (the mixture components), the $\boldsymbol{\theta}_j$ are the parameters of the $j$th mixture density, and the $\alpha_j$ are the mixture coefficients (i.e. in what proportion the $j$th density contributes). Much as we did for Example 1 we need to compute $f_{X_i|\mathbf{Y}_i}\left(x_i|\mathbf{y}_i;\boldsymbol{\theta}^{(r)}\right)=P\left(X_i=j|\mathbf{y}_i;\boldsymbol{\theta}^{(r)}\right)$ in order to compute the conditional expectations

$$\begin{aligned} P\left(X_i=j|\mathbf{y}_i;\boldsymbol{\theta}^{(r)}\right) &= \frac{f_{X_i,\mathbf{Y}_i}\left(j,\mathbf{y}_i;\boldsymbol{\theta}^{(r)}\right)}{f_{\mathbf{Y}_i}(\mathbf{y}_i;\theta)} \\ &= \frac{g_j\left(\mathbf{y}_i;\boldsymbol{\theta}_j^{(r)}\right)\alpha_j^{(r)}}{\sum_{k=1}^{m}\left(g_k\left(\mathbf{y}_i;\boldsymbol{\theta}_k^{(r)}\right)\alpha_k^{(r)}\right)} \end{aligned}$$

So again $X_i|\mathbf{Y}_i$ is categorical random variable. By i.i.d

$$P\left((X_1=j_1,\ldots,X_n=j_n)\,|\,(\mathbf{y}_1,\ldots,\mathbf{y}_n);\boldsymbol{\theta}^{(r)}\right)=\prod_{i=1}^{n}\frac{g_j\left(\mathbf{y}_i;\boldsymbol{\theta}_j^{(r)}\right)\alpha_{j_i}^{(r)}}{\sum_{k=1}^{m}\left(g_k\left(\mathbf{y}_i;\boldsymbol{\theta}_k^{(r)}\right)\alpha_k^{(r)}\right)}$$

Now we just need to take the expectation of the log-likelihood against this conditional density

$$E\left[l((X_1,\ldots,X_n),(\mathbf{y}_1,\ldots,\mathbf{y}_n);\boldsymbol{\theta})\,|\,(\mathbf{y}_1,\ldots,\mathbf{y}_n);\boldsymbol{\theta}^{(r)}\right]=$$

$$\sum_{j_1=1}^{m}\sum_{j_2=1}^{m}\cdots\sum_{j_n=1}^{m}\left(\left[\sum_{k=1}^{n}\left\{\log\alpha_{j_k}+\log\left[g_{j_k}(\mathbf{y}_k;\boldsymbol{\theta}_{j_k})\right]\right\}\right]\prod_{i=1}^{n}\frac{g_j\left(\mathbf{y}_i;\boldsymbol{\theta}_j^{(r)}\right)\alpha_j^{(r)}}{\sum_{k=1}^{m}\left(g_k\left(\mathbf{y}_i;\boldsymbol{\theta}_k^{(r)}\right)\alpha_k^{(r)}\right)}\right)$$

Pretty ugly right? Suffice it to say you're not actually taking this expectation. So let

$$\gamma_{ij}^{(r)}\coloneqq P\left(X_i=j|\mathbf{y}_i;\boldsymbol{\theta}^{(r)}\right)=\frac{g_j\left(\mathbf{y}_i;\boldsymbol{\theta}_j^{(r)}\right)\alpha_j^{(r)}}{\sum_{k=1}^{m}\left(g_k\left(\mathbf{y}_i;\boldsymbol{\theta}_k^{(r)}\right)\alpha_k^{(r)}\right)}$$

and a helpful lemma:

**Lemma 5.** *For i.i.d incomplete samples $Y_i$, with completions $X_i$*

$$E\left[l\left(\left(X_1,\ldots,X_n\right),\left(\mathbf{y}_1,\ldots,\mathbf{y}_n\right);\boldsymbol{\theta}\right)|\left(\mathbf{y}_1,\ldots,\mathbf{y}_n\right);\boldsymbol{\theta}^{(r)}\right]=\sum_{i=1}^{n}E\left[l\left(X_i|\mathbf{y}_i;\boldsymbol{\theta}\right)|\mathbf{y}_i;\boldsymbol{\theta}^{(r)}\right]$$

*where $\mathbf{y}_i$ is the ith sample.*

*Proof.* By i.i.d

$$E\left[l\left(\left(X_1,\ldots,X_n\right),\left(\mathbf{y}_1,\ldots,\mathbf{y}_n\right);\boldsymbol{\theta}\right)|\left(\mathbf{y}_1,\ldots,\mathbf{y}_n\right);\boldsymbol{\theta}^{(r)}\right]=$$

$$E\left[\log\left[\prod_{i=1}^{n}f\left(X_i|\left(\mathbf{y}_1,\ldots,\mathbf{y}_n\right);\boldsymbol{\theta}\right)\right]|\left(\mathbf{y}_1,\ldots,\mathbf{y}_n\right);\boldsymbol{\theta}^{(r)}\right]=$$

$$E\left[\sum_{i=1}^{n}\log\left[f\left(X_i|\left(\mathbf{y}_1,\ldots,\mathbf{y}_n\right);\boldsymbol{\theta}\right)\right]|\left(\mathbf{y}_1,\ldots,\mathbf{y}_n\right);\boldsymbol{\theta}^{(r)}\right]=(\text{ since } X_i\perp X_j \text{ for } i\neq j)$$

$$\sum_{i=1}^{n}E\left[\log\left[f\left(X_i|\left(\mathbf{y}_1,\ldots,\mathbf{y}_n\right);\boldsymbol{\theta}\right)\right]|\left(\mathbf{y}_1,\ldots,\mathbf{y}_n\right);\boldsymbol{\theta}^{(r)}\right]=(\text{ since } P\left(X_i=x_i|\left(\mathbf{Y}_1,\ldots,\mathbf{Y}_n\right)\right)=P\left(X_i=x_i|\mathbf{Y}_i\right))$$

$$\sum_{i=1}^{n}E\left[\log\left[f\left(X_i|\mathbf{y}_i;\boldsymbol{\theta}\right)\right]|\mathbf{y}_i;\boldsymbol{\theta}^{(r)}\right]=$$

$$\sum_{i=1}^{n}E\left[l\left(X_i|\mathbf{y}_i;\boldsymbol{\theta}\right)|\mathbf{y}_i;\boldsymbol{\theta}^{(r)}\right]$$

□

Therefore we only need to compute $E\left[l\left(X_i|\mathbf{y}_i;\boldsymbol{\theta}\right)|\mathbf{y}_i;\boldsymbol{\theta}^{(r)}\right]$ to perform the E-step. Let $j_i\to j$ for convenience and $(\mu_j,\Sigma_j)$ be the parameters for the $j$th Gaussian. Actually let's take the 1 dimensional Gaussian case[9] so $(\mu_j,\Sigma_j)=\left(\mu_j,\sigma_j^2\right)$ and $\mathbf{y}_i\to y_i$.

$$\begin{aligned}E\left[l\left(X_i|y_i;\boldsymbol{\theta}\right)|y_i;\boldsymbol{\theta}^{(r)}\right] &= \sum_{j=1}^{m}\left(\left[\log\alpha_j+\log\left(g_j\left(y_i;\boldsymbol{\theta}_j\right)\right)\right]\gamma_{ij}^{(r)}\right)\\ &= \sum_{j=1}^{m}\left(\left[\log\alpha_j+\left(-\frac{1}{2}\left(\log\left(\sigma_j^2\right)+\log\left(2\pi\right)\right)-\frac{1}{2\sigma_j^2}\left(y_i-\mu_j\right)^2\right)\right]\gamma_{ij}^{(r)}\right)\end{aligned}$$

Then

$$\begin{aligned}E\left[l\left(\left(X_1,\ldots,X_n\right)|\left(y_1,\ldots,y_n\right);\boldsymbol{\theta}\right)|\left(y_1,\ldots,y_n\right);\boldsymbol{\theta}^{(r)}\right] &=\\ \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\left[\log\alpha_j+\log\left(g_j\left(y_i;\boldsymbol{\theta}_j\right)\right)\right]\gamma_{ij}^{(r)}\right) &=\\ \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\left[\log\alpha_j+\left(-\frac{1}{2}\left(\log\left(\sigma_j^2\right)+\log\left(2\pi\right)\right)-\frac{1}{2\sigma_j^2}\left(y_i-\mu_j\right)^2\right)\right]\gamma_{ij}^{(r)}\right)\end{aligned}$$

[9] [?] is a good place for the general case.

Define

$$n_j^{(r)} = \sum_{i=1}^{n} \gamma_{ij}^{(r)} = \sum_{i=1}^{n} P\left(X_i = j | \mathbf{y}_i; \boldsymbol{\theta}^{(r)}\right)$$

This is something like the portion of the samples that were generated by the $j$th component of the density[10]. The expectation becomes

$$\begin{aligned}
E\left[l\left((X_1,\ldots,X_n)\,|\,(y_1,\ldots,y_n);\boldsymbol{\theta}\right)|\,(y_1,\ldots,y_n);\boldsymbol{\theta}^{(r)}\right] &= \\
\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\left[\log\alpha_j + \left(-\frac{1}{2}\left(\log\left(\sigma_j^2\right)+\log\left(2\pi\right)\right) - \frac{1}{2\sigma_j^2}\left(y_i-\mu_j\right)^2\right)\right]\gamma_{ij}^{(r)}\right) &= \\
\sum_{i=1}^{n}\sum_{j=1}^{m}\log\alpha_j\gamma_{ij}^{(r)} - \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{2}\left(\log\left(\sigma_j^2\right)+\log\left(2\pi\right)\right)\gamma_{ij}^{(r)} - \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{2\sigma_j^2}\left(y_i-\mu_j\right)^2\gamma_{ij}^{(r)} &= \\
\sum_{j=1}^{m} n_j^{(r)}\log\alpha_j - \frac{1}{2}\sum_{j=1}^{m} n_j^{(r)}\left(\log\left(\sigma_j^2\right)+\log\left(2\pi\right)\right) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(y_i-\mu_j\right)^2}{\sigma_j^2}\gamma_{ij}^{(r)} &
\end{aligned}$$

This is the expression that needs to be maximized with respect to $\alpha_j, \mu_j, \sigma_j^2$[11]. The fully specified maximization problem is

$$\begin{aligned}
&\max_{\alpha_j,\mu_j,\sigma_j^2} E\left[l\left((X_1,\ldots,X_n)\,|\,(y_1,\ldots,y_n);\boldsymbol{\theta}\right)|\,(y_1,\ldots,y_n);\boldsymbol{\theta}^{(r)}\right] \\
&\qquad\Longleftrightarrow \\
&\max_{\alpha_j,\mu_j,\sigma_j^2} \sum_{j=1}^{m} n_j^{(r)}\log\alpha_j - \frac{1}{2}\sum_{j=1}^{m} n_j^{(r)}\left(\log\left(\sigma_j^2\right)+\log\left(2\pi\right)\right) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(y_i-\mu_j\right)^2}{\sigma_j^2}\gamma_{ij}^{(r)} \\
&\text{subject to} \sum_{j=1}^{m}\alpha_j = 1,\ \alpha_j \geq 0
\end{aligned}$$

This looks tough but because it's a bunch of uncoupled (somewhat) terms summed together so it's easier than the original log-likelihood

$$\log\left[\sum_{j=1}^{m}\alpha_j g_j\left(\mathbf{y}_i;\boldsymbol{\theta}_j\right)\right]$$

The local maxima for $\alpha_j, \mu_j, \sigma_j^2$ become the new iterates, ie. $\alpha_j^{(r+1)}, \mu_j^{(r+1)}, \left(\sigma_j^2\right)^{(r+1)}$. To perform the constrained maximization (only the $\alpha_j$) we use Lagrange multipliers; form the Lagrangian

$$\mathcal{L}\left(\boldsymbol{\alpha},\lambda\right) = \sum_{j=1}^{m} n_j^{(r)}\log\alpha_j + \lambda\left(\sum_{j=1}^{m}\alpha_j - 1\right)$$

[10] Why? Because $\sum_{j=1}^{m} n_j = \sum_{j=1}^{m}\sum_{i=1}^{n}\gamma_{ij} = \sum_{i=1}^{n} 1 = n$.

[11] Pay attention to the difference between the arguments that come from the log $\alpha_j, \mu_j, \sigma_j^2$ and the current estimates of the parameters, those with iterate superscripts $(r)$ that come from conditional density, with respect to which the expectation was taken.

Then computing derivatives and setting to zero to find $\alpha_j$

$$\frac{\partial \mathcal{L}}{\partial \alpha_j} = \frac{n_j^{(r)}}{\alpha_j} + \lambda$$

Therefore the critical[12] $\alpha_j$ satisfy

$$\alpha_j = -\frac{n_j^{(r)}}{\lambda}$$

To eliminate $\lambda$ use the equality constraint, i.e. $\sum_{j=1}^m \alpha_j = 1$ implies $\lambda = -\sum_{j=1}^m n_j^{(r)}$ and hence

$$\alpha_j^{(r+1)} = \frac{n_j^{(r)}}{\sum_{j=1}^m n_j^{(r)}} = \frac{n_j^{(r)}}{n}$$

So $\alpha_j^{(r+1)}$ is just the current best estimate for how many of the samples were generated by the $j$th component of the mixture density[13]. To compute the updates $\mu_j^{(r+1)}$ we just need to do the standard maximization of $E\left[l\left((X_1,\ldots,X_n)\,|\,(y_1,\ldots,y_n);\boldsymbol{\theta}\right)|\,(y_1,\ldots,y_n);\boldsymbol{\theta}^{(r)}\right]$ since the $\mu_j$ are unconstrained. Therefore

$$\frac{\partial}{\partial \mu_j} E\left[l\left((X_1,\ldots,X_n)\,|\,(y_1,\ldots,y_n);\boldsymbol{\theta}\right)|\,(y_1,\ldots,y_n);\boldsymbol{\theta}^{(r)}\right] = \sum_{i=1}^n \frac{(y_i-\mu_j)}{\sigma_j^2}\gamma_{ij}^{(r)} = 0$$

Note the sum of $j$ disappears because only the $j$th term is non-zero. Then simplifying further

$$\begin{aligned}
\sum_{i=1}^n \frac{(y_i-\mu_j)}{\sigma_j^2}\gamma_{ij}^{(r)} &= \frac{1}{\sigma_j^2}\left(\sum_{i=1}^n \gamma_{ij}^{(r)} y_i - \mu_j \sum_{i=1}^n \gamma_{ij}^{(r)}\right) \\
&= \frac{1}{\sigma_j^2}\left(\sum_{i=1}^n \gamma_{ij}^{(r)} y_i - \mu_j n_j^{(r)}\right) = 0 \\
&\Rightarrow \\
\mu_j^{(r+1)} &= \frac{1}{n_j^{(r)}}\sum_{i=1}^n \gamma_{ij}^{(r)} y_i
\end{aligned}$$

Finally doing the same kind of thing for $\left(\sigma_j^2\right)^{(r+1)}$

$$\frac{\partial}{\partial \sigma_j^2} E\left[l\left((X_1,\ldots,X_n)\,|\,(y_1,\ldots,y_n);\boldsymbol{\theta}\right)|\,(y_1,\ldots,y_n);\boldsymbol{\theta}^{(r)}\right] = -\frac{1}{2}\frac{n_j^{(r)}}{\sigma_j^2} + \frac{1}{2}\sum_{i=1}^n \frac{\left(y_i-\mu_j^{(r+1)}\right)^2}{\left(\sigma_j^2\right)^2}\gamma_{ij}^{(r)} = 0$$

where $\mu_j$ now gets a superscript because it's already been updated. Hence

$$\left(\sigma_j^2\right)^{(r+1)} = \frac{1}{n_j^{(r+1)}}\sum_{i=1}^n \left(y_i-\mu_j^{(r+1)}\right)^2 \gamma_{ij}^{(r)}$$

[12] You need to do the convex analysis (second-deriative test) to determine whether $\alpha_j$ are maxima but they are and I'm not going to.

[13] Pretty much the epitome of the MLE for a parameter.

In summary the update equations for a univariate GMM

$$\alpha_j^{(r+1)} = \frac{n_j^{(r)}}{n}$$
$$\mu_j^{(r+1)} = \frac{1}{n_j^{(r)}} \sum_{i=1}^{n} \gamma_{ij}^{(r)} y_i$$
$$\left(\sigma_j^2\right)^{(r+1)} = \frac{1}{n_j^{(r+1)}} \sum_{i=1}^{n} \left(y_i - \mu_j^{(r+1)}\right)^2 \gamma_{ij}^{(r)}$$

where

$$\gamma_{ij}^{(r)} = \frac{g_j\left(\mathbf{y}_i; \boldsymbol{\theta}_j^{(r)}\right) \alpha_j^{(r)}}{\sum_{k=1}^{m} \left(g_k\left(\mathbf{y}_i; \boldsymbol{\theta}_k^{(r)}\right) \alpha_k^{(r)}\right)}$$
$$n_j^{(r)} = \sum_i \gamma_{ij}^{(r)}$$

For multivariate GMM the only difference is $\sigma_j^2 \to \Sigma_j$ and so you need to use matrix derivatives. Consult [**?**] for the full derivation

$$\Sigma_j^{(r+1)} = \frac{1}{n_j^{(r+1)}} \sum_{i=1}^{n} \left(\mathbf{y}_i - \boldsymbol{\mu}_j^{(r+1)}\right) \left(\mathbf{y}_i - \boldsymbol{\mu}_j^{(r+1)}\right)^T \gamma_{ij}^{(r)}$$