

**Clarifying the Relationship of $P(X, Z|\Theta)$ and $P(X|\Theta)$ in EM
(and along the way seeing how cool the 1-of-K representation is)**

Jan 29, 2012

In class, the question arose about how the labeled joint likelihood $P(X, Z|\Theta)$ in EM is related to the original unlabeled likelihood $P(X|\Theta)$. This note will show that, as the notation suggests, $P(X|\Theta)$ is just the marginal distribution of $P(X, Z|\Theta)$. That is,

$$P(X|\Theta) = \sum_Z P(X, Z|\Theta) .$$

Recall that we have a set of N datapoints $X = \{x_1, x_2, \dots, x_N\}$ thought to come from a K -component Gaussian mixture model with parameters $\Theta = \{w_1, w_2, \dots, w_K, \mu_1, \mu_2, \dots, \mu_K, \Sigma_1, \Sigma_2, \dots, \Sigma_K\}$ where w_k, μ_k, Σ_k is the mixing weight, mean vector and covariance matrix for the k th Gaussian component. We want to estimate the parameters Θ using maximum likelihood estimation. MLE starts with assuming all datapoints x_n are independent and identically distribution, and forms the likelihood function

$$P(X|\Theta) = \prod_{n=1}^N \sum_{k=1}^K w_k N(x_n|\mu_k, \Sigma_k) .$$

Also recall that we have difficulty carrying out MLE in this case because we can't easily take the log of this formula to form the log likelihood. Informally speaking, the log operation gets "blocked" by the summation over K so that it cannot get into the interior of the sum to simplify the exponential in $N(x_n|\mu_k, \Sigma_k)$. To overcome this problem, for each datapoint x_n we hypothesized a vector of K latent variables $z_n = [z_{n,1}, z_{n,2}, \dots, z_{n,K}]$ such that, if datapoint n comes from Gaussian component k , the value of $z_{n,k}$ will be 1 and all the other $K-1$ elements will be 0. This so-called "1 of K " vector representation can be interpreted as a label saying which component x_n belongs to.

Assuming some oracle gives us the correct values of the $N \times K$ variables $Z = \{z_1, z_2, \dots, z_N\}$, the new, labeled joint likelihood function becomes

$$P(X, Z|\Theta) = \prod_{n=1}^N \prod_{k=1}^K [w_k N(x_n|\mu_k, \Sigma_k)]^{z_{n,k}} .$$

The summation is gone, and taking the log likelihood can now proceed without difficulty.

Before showing how to sum this expression over all values of Z to get a marginal distribution depending only on X , we first take a closer look at how a 1-of- K vector serves to "pick out" a specific component in this formula. First note that for a single data point x_n ,

$$\prod_{k=1}^K [w_k N(x_n|\mu_k, \Sigma_k)]^{z_{n,k}} = [w_1 N(x_n|\mu_1, \Sigma_1)]^{z_{n,1}} [w_2 N(x_n|\mu_2, \Sigma_2)]^{z_{n,2}} \dots [w_K N(x_n|\mu_K, \Sigma_K)]^{z_{n,K}} .$$

Now, without loss of generality, assume that x_n is labeled as belonging to Gaussian component 1. Then in the 1-of- K representation, $z_{n,1} = 1$ while $z_{n,2} = z_{n,3} = \dots = z_{n,K} = 0$. Plugging this in, we

see that above product reduces to

$$\prod_{k=1}^K [w_k N(x_n|\mu_k, \Sigma_k)]^{z_{n,k}} = w_1 N(x_n|\mu_1, \Sigma_1)$$

where just the relevant factor for Gaussian component 1 has been picked out. In general, if we know that $z_{n,j} = 1$, the product selects the j th component factor $w_j N(x_n|\mu_j, \Sigma_j)$.

Define summation of a function over a set to be applying the function to each value in the set and summing up the results, that is:

$$\sum_{i \in \{a,b,c\}} f(i) = f(a) + f(b) + f(c)$$

and consider that 1-of-K vector z_n takes values in the set $\{[1, 0, \dots, 0], [0, 1, \dots, 0], \dots, [0, 0, \dots, 1]\}$. Combining this with the above discussion, we can see that

$$\sum_{z_n} \prod_{k=1}^K [w_k N(x_n|\mu_k, \Sigma_k)]^{z_{n,k}} = \sum_{k=1}^K w_k N(x_n|\mu_k, \Sigma_k) .$$

Finally, let's consider summing $P(X, Z|\Theta)$ over all possible values of Z . This means over all possible combinations of values that the 1-of-K vectors z_1, z_2, \dots, z_n can take. Therefore

$$\sum_Z P(X, Z|\Theta) = \sum_{z_1} \sum_{z_2} \dots \sum_{z_n} P(X, Z|\Theta) .$$

To simplify this, note that

$$\sum_a \sum_b \sum_c f(a)f(b)f(c) = \left(\sum_a f(a) \right) \left(\sum_b f(b) \right) \left(\sum_c f(c) \right) .$$

We thus have

$$\begin{aligned} \sum_Z P(X, Z|\Theta) &= \sum_{z_1} \sum_{z_2} \dots \sum_{z_n} \left(\prod_{n=1}^N \prod_{k=1}^K [w_k N(x_n|\mu_k, \Sigma_k)]^{z_{n,k}} \right) \\ &= \prod_{n=1}^N \left(\sum_{z_n} \prod_{k=1}^K [w_k N(x_n|\mu_k, \Sigma_k)]^{z_{n,k}} \right) \\ &= \prod_{n=1}^N \sum_{k=1}^K w_k N(x_n|\mu_k, \Sigma_k) = P(X|\Theta) . \end{aligned}$$

This establishes the relationship we were trying to show.

Example. Consider 2 data points and 2 Gaussian clusters. To simplify the equations, define the shorthand notation $[f_{nk}] = w_k N(x_n | \mu_k, \Sigma_k)$.

$$\begin{aligned}
 P(X, Z | \Theta) &= \prod_{n=1}^2 \prod_{k=1}^2 [w_k N(x_n | \mu_k, \Sigma_k)]^{z_{n,k}} \\
 &= [f_{11}]^{z_{1,1}} [f_{12}]^{z_{1,2}} [f_{21}]^{z_{2,1}} [f_{22}]^{z_{2,2}} \\
 \\
 \sum_{z_1} \sum_{z_2} P(X, Z | \Theta) &= \sum_{z_1} \sum_{z_2} [f_{11}]^{z_{1,1}} [f_{12}]^{z_{1,2}} [f_{21}]^{z_{2,1}} [f_{22}]^{z_{2,2}} \\
 &= \sum_{z_1} \left([f_{11}]^{z_{1,1}} [f_{12}]^{z_{1,2}} [f_{21}] + [f_{11}]^{z_{1,1}} [f_{12}]^{z_{1,2}} [f_{22}] \right) \\
 &= \left([f_{11}] [f_{21}] + [f_{11}] [f_{22}] + [f_{12}] [f_{21}] + [f_{12}] [f_{22}] \right) \\
 &= \left([f_{11}] + [f_{12}] \right) \left([f_{21}] + [f_{22}] \right) \\
 &= \prod_{n=1}^2 \sum_{k=1}^2 [f_{nk}] \\
 &= \prod_{n=1}^2 \sum_{k=1}^2 w_k N(x_n | \mu_k, \Sigma_k) \\
 &= P(X | \Theta)
 \end{aligned}$$

Final Thoughts. The subtext of the question that was raised in class is: why do we believe that applying MLE to $P(X, Z | \Theta)$ is equivalent to solving the initial problem of computing the MLE of $P(X | \Theta)$? For example, is the globally optimal value Θ_1^* we get from solving one of them going to be the same as the globally optimal value Θ_2^* we get from solving the other one? I think it is at least plausible that the two solutions are closely related to each other... our new problem is solving for a superset of variables, the labels AND the mixture component parameters, and we can then just ignore the values of the label variables if we don't want them (actually, having those soft labels is a useful extra bit of information in many problems, and worth having). But it is kind of a moot issue, because the estimated value $\hat{\Theta}$ computed by EM is only guaranteed to be a local optimum, not the global optimum, and we don't know how to solve the original MLE problem anyways. At least our artificially-created new problem gives us a way to proceed.