

GMM Appendix B: Derivation of the Complete Data Likelihood

Starting Point: The Generative Model¹

Recall that the GMM generates each point by:

1. **First:** Choose cluster k with probability π_k
2. **Then:** Generate point from that cluster: $x_i \sim \mathcal{N}(\mu_k, \Sigma_k)$

1. Courtesy of Claude.ai

Step 1: Joint Probability with Known Assignment

If we **know** that point x_i was generated by cluster z_i , the joint probability is:

$$p(x_i, z_i \mid \theta) = p(z_i \mid \theta) \cdot p(x_i \mid z_i, \theta) = \pi_{z_i} \cdot \mathcal{N}(x_i \mid \mu_{z_i}, \Sigma_{z_i})$$

This is just applying the chain rule: probability of choosing cluster z_i times probability of generating x_i from that cluster.

Summing this joint probability over the K possible values of z_i gives back the mixture density from the lecture:

$$p(x_i \mid \theta) = \sum_{k=1}^K p(x_i, z_i = k \mid \theta) = \sum_{k=1}^K \pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k).$$

So the observed-data likelihood is the complete-data likelihood with the unknown z_i averaged out. That averaging is where the sum inside the log comes from.

Step 2: Complete Data Likelihood (All Points)

For all n independent data points:

$$p(X, Z \mid \theta) = \prod_{i=1}^n \pi_{z_i} \cdot \mathcal{N}(x_i \mid \mu_{z_i}, \Sigma_{z_i})$$

Step 3: Take the Log

$$\log p(X, Z \mid \theta) = \sum_{i=1}^n [\log \pi_{z_i} + \log \mathcal{N}(x_i \mid \mu_{z_i}, \Sigma_{z_i})]$$

Step 4: The Indicator Function Trick

Here's the key insight! Since z_i takes on exactly one value from $\{1, 2, \dots, K\}$, we can rewrite π_{z_i} using an indicator function:

$$\log \pi_{z_i} = \sum_{k=1}^K \mathbb{1}(z_i = k) \log \pi_k$$

Why? Because only **one** term in this sum is non-zero (when $k = z_i$), and it equals $\log \pi_{z_i}$.

Similarly:

$$\log \mathcal{N}(x_i \mid \mu_{z_i}, \Sigma_{z_i}) = \sum_{k=1}^K \mathbb{1}(z_i = k) \log \mathcal{N}(x_i \mid \mu_k, \Sigma_k)$$

Step 5: Final Form

Combining everything:

$$\log p(X, Z \mid \theta) = \sum_{i=1}^n \sum_{k=1}^K \mathbb{1}(z_i = k) [\log \pi_k + \log \mathcal{N}(x_i \mid \mu_k, \Sigma_k)]$$

Compare it with the observed-data log-likelihood that the lecture tried to maximize directly:

$$\ell(\theta) = \log p(X \mid \theta) = \sum_{i=1}^n \log \left(\sum_{k=1}^K \pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k) \right).$$

The two differ only in where the sum over clusters sits: **inside** the log in the observed-data version, **outside** it in the complete-data version. Inside, every parameter is tangled up with every other one in each term. Outside, each term involves a single cluster's parameters, so the expression separates by cluster.

Why Use This Form?

As mentioned earlier, this form is **much easier to work with** because:

- The log operates on individual terms (not on a sum)
- In the E-step, we replace the indicator $\mathbb{1}(z_i = k)$ with its expected value (the responsibility γ_{ik}); see [Appendix C](#)
- In the M-step, the separated form gives closed-form updates, one cluster at a time; see [Appendix D](#)