

GMM Appendix D: Deriving the M-Step

Starting Point: The Expected Complete-Data Log-Likelihood

Appendix C ended with the Q -function, the expected complete-data log-likelihood under the responsibilities computed in the E-step:

$$Q(\theta \mid \theta^{(t)}) = \sum_{i=1}^n \sum_{k=1}^K \gamma_{ik} [\log \pi_k + \log \mathcal{N}(x_i \mid \mu_k, \Sigma_k)], \quad N_k = \sum_{i=1}^n \gamma_{ik}.$$

The responsibilities γ_{ik} were computed from the *current* parameters $\theta^{(t)}$ and are **constants** in what follows. The M-step chooses new parameters $\theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$ to maximize Q . The lecture states the answer:

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^n \gamma_{ik} x_i, \quad \Sigma_k = \frac{1}{N_k} \sum_{i=1}^n \gamma_{ik} (x_i - \mu_k)(x_i - \mu_k)^T, \quad \pi_k = \frac{N_k}{n}.$$

This appendix derives it. Nothing here is harder than the single-Gaussian derivation in

Why It Separates

Regroup Q by cluster:

$$Q(\theta \mid \theta^{(t)}) = \underbrace{\sum_{k=1}^K N_k \log \pi_k}_{\text{depends only on the } \pi_k} + \underbrace{\sum_{k=1}^K \sum_{i=1}^n \gamma_{ik} \log \mathcal{N}(x_i \mid \mu_k, \Sigma_k)}_{\text{depends only on } \mu_k \text{ and } \Sigma_k}.$$

Each μ_k and Σ_k appears in exactly one term, so maximizing Q over all the parameters at once is the same as maximizing each cluster's term on its own. That is what the sum inside the log had prevented: in $\ell(\theta)$, every parameter appeared inside every term and the derivatives were coupled. The mixing weights are coupled to each other only through the constraint $\sum_k \pi_k = 1$, handled last.

The Mean and Variance in One Dimension

Take $d = 1$ first, so $\Sigma_k = \sigma_k^2$. Then

$$\log \mathcal{N}(x_i \mid \mu_k, \sigma_k^2) = -\frac{1}{2} \log(2\pi) - \log \sigma_k - \frac{(x_i - \mu_k)^2}{2\sigma_k^2},$$

and cluster k 's term in Q is

$$Q_k(\mu_k, \sigma_k) = \sum_{i=1}^n \gamma_{ik} \left[-\log \sigma_k - \frac{(x_i - \mu_k)^2}{2\sigma_k^2} \right] + \text{const.}$$

Mean. Differentiate with respect to μ_k and set to zero:

$$\frac{\partial Q_k}{\partial \mu_k} = \sum_{i=1}^n \gamma_{ik} \frac{x_i - \mu_k}{\sigma_k^2} = 0 \implies \sum_{i=1}^n \gamma_{ik} x_i = \mu_k \sum_{i=1}^n \gamma_{ik} \implies \mu_k = \frac{1}{N_k} \sum_{i=1}^n \gamma_{ik} x_i$$

Higher Dimensions

The same two steps with Σ_k in place of σ_k^2 . Write

$$\log \mathcal{N}(x \mid \mu, \Sigma) = -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu),$$

and use three facts from matrix calculus, valid for symmetric positive-definite Σ :

$$\nabla_{\mu} (x - \mu)^T \Sigma^{-1} (x - \mu) = -2 \Sigma^{-1} (x - \mu), \quad \nabla_{\Sigma} \log |\Sigma| = \Sigma^{-1}, \quad \nabla_{\Sigma} (x -$$

Mean. $\nabla_{\mu_k} Q_k = \sum_i \gamma_{ik} \Sigma_k^{-1} (x_i - \mu_k) = 0$. Multiply through by Σ_k and solve: $\mu_k = \frac{1}{N_k} \sum_i \gamma_{ik} x_i$, unchanged.

Covariance. With μ_k fixed,

$$\nabla_{\Sigma_k} Q_k = \sum_{i=1}^n \gamma_{ik} \left[-\frac{1}{2} \Sigma_k^{-1} + \frac{1}{2} \Sigma_k^{-1} (x_i - \mu_k)(x_i - \mu_k)^T \Sigma_k^{-1} \right] = 0.$$

Multiply on the left and on the right by Σ_k :

The Mixing Weights

What remains is

$$\max_{\pi} \sum_{k=1}^K N_k \log \pi_k \quad \text{subject to} \quad \sum_{k=1}^K \pi_k = 1, \quad \pi_k \geq 0.$$

Without the constraint every π_k would run off to infinity, so we need a Lagrange multiplier:

$$\mathcal{L}(\pi, \lambda) = \sum_{k=1}^K N_k \log \pi_k + \lambda \left(\sum_{k=1}^K \pi_k - 1 \right).$$

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = \frac{N_k}{\pi_k} + \lambda = 0 \implies \pi_k = -\frac{N_k}{\lambda}.$$

Summing both sides over k and using the constraint, $1 = -\frac{1}{\lambda} \sum_k N_k = -\frac{n}{\lambda}$, since the responsibilities of each point sum to 1 and there are n points. Hence $\lambda = -n$ and

Summary and Sanity Checks

$$\mu_k = \frac{1}{N_k} \sum_i \gamma_{ik} x_i, \quad \Sigma_k = \frac{1}{N_k} \sum_i \gamma_{ik} (x_i - \mu_k)(x_i - \mu_k)^T, \quad \pi_k = \frac{N_k}{n}$$

- $K = 1$. Every $\gamma_{i1} = 1$ and $N_1 = n$: the formulas reduce to the single-Gaussian MLE of lecture 05.
- **Labels known.** Every $\gamma_{ik} \in \{0, 1\}$: the formulas reduce to the per-cluster mean, covariance and fraction, the “if we knew the labels” case in the lecture. Under equal spherical covariances, the mean update is the k -means centroid step.
- **Closed form.** Because Q separates, the M-step is a direct maximization, not an inner iteration. The only iteration in EM comes from the responsibilities depending on the parameters, which is why the E-step and M-step have to alternate.

What this appendix does *not* show is that raising Q also raises the observed-data log-likelihood ℓ , which is what makes the alternation converge. That is a separate argument.