Notes on Semi-Supervised Expectation Maximization

Christoph Heindl
PROFACTOR GmbH, Austria
christoph.heindl@gmail.com

Josef Scharinger
JKU, Austria
josef.scharinger@jku.at

Abstract

This work considers the Expectation Maximization (EM) algorithm in the semi-supervised setting. First, the general form for semi-supervised version of maximum likelihood is derived from the Latent Variable Model (LVM). Since the involved integrals are usually intractable, a surrogate objective function based on the Evidence Lower Bound (ELBO) is introduced. Next, we derive the equations of the semi-supervised EM. Finally, the concrete equations for a fitting a Gaussian Mixture Model (GMM) using labeled and unlabeled data are deduced.

1 Introduction

The EM algorithm, first formalized by Dempster et al. [1], is a statistical method for maximum likelihood parameter estimation. It is particularly useful when the model contains latent variables. This work derives the EM equations for the semi-supervised setting, where we can group the set of random variables into fully and partially observed ones.

Consider a generative latent variable model (LVM) as shown in Figure 1. We assume *iid* observations $\{(x_i, z_i)\}_{i \leq N}$ and partial observations $\{\tilde{x}_j\}_{j \leq M}$. The marginal log-likelihood of the generative probabilistic model associated with the observations is given by

$$\begin{aligned}
\log p(X, Z, \tilde{X}|\theta) &= \log \int p(X, Z, \tilde{X}, \tilde{Z}|\theta)\, d\tilde{Z} \\
&= \log \int p(X, Z|\theta) p(\tilde{X}, \tilde{Z}|\theta)\, d\tilde{Z} \\
&= \log p(X, Z|\theta) \int p(\tilde{X}, \tilde{Z}|\theta)\, d\tilde{Z} \\
&= \log p(X, Z|\theta) + \log \int p(\tilde{X}, \tilde{Z}|\theta)\, d\tilde{Z}, \qquad (1)
\end{aligned}$$

where we have made use of the independence assumptions of our model, abbreviated $\{X_i\}_{i \leq N}$ by X and similar for the other types of random variables. Equation 1 is called the generative approach to semi-supervised learning. Often, this equation is seen with an additional balancing factor

$$\log p(X, Z|\theta) + \lambda \log \int p(\tilde{X}, \tilde{Z}|\theta)\, d\tilde{Z}.$$

In semi-supervised learning then seeks to maximize the marginal likelihood by estimating θ so that

$$\theta^* = \arg\max_{\theta} \; \log p(X, Z, \tilde{X}|\theta) \tag{2}$$

$$= \arg\max_{\theta} \; \left[\log p(X, Z|\theta) + \log \int p(\tilde{X}, \tilde{Z}|\theta)\, d\tilde{Z}\right]. \tag{3}$$

In practice the integral of the second term on the right hand side (rhs) is often intractable. We therefore seek to find a surrogate objective that is tractable—the evidence lower bound (ELBO).

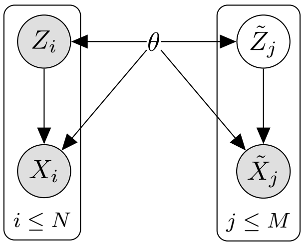

Figure 1: Generative latent variable model for the semi-supervised case.

2 ELBO in the Semi-Supervised Setting

Consider the following inequality derived from Equation 1.

$$
\begin{aligned}
\log p(X, Z, \tilde{X}|\theta) &= \log \int p(X, Z, \tilde{X}, \tilde{Z}|\theta)\, d\tilde{Z} \\
&= \log \int q(\tilde{Z}) \frac{p(X, Z, \tilde{X}, \tilde{Z}|\theta)}{q(\tilde{Z})}\, d\tilde{Z} \\
&\geq \int q(\tilde{Z}) \log \frac{p(X, Z, \tilde{X}, \tilde{Z}|\theta)}{q(\tilde{Z})}\, d\tilde{Z} && \text{by Jensen's ineq.} \\
&= \int q(\tilde{Z}) \log \frac{p(X, Z|\theta) p(\tilde{X}, \tilde{Z}|\theta)}{q(\tilde{Z})}\, d\tilde{Z} && \text{by LVM assumptions} \\
&= \int q(\tilde{Z}) \left[\log p(X, Z|\theta) + \log \frac{p(\tilde{X}, \tilde{Z}|\theta)}{q(\tilde{Z})} \right] d\tilde{Z} \\
&= \int q(\tilde{Z}) \log p(X, Z|\theta)\, d\tilde{Z} + \int q(\tilde{Z}) \log \frac{p(\tilde{X}, \tilde{Z}|\theta)}{q(\tilde{Z})}\, d\tilde{Z} \\
&= \log p(X, Z|\theta) \underbrace{\int q(\tilde{Z})\, d\tilde{Z}}_{=1} + \int q(\tilde{Z}) \log \frac{p(\tilde{X}, \tilde{Z}|\theta)}{q(\tilde{Z})}\, d\tilde{Z} \\
&= \log p(X, Z|\theta) + \int q(\tilde{Z}) \log \frac{p(\tilde{X}, \tilde{Z}|\theta)}{q(\tilde{Z})}\, d\tilde{Z} \\
&= \log p(X, Z|\theta) + \underbrace{\mathbb{E}_{\tilde{Z} \sim q(\tilde{Z})} \log \frac{p(\tilde{X}, \tilde{Z}|\theta)}{q(\tilde{Z})}}_{\text{ELBO}(q,\theta)} && (4) \\
&= \mathcal{F}(q, \theta). && (5)
\end{aligned}
$$

In the above we abbreviated $q(\tilde{Z}|X, Z, \tilde{X}, \theta)$ by $q(\tilde{Z})$ to avoid clutter.

3 EM in the Semi-Supervised Setting

The expectation maximization algorithm iteratively maximizes

$$(\hat{q}, \hat{\theta}) = \arg\max_{q,\theta}\ \mathcal{F}(q, \theta).$$

Let us introduce a 'time' dependency on parameters θ^t and the form of $q(\tilde{Z})^t$. In the **E-step** we optimize

$$q^{t+1} = \arg\max_{q}\ \mathcal{F}(q, \theta^t).$$

It is well known that choosing

$$q^{t+1}(\tilde{Z}) = p(\tilde{Z}|X, Z, \tilde{X}, \theta^t)$$

makes the ELBO tight(i.e turn Inequality 4 into an equality). In the LVM case the above simplifies due to independence assumptions as follows

$$q^{t+1}(\tilde{Z}) = p(\tilde{Z}|\tilde{X}, \theta^t). \tag{6}$$

Remark Plugging Equation 6 into $\mathcal{F}(q^{t+1}, \theta^t)$ leads to an equality between the marginal log-likelihood and the left hand side (lhs), i.e. makes the bound is tight. We can also see this by starting from the KL-divergence as follows

$$KL(q^{t+1}(\tilde{Z}|\tilde{X}, X, Z, \theta^t), p(\tilde{Z}|\tilde{X}, X, Z, \theta^t))$$

which leads to

$$\log p(X, Z, \tilde{X}|\theta^t) = \log p(X, Z|\theta^t) + ELBO(q^{t+1}, \theta^t) + KL(q^{t+1}(\tilde{Z}|\tilde{X}, X, Z, \theta^t), p(\tilde{Z}|\tilde{X}, X, Z, \theta^t)).$$

Setting $q^{t+1}(\tilde{Z}) = p(\tilde{Z}|\tilde{X}, X, Z, \theta^t)$ leads to a vanishing KL term and we arrive at $\mathcal{F}(q^{t+1}, \theta^t)$—therefore holding with equality.

Remark Note that in the E-Step the term involving supervised data $\log p(X, Z|\theta)$ does not depend on q and thus does not influence the shape of q. Intuitively this happens because (1) we explicity introduce a distribution q only over the unobserved variables $\tilde{Z}$ and (2) the independence assumptions of the LVM leads to this specific factorization.

In the **M-Step** we optimize

$$\begin{aligned}
\theta^{t+1} &= \arg\max_\theta \; \mathcal{F}(q^{t+1}, \theta) && (7)\\
&= \arg\max_\theta \; \log p(X, Z|\theta) + \mathbb{E}_{\tilde{Z}\sim q^{t+1}(\tilde{Z})} \log \frac{p(\tilde{X}, \tilde{Z}|\theta)}{q^{t+1}(\tilde{Z})}\\
&= \arg\max_\theta \; \log p(X, Z|\theta) + \mathbb{E}_{\tilde{Z}\sim q^{t+1}(\tilde{Z})} \log p(\tilde{X}, \tilde{Z}|\theta) \underbrace{- \mathbb{E}_{\tilde{Z}\sim q^{t+1}(\tilde{Z})} \log q^{t+1}(\tilde{Z})}_{H(q^{t+1})}\\
&= \arg\max_\theta \; \log p(X, Z|\theta) + \mathbb{E}_{\tilde{Z}\sim q^{t+1}(\tilde{Z})} \log p(\tilde{X}, \tilde{Z}|\theta). && (8)
\end{aligned}$$

Here we were able to drop the entropy of $q^{t+1}(\tilde{Z})$ denoted by $H(q^{t+1})$, because q^{t+1} depends only on θ^t which is considered fixed when optimizing for θ. Both steps are iterated until convergence.

Remark Comparing the semi-supervised EM variant to the classical unsupervised one leads to only minor differences: The E-Step is practically the same, as both construct q over $\tilde{Z}$. The difference in the M-Step is that the classical variant misses the additive term $\log p(x, z|\theta)$ in Equation 8.

Remark EM assumes we can compute

$$q^{t+1}(\tilde{Z}) = p(\tilde{Z}|\tilde{X}, \theta^t).$$

If $p(\tilde{Z}|\tilde{X}, \theta^t)$ is intractable, we can resort to *variational EM*, that uses an arbitrary distribution q that is tractable. The E-step then becomes an optimization problem on its own.

3.1 Simplifications by Fully Factorized q

Here we want to focus on how $\mathcal{F}(q, \theta)$ simplifies in the E-step and M-step of EM algorithm when we assuming an LVM model and a fully factorized $q(\tilde{Z})$. The joint distribution in the LVM model, given parameters θ, factors as follows

$$p(X, Z, \tilde{X}, \tilde{Z}|\theta) = \prod_{i=1}^{N} p(X_i, Z_i|\theta) \prod_{j=1}^{M} p(\tilde{X}_j, \tilde{Z}_j|\theta).$$

In addition we assume that $q(\tilde{Z})$ factors as

$$q^{t+1}(\tilde{Z}) = \prod_{j=1}^{M} q_j^{t+1}(\tilde{Z}_j).$$

To avoid clutter we drop the time index from q in subsequent steps.

3.1.1 M-Step

By Equation 8 the M-Step optimizes

$$\theta^{t+1} = \arg\max_{\theta} \log p(X, Z|\theta) + \mathbb{E}_{\tilde{Z} \sim q(\tilde{Z})} \log p(\tilde{X}, \tilde{Z}|\theta).$$

For convenience in the following derivations, we rewrite the above equation using two helper functions

$$\theta^{t+1} = \arg\max_{\theta} A(\theta) + B(\theta).$$

Next, we study the factorization of each function separately. The factorization of function $A(\theta)$ is trivial as it does not involve q

$$A(\theta) = \log p(X, Z|\theta) = \sum_{i=1}^{N} [\log p(Z_i|\theta) + \log p(X_i|Z_i, \theta)] \tag{9}$$

The simplification of $B(\theta)$ is more involved

$$
\begin{aligned}
B(\theta) &= \mathbb{E}_{\tilde{Z}\sim q(\tilde{Z})} \log p(\tilde{X},\tilde{Z}|\theta) \\
&= \int_{\tilde{Z}} q(\tilde{Z}) \log p(\tilde{X},\tilde{Z}|\theta)\, d\tilde{Z} \\
&= \int_{\tilde{Z}_1} \cdots \int_{\tilde{Z}_M} \prod_{j=1}^{M} q_j(\tilde{Z}_j) \log \prod_{j=1}^{M} p(\tilde{X}_j,\tilde{Z}_j|\theta)\, d\tilde{Z}_M \ldots d\tilde{Z}_1 \\
&= \int_{\tilde{Z}_1} \cdots \int_{\tilde{Z}_M} \prod_{j=1}^{M} q_j(\tilde{Z}_j) \sum_{j=1}^{M} \log p(\tilde{X}_j,\tilde{Z}_j|\theta)\, d\tilde{Z}_M \ldots d\tilde{Z}_1 \\
&= \int_{\tilde{Z}_1} \cdots \int_{\tilde{Z}_M} \sum_{j=1}^{M} \left(\prod_{k=1}^{M} q_k(\tilde{Z}_k)\right) \log p(\tilde{X}_j,\tilde{Z}_j|\theta)\, d\tilde{Z}_M \ldots d\tilde{Z}_1 \\
&= \int_{\tilde{Z}_1} \cdots \int_{\tilde{Z}_M} \Big[q_1(\tilde{Z}_1) q_2(\tilde{Z}_2) \cdots q_M(\tilde{Z}_M) \log p(\tilde{X}_1,\tilde{Z}_1|\theta) + \cdots \\
&\qquad + q_1(\tilde{Z}_1) q_2(\tilde{Z}_2) \cdots q_M(\tilde{Z}_M) \log p(\tilde{X}_M,\tilde{Z}_M|\theta) \Big]\, d\tilde{Z}_M \ldots d\tilde{Z}_1 && \text{eXpand} \\
&= \int_{\tilde{Z}_1} \cdots \int_{\tilde{Z}_{M-1}} \Big[\int_{\tilde{Z}_M} q_1(\tilde{Z}_1) q_2(\tilde{Z}_2) \cdots q_M(\tilde{Z}_M) \log p(\tilde{X}_1,\tilde{Z}_1|\theta)\, d\tilde{Z}_M + \cdots \\
&\qquad + \int_{\tilde{Z}_M} q_1(\tilde{Z}_1) q_2(\tilde{Z}_2) \cdots q_M(\tilde{Z}_M) \log p(\tilde{X}_M,\tilde{Z}_M|\theta)\, d\tilde{Z}_M \Big]\, d\tilde{Z}_{M-1} \ldots d\tilde{Z}_1 && \text{distr. integra} \\
&= \int_{\tilde{Z}_1} \cdots \int_{\tilde{Z}_{M-1}} \Bigg[q_1(\tilde{Z}_1) q_2(\tilde{Z}_2) \cdots q_{M-1}(\tilde{Z}_{M-1}) \log p(\tilde{X}_1,\tilde{Z}_1|\theta) \underbrace{\int_{\tilde{Z}_M} q_M(\tilde{Z}_M)\, d\tilde{Z}_M}_{=1} + \cdots \\
&\qquad + q_1(\tilde{Z}_1) q_2(\tilde{Z}_2) \cdots q_{M-1}(\tilde{Z}_{M-1}) \int_{\tilde{Z}_M} q_M(\tilde{Z}_M) \log p(\tilde{X}_M,\tilde{Z}_M|\theta)\, d\tilde{Z}_M \Bigg]\, d\tilde{Z}_{M-1} \ldots d\tilde{Z}_1 \\
&= \int_{\tilde{Z}_1} \cdots \int_{\tilde{Z}_{M-1}} \prod_{j=1}^{M-1} q_j(\tilde{Z}_j) \left[\sum_{j=1}^{M-1} \log p(\tilde{X}_j,\tilde{Z}_j|\theta) + \int_{\tilde{Z}_M} q_M(\tilde{Z}_M) \log p(\tilde{X}_M,\tilde{Z}_M|\theta)\, d\tilde{Z}_M \right] d\tilde{Z}_{M-1} \ldots d\tilde{Z}_1 \\
&= \int_{\tilde{Z}_1} \cdots \int_{\tilde{Z}_{M-1}} \prod_{j=1}^{M-1} q_j(\tilde{Z}_j) \sum_{j=1}^{M-1} \log p(\tilde{X}_j,\tilde{Z}_j|\theta)\, d\tilde{Z}_{M-1} \ldots d\tilde{Z}_1 \\
&\qquad + \int_{\tilde{Z}_M} q_M(\tilde{Z}_M) \log p(\tilde{X}_M,\tilde{Z}_M|\theta)\, d\tilde{Z}_M \underbrace{\int_{\tilde{Z}_1} \cdots \int_{\tilde{Z}_{M-1}} \prod_{j=1}^{M-1} q_j(\tilde{Z}_j)\, d\tilde{Z}_{M-1} \ldots d\tilde{Z}_1}_{=1} \\
&= \int_{\tilde{Z}_1} \cdots \int_{\tilde{Z}_{M-1}} \prod_{j=1}^{M-1} q_j(\tilde{Z}_j) \sum_{j=1}^{M-1} \log p(\tilde{X}_j,\tilde{Z}_j|\theta)\, d\tilde{Z}_{M-1} \ldots d\tilde{Z}_1 \\
&\qquad + \int_{\tilde{Z}_M} q_M(\tilde{Z}_M) \log p(\tilde{X}_M,\tilde{Z}_M|\theta)\, d\tilde{Z}_M. && (10)
\end{aligned}
$$

Equation 10 tells us that the Mth term can be separated from the remaining $M-1$ terms. Repeating the same argument iteratively leads to

$$B(\theta) = \mathbb{E}_{\tilde{Z}\sim q(\tilde{Z})} \log p(\tilde{X}, \tilde{Z}|\theta) = \sum_{j=1}^{M} \int_{\tilde{Z}_j} q_j(\tilde{Z}_j) \log p(\tilde{X}_j, \tilde{Z}_j|\theta)\, d\tilde{Z}_j. \tag{11}$$

Finally, we can solve for θ^{t+1} by setting the gradient to zero

$$\nabla_\theta \mathcal{F}(q^{t+1}, \theta) = \mathbf{0}. \tag{12}$$

4 Gaussian Mixture Model

In this section we apply the EM to learn a finite Gaussian Mixture Model (GMM) in the semi-supervised setting. The repository `https://github.com/cheind/semi-supervised-em` contains exemplary source code.

The generative process of a GMM with K is

$$Z|\theta \sim Cat(\alpha_1 \ldots \alpha_K) \tag{13}$$

$$X|Z, \theta \sim N(\mu_z, \sigma_z^2). \tag{14}$$

From Figure 1 we see that the above equations are the same for random variables $\tilde{X}, \tilde{Z}$. The parameters θ of the model are

$$\theta = \left\{\alpha_1 \ldots \alpha_K, \mu_1 \ldots \mu_K, \sigma_1^2 \ldots \sigma_K^2\right\}. \tag{15}$$

4.1 E-Step

By Equation 6 we seek to find

$$q^{t+1}(\tilde{Z}) = p(\tilde{Z}|\tilde{X}, \theta^t).$$

Using the factoring assumptions of q, it suffices to find

$$q_j^{t+1}(\tilde{Z}_j) = p(\tilde{Z}_j|\tilde{X}_j, \theta^t).$$

Recall the joint distribution of partial observed random variables according to our model

$$p(\tilde{Z}_j, \tilde{X}_j|\theta^t) = p(\tilde{Z}_j|\theta^t)p(\tilde{X}_j|\tilde{Z}_j, \theta^t).$$

By Bayes rule

$$p(\tilde{Z}_j|\tilde{X}_j, \theta^t) = \frac{p(\tilde{Z}_j|\theta^t)p(\tilde{X}_j|\tilde{Z}_j, \theta^t)}{p(\tilde{X}_j|\theta^t)} \tag{16}$$

$$= \frac{p(\tilde{Z}_j|\theta^t)p(\tilde{X}_j|\tilde{Z}_j, \theta^t)}{\sum_{k=1}^{K} p(\tilde{Z}_k = k|\theta^t)p(\tilde{X}_j|\tilde{Z}_k = k, \theta^t)} \tag{17}$$

$$= q^{t+1}(\tilde{Z}_j|\tilde{X}_j, \theta^t). \tag{18}$$

$p(\tilde{Z}_j|\tilde{X}_j, \theta^t)$ is called the responsibility and corresponds to soft-assignments of data to components given the current parameter estimates.

4.2 M-Step

In the M-Step we solve for θ^{t+1} by Equation 12

$$\nabla_\theta \mathcal{F}(q^{t+1}, \theta) = \mathbf{0}.$$

We consider parameters in the following order μ_k, σ_k^2 and finally α_k.

4.2.1 M-Step with respect to μ_k

Recalling Equation 8, substituting results from Equation 9, Equation 11 and taking partial derivative with respect to μ_k and setting to zero, gives

$$\begin{aligned}
\frac{\partial \mathcal{F}(q^{t+1}, \theta)}{\partial \mu_k} &= \frac{\partial}{\partial \mu_k} \sum_{i=1}^{N} \log p(X_i{=}x_i, Z_i{=}z_i|\theta) + \frac{\partial}{\partial \mu_k} \sum_{j=1}^{M} \sum_{z=1}^{K} q_j(\tilde{Z}_j{=}z) \log p(\tilde{X}_j{=}\tilde{x}_j, \tilde{Z}_j{=}z|\theta) \\
&= \frac{\partial}{\partial \mu_k} \sum_{i=1}^{N} \delta_{z_i k} \log p(X_i{=}x_i|Z_i{=}k, \theta) + \frac{\partial}{\partial \mu_k} \sum_{j=1}^{M} q_j(\tilde{Z}_j{=}k) \log p(\tilde{X}_j{=}\tilde{x}_j|\tilde{Z}_j{=}k, \theta) \qquad (19) \\
&= \sum_{i=1}^{N} \delta_{z_i k} \frac{\partial}{\partial \mu_k} \left[-\log\left(\sqrt{2\pi\sigma_k^2}\right) - \frac{1}{2\sigma_k^2}(x_i - \mu_k)^2 \right] \\
&\quad + \sum_{j=1}^{M} q_j(\tilde{Z}_j{=}k) \frac{\partial}{\partial \mu_k} \left[-\log\left(\sqrt{2\pi\sigma_k^2}\right) - \frac{1}{2\sigma_k^2}(\tilde{x}_j - \mu_k)^2 \right] \qquad (20) \\
&= \frac{1}{\sigma_k^2} \sum_{i=1}^{N} \delta_{z_i k}(x_i - \mu_k) + \frac{1}{\sigma_k^2} \sum_{j=1}^{M} q_j(\tilde{Z}_j{=}k)(\tilde{x}_j - \mu_k) = 0 \\
\Leftrightarrow \mu_k &= \frac{\sum_{i=1}^{N} \delta_{z_i k} x_i + \sum_{j=1}^{M} q_j(\tilde{Z}_j{=}k)\tilde{x}_j}{\sum_{i=1}^{N} \delta_{z_i k} + \sum_{j=1}^{M} q_j(\tilde{Z}_j{=}k)}. \qquad (21)
\end{aligned}$$

Equation 19 made use of properties of log and dropping terms that does not depend on μ_k. In Equation 20 probabilities are substituted for concrete GMM densities.

4.2.2 M-Step with respect to σ_k^2

The derivation is similar as for μ_k. Starting from Equation 20 we have

$$
\begin{aligned}
\frac{\partial \mathcal{F}(q^{t+1},\theta)}{\partial \sigma_k^2} &= \sum_{i=1}^{N} \delta_{z_i k} \frac{\partial}{\partial \mu_k}\left[-\log\left(\sqrt{2\pi\sigma_k^2}\right) - \frac{1}{2\sigma_k^2}(x_i-\mu_k)^2\right] \\
&\quad + \sum_{j=1}^{M} q_j(\tilde{Z}_j{=}k)\frac{\partial}{\partial \mu_k}\left[-\log\left(\sqrt{2\pi\sigma_k^2}\right) - \frac{1}{2\sigma_k^2}(\tilde{x}_j-\mu_k)^2\right] \\
&= \sum_{i=1}^{N} \delta_{z_i k} \frac{\partial}{\partial \mu_k}\left[-\frac{1}{2}\log(2\pi\sigma_k^2) - \frac{1}{2\sigma_k^2}(x_i-\mu_k)^2\right] \\
&\quad + \sum_{j=1}^{M} q_j(\tilde{Z}_j{=}k)\frac{\partial}{\partial \mu_k}\left[-\frac{1}{2}\log(2\pi\sigma_k^2) - \frac{1}{2\sigma_k^2}(\tilde{x}_j-\mu_k)^2\right] \\
&= \sum_{i=1}^{N} \delta_{z_i k}\left[-\frac{1}{2\sigma_k^2} + \frac{1}{2\sigma_k^4}(x_i-\mu_k)^2\right] + \sum_{j=1}^{M} q_j(\tilde{Z}_j{=}k)\left[-\frac{1}{2\sigma_k^2} + \frac{1}{2\sigma_k^4}(\tilde{x}_j-\mu_k)^2\right] \qquad (22) \\
&= \sum_{i=1}^{N} \delta_{z_i k}\left[-\frac{\sigma_k^2}{2\sigma_k^4} + \frac{1}{2\sigma_k^4}(x_i-\mu_k)^2\right] + \sum_{j=1}^{M} q_j(\tilde{Z}_j{=}k)\left[-\frac{\sigma_k^2}{2\sigma_k^4} + \frac{1}{2\sigma_k^4}(\tilde{x}_j-\mu_k)^2\right] \\
&= -\frac{\sigma_k^2}{2\sigma_k^4}\sum_{i=1}^{N}\delta_{z_i k} + \frac{1}{2\sigma_k^4}\sum_{i=1}^{N}\delta_{z_i k}(x_i-\mu_k)^2 - \frac{\sigma_k^2}{2\sigma_k^4}\sum_{j=1}^{M} q_j(\tilde{Z}_j{=}k) + \frac{1}{2\sigma_k^4}\sum_{j=1}^{M} q_j(\tilde{Z}_j{=}k)(\tilde{x}_j-\mu_k)^2 \\
&= -\frac{\sigma_k^2}{2\sigma_k^4}\left[\sum_{i=1}^{N}\delta_{z_i k} + \sum_{j=1}^{M} q_j(\tilde{Z}_j{=}k)\right] + \frac{1}{2\sigma_k^4}\left[\sum_{i=1}^{N}\delta_{z_i k}(x_i-\mu_k)^2 + \sum_{j=1}^{M} q_j(\tilde{Z}_j{=}k)(\tilde{x}_j-\mu_k)^2\right] = 0 \\
\Leftrightarrow \sigma_k^2 &= \frac{\sum_{i=1}^{N}\delta_{z_i k}(x_i-\mu_k)^2 + \sum_{j=1}^{M} q_j(\tilde{Z}_j{=}k)(\tilde{x}_j-\mu_k)^2}{\sum_{i=1}^{N}\delta_{z_i k} + \sum_{j=1}^{M} q_j(\tilde{Z}_j{=}k)}. \qquad (23)
\end{aligned}
$$

In Equation 22 the fourth-power appears because $r=\sigma^2$, $\frac{\partial}{\partial r}\frac{1}{r} = -\frac{1}{r^2} = -\frac{1}{\sigma^4}$.

4.2.3 M-Step with respect to α_k

When optimizing α_k we need to take into account the constraint that $\alpha_1 \dots \alpha_K$ needs to be a valid probability mass function and thus needs to sum to one. By the method of Lagrangian multipliers we have

$$\mathcal{L}(q^{t+1},\theta,\lambda) = \mathcal{F}(q^{t+1},\theta) + \lambda\left(\sum_{z=1}^{K}\alpha_z - 1\right),$$

and thus

$$\nabla_{\{\alpha_k,\lambda\}}\mathcal{L}(q^{t+1},\theta,\lambda) = \mathbf{0}.$$

Note that the PMF of the categorical distribution is given by

$$p(Z=z|\theta) = \prod_{i=1}^{K}\alpha_i^{\delta_{zi}}. \qquad (24)$$

Then, the partial derivative with respect to α_k is given by

$$\begin{aligned}
\frac{\partial \mathcal{L}(q^{t+1}, \theta, \lambda)}{\partial \alpha_k} &= \frac{\partial}{\partial \alpha_k}\left[\mathcal{F}(q^{t+1}, \theta) + \lambda\left(\sum_{z=1}^{K} \alpha_z - 1\right)\right] \\
&= \frac{\partial}{\partial \alpha_k} \sum_{i=1}^{N} \log p(X_i = x_i, Z_i = z_i | \theta) \\
&\quad + \frac{\partial}{\partial \alpha_k} \sum_{j=1}^{M} \sum_{z=1}^{K} q_j(\tilde{Z}_j = z) \log p(\tilde{X}_j = \tilde{x}_j, \tilde{Z}_j = z | \theta) \\
&\quad + \frac{\partial}{\partial \alpha_k} \lambda \left(\sum_{z=1}^{K} \alpha_z - 1\right) \\
&= \frac{\partial}{\partial \alpha_k} \sum_{i=1}^{N} \delta_{z_i k} \log p(Z_i = k | \theta) \\
&\quad + \frac{\partial}{\partial \alpha_k} \sum_{j=1}^{M} q_j(\tilde{Z}_j = k) \log p(\tilde{Z}_j = k | \theta) \\
&\quad + \lambda \\
&= \frac{\partial}{\partial \alpha_k} \sum_{i=1}^{N} \delta_{z_i k} \log \prod_{l=1}^{K} \alpha_l^{\delta_{kl}} \\
&\quad + \frac{\partial}{\partial \alpha_k} \sum_{j=1}^{M} q_j(\tilde{Z}_j = k) \log \prod_{l=1}^{K} \alpha_l^{\delta_{kl}} \\
&\quad + \lambda \\
&= \sum_{i=1}^{N} \delta_{z_i k} \frac{\partial}{\partial \alpha_k} \log \alpha_k + \sum_{j=1}^{M} q_j(\tilde{Z}_j = k) \frac{\partial}{\partial \alpha_k} \log \alpha_k + \lambda \\
&= \frac{1}{\alpha_k} \sum_{i=1}^{N} \delta_{z_i k} + \frac{1}{\alpha_k} \sum_{j=1}^{M} q_j(\tilde{Z}_j = k) + \lambda = 0 \\
&= \frac{1}{\alpha_k} \left[\sum_{i=1}^{N} \delta_{z_i k} + \sum_{j=1}^{M} q_j(\tilde{Z}_j = k)\right] + \lambda = 0 \\
&\Leftrightarrow \alpha_k = -\frac{\sum_{i=1}^{N} \delta_{z_i k} + \sum_{j=1}^{M} q_j(\tilde{Z}_j = k)}{\lambda} = -\frac{N_k}{\lambda}.
\end{aligned} \tag{25}$$

Equating our original constraint $\sum_{z=1}^{K} \alpha_z = 1$ together with $\sum_{z=1}^{K} -\frac{N_z}{\lambda} = 1$ we solve for λ as follows

$$
\begin{aligned}
-\sum_{z=1}^{K} \frac{N_z}{\lambda} &= 1 \\
-\frac{1}{\lambda} \sum_{z=1}^{K} N_z &= 1 \\
\lambda &= -\sum_{z=1}^{K} N_z. \qquad (26)
\end{aligned}
$$

Plugging Equation 26 into Equation 25 yields

$$\alpha_k = \frac{N_k}{\sum_{z=1}^{K} N_z}. \qquad (27)$$

References

[1] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.