Variational EM-Marginal Likelihood with MCMC Notes

Tony

March 2022

1 Introduction

In these notes I write about the Variational Expectation Maximization/Empirical Bayes/Type II maximum likelihood method. I start this post by writing about the Expectation Maximization method (Dempster et al., 1977) and from there I will move on to the Variational method[1].

2 Derivation of Expectation Maximization

Expectation Maximization (EM) is the most known algorithm for iteratively optimizing the Gaussian Mixture Model method. It finds the Maximum Likelihood Estimates of model variables in the presence of latent variables in this model. Here we provide the derivation of the EM algorithm and furthermore the intuition behind the equations[2]. More details on the intuition are provided in Section 3.

In general the goal is to maximize the log-likelihood $l(x) = \log p(x|\theta) = \log \sum_z p(x, z|\theta)$ via gradient descent updates. There are times where we do not want to optimize a method via using gradients for various reasons. In these cases it is useful to use other routines like EM. Let us provide the derivation of ELBO from log-likelihood as follows:

$$\begin{aligned} l(x) = \log p(x|\theta) &= \log \sum_z p(x, z|\theta) \\ &= \log \sum_z q(z|x,\theta) \frac{p(x,z|\theta)}{q(z|x,\theta)} \\ &\geq \sum_z q(z|x,\theta) \log \frac{p(x,z|\theta)}{q(z|x,\theta)} \equiv ELBO(q,\theta) \end{aligned} \tag{1}$$

where $q(z|x, \theta)$ is a distribution of latent variables z. What we are doing here is that instead of directly maximizing the log-likelihood $l(x)$, EM is maximizing the $ELBO(q, \theta)$ with coordinate ascent with the E and M steps, where:

[1] Part of these notes was inspired by these lecture notes.

[2] The derivation was taken from these lecture notes, and these ones.

$$\textbf{E-Step: } q_{(i+1)} = \arg\max_{q} ELBO(q, \theta_{(i)})$$
$$\textbf{M-Step: } \theta_{(i+1)} = \arg\max_{\theta} ELBO(q_{(i+1)}, \theta)$$

We iterate between E and M steps. It can been observed that in E-step above we have a fixed θ and we optimized in the space of distributions q, which might be time-consuming and difficult, since the space of distributions q is very vast. Therefore we go ahead and show that $q_{(i+1)} = p(z|z, \theta_{(i)})$. Basically we need to mazimize $ELBO(q, \theta)$, which is equivalent to maximizing $\sum_z q(z|x,\theta) \log p(x, z|\theta) - \sum_z q(z|x,\theta) \log q(z|x,\theta)$. The solution of that is $q(z|x,\theta) = p(z|x,\theta)$. Let us restate the maximization problem of $\sum_z q(z|x,\theta) \log p(x, z|\theta) - \sum_z q(z|x,\theta) \log q(z|x,\theta)$ as follows:

$$\begin{aligned} &\max_{q(z|x,\theta)} \sum_z q(z|x,\theta) \log p(x,z|\theta) - \sum_z q(z|x,\theta) \log q(z|x,\theta) \\ &\text{s.t. } q(z|x,\theta) \geq 0, \sum_z q(z|x,\theta) = 1 \end{aligned}$$

We solve this with the use of Lagrangian mutlipliers and we have:

$$\mathcal{L} = \sum_z q(z|x,\theta) \log p(x,z|\theta) - \sum_z q(z|x,\theta) \log q(z|x,\theta) - \sum_z \lambda_z q(z|x,\theta) - \nu(1 - \sum_z q(z|x,\theta)) \quad (2)$$

We take the derivative of Eq. (2) and we have:

$$\frac{d\mathcal{L}}{dq(z|x,\theta)} = \log p(x,z|\theta) - \log q(z|x,\theta) - 1 - \lambda_z - \nu = 0 \quad (3)$$

We take the derivative of Eq. (2) w.r.t nu and λ_z and we get equations similar to Eq. (3). We solve the systemf of equations and eventually we end up with:

$$q(z|x,\theta) \propto p(x,z|\theta) \quad (4)$$

which is true and at the same time $q(z|x,\theta)$ is a normalized distribution if $q(z|x,\theta) = p(z|x,\theta)$. So we have proved in the expectation step that we maximize the expectation w.r.t $q(z|x,\theta)$ if $q(z|x,\theta) = p(z|x,\theta)$.

We continue from Eq. (1):

$$\begin{aligned} l(x) &\geq ELBO(q,\theta) \\ &= \sum_z q(z|x,\theta) \log \frac{p(x,z|\theta)}{q(z|x,\theta)} \\ &= \sum_z q(z|x,\theta) \log p(x,z|\theta) - \sum_z q(z|x,\theta) \log q(z|x,\theta) \\ &= Q(\theta|\theta_{(i)}) + H(q) \end{aligned}$$

where we have shown that maximizing $ELBO(q,\theta)$ is the same as maximizing the expectation of the log-likelihood $\sum_z q(z|x,\theta) \log p(x,z|\theta)$, where the maximum of that is when $q(z|x,\theta) = p(z|x,\theta)$, so the maximum expectation is then $\sum_z p(z|x,\theta)\log p(x,z|\theta)$. There we rewrite the E and M steps as follows:

$$\begin{aligned}\textbf{E-Step: } & Q(\theta,\theta_{(i)}) = \mathbb{E}_{p(z|x,\theta_{(i)})}[\log p(x,z|\theta)]\\ \textbf{M-Step: } & \theta_{(i+1)} = \arg\max_{\theta}\mathbb{E}_{p(z|x,\theta_{(i)})}[\log p(x,z|\theta)]\end{aligned}$$

This was more abstract so in the next sections we provide more specific examples for the EM algorithm.

3 Expectation Maximization for Gaussian Mixture Model

We start by briefly presenting Gaussian Mixture model (GMM). Gaussian Mixture Model is finite mixture model where each of the K components is a Guassian density with parameters mean μ_k and covariance matrix Σ_k. Then the Gaussian Mixture model is defined as:

$$p(x_i|\Theta) = \sum_{k=1}^{K} w_k p_k(x_i|z_{ik}=1,\theta_k) \tag{5}$$

where $p_k(x_i|\theta_k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}}\exp\left[-\frac{1}{2}(x_i-\mu_k)^\intercal\Sigma_k^{-1}(x_i-\mu_k)\right]$ is multivariate Gaussian density and a component of the Gaussian Mixture model, and w_k is the weight of component k, which is the probability that an instance x was generated by that component k.

In order to train a Gaussian Mixture Model we use the maximum likelihood. We define the likelihood as the following:

$$l(\theta) = P(D|\Theta) = \sum_{i=1}^{N}\log\left(\sum_{k=1}^{K} w_k p(x_i|z_{ik}=1,\theta_k)\right) \tag{6}$$

where D is that data that we are using.

In general the Gaussian Mixture model assumes the data instances were generated from a mixture of a finite number of Gaussian distributions, of which distributions we don't know the parameters. On the one hand, Guassian Mixture model is the fastest algorithm to learn mixture models, and since it maximizes only the likelihood, it does not have any biases for the mean of the distributions or the structure sizes. On the other hand when there exists a lot of data the estimation of the covariance matrices of the distributions might become intractable and end up in a singular solution (GMM is notorious for diverging and finding solution with infinite likelihood). Moreover, GMM uses all the components that we have defined, and some of them might be redundant, which introduces the need of statistical criteria in order to find the right number of components[3]

Going back to Eq. (6), the logical step to solve this equation is to take partial derivatives of the log of the likelihood with respect to all the model parameters, set them to 0 and apply gradient descent to these equations. For that we need to set hyperparameters and some constraints that the sum of the weights a_k is equal to 0. Here we circle back to an easier solution for training GMMs that we mentioned earlier, the EM algorithm. EM iterates between between two steps, the Expectation step

[3] More information on the GMM can be found in this scikit-learn page.

(E-step) and the Maximization step (M-step). The **E-step** finds the probabilities of different data X given the observed data y (targets) and the parameter estimates θ. The log-likelihoods are also computed on this step. More formally the E-step computes the expected log-likelihood with respect to the probability of the hidden variables given the data X and fixed values of θs. **M-step** tries to maximize the expectation of the E-step, where it finds new parameter values for the θs that give the maximum expected log-likelihoods.

3.1 Expectation Step (E-step)

We write again the full likelihood:

$$L(x, z|\theta) = \prod_{i=1}^{N}\prod_{k=1}^{K} \left[w_k p_k(x_i|\theta_k)\right]^{\mathbb{I}(z_i=k)} \tag{7}$$

and we take the log of Eq. (7) in order to find the log-likelihood:

$$\log L(x, z|\theta) = \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{I}(z_i = k) \log\left[w_k p_k(x_i|\theta_k)\right] \tag{8}$$

Let θ be the unknown parameters of the GMM and θ_n be the estimates of these parameters from the last iteration. We define the expectation for the E-step:

$$Q(\theta, \theta_n) = \mathbb{E}_{Z|X,\theta_n}\left[\log L(x, z|\theta)\right] \tag{9}$$

$$= \mathbb{E}_{Z|X,\theta_n}\left[\sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{I}(z_i = k) \log\left[w_k p_k(x_i|\theta_k)\right]\right] \tag{10}$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{E}_{Z|X,\theta_n}\left[\mathbb{I}(z_i = k)\right] \log\left[w_k p_k(x_i|\theta_k)\right] \tag{11}$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} p(z_i = k|x_i, \theta_n) \log\left[w_k p_k(x_i|\theta_k)\right] \tag{12}$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} \frac{p(x_i|z_i = k, \theta_n)p(z_i = k)}{\sum_{l=1}^{K} p(x_i|z_i = l, \theta_n)p(z_i = l)} \log\left[w_k p_k(x_i|\theta_k)\right] \tag{13}$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} \frac{p(x_i|z_i = k, \theta_n)w_k}{\sum_{l=1}^{K} p(x_i|z_i = l, \theta_n)w_l} \log\left[w_k p_k(x_i|\theta_k)\right] \tag{14}$$

$$\tag{15}$$

On the above equation between the last and the before the last equation we use the Bayes rule. We set $w_{ik} = \frac{p(x_i|z_i=k,\theta_n)w_k}{\sum_{l=1}^{K} p(x_i|z_i=l,\theta_n)w_l}$ and we finally get the expectation that we are looking to maximize:

$$Q(\theta, \theta_n) = \sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik} \log\left[w_k p_k(x_i|\theta_k)\right] \tag{16}$$

In the E-step for GMMs, we need to find the ”membership ”weights” w_{ik} in order to compute the expectation for the M-step. The way that we compute the ”membership ”weights” w_{ik} in GMM is by computing the cluster weight times the probability of the cluster that it is assigned, divided by the sum of all the cluster weights times their probabilities. This fraction makes up each point’s weight for its assigned cluster.

3.2 Maximization Step (M-step)

In the M-step we find the parameter values for θ that maximize the expectation $Q(\theta, \theta_n)$ from the E-step. Basically we find:

$$\theta_{n+1} = \arg\max_{\theta} Q(\theta, \theta_n) \tag{17}$$

This can be solved by obtaining the MLE of the parameters θ. If we expand Eq. (16) we get:

$$Q(\theta, \theta_n) = \sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik}\left[\log w_k - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x_i - \mu_k)^{\intercal}\Sigma_k^{-1}(x_i - \mu_k) - \frac{d}{2}\log(2\pi)\right] \tag{18}$$

We observe that θs have closed-form solutions since Eq. (18) appears to be in the form of a weighted MLE for a normal distribution. w_{ik} and μ/Σ appear in separate linear terms so they can be maximized independently. We find the MLE for every parameter θ, w_k, μ and Σ and we get the following:

$$w_k = \frac{\sum_{i=1}^{n} w_{ik}}{n} \tag{19}$$

$$\mu_k = \frac{\sum_{i=1}^{n} w_{ik}x_i}{\sum_{i=1}^{n} w_{ik}} \tag{20}$$

$$\Sigma_k = \frac{\sum_{i=1}^{n} w_{ik}(x_i - \mu_k)(x_i - \mu_k)^{\intercal}}{\sum_{i=1}^{n} w_{ik}} \tag{21}$$

We can also think of EM for GMM in a more intuitive way. For instance The weight of each cluster is the sum of the weights of the data points assigned to it, divided by the number of all data points. In the nominator we use the weights of the data points, rather than just the number of points because every data point has a probability of being assigned in a specific cluster. Same goes for the mean and the variance. It is similar to computing the empirical average and variance but here we multiply and divide by the assigned weights.

We terminate the iterative EM algorithm by detecting convergence. The way to do it is when the value of the log-likelihood or the average of the ”membership weights” changes by less than a small threshold.

4 Expectation Maximization for Exponential families

The probability density of an exponential family is the following:

$$\begin{aligned} p_\theta(x) &= h(x)\exp(\theta^{\intercal}T(x) - A(\theta)) \\ &= h(x)\exp(-A(\theta))\exp(\theta^{\intercal}T(x)) \end{aligned}$$

where we set $c(\theta) = \exp(-A(\theta)) \Rightarrow A(\theta) = -\log c(\theta)$, and therefore we have $h(x)c(\theta)\exp(\theta^\intercal T(x))$. In general $A(\theta)$ is called the log-partition function since it is the logarithm of a normalization factor. If not, then $p(x)$ would not be a probability distribution. $A(\theta)$ is defined as follows:

$$A(\theta) = \log \underbrace{\left[\int h(x)\exp(\theta^\intercal T(x))dx\right]}_{Q(\theta)} \tag{22}$$

The first derivative of $A(\theta)$ is the following:

$$\begin{aligned}
\frac{dA(\theta)}{d\theta} &= \frac{1}{Q(\theta)}\frac{dQ(\theta)}{d\theta} = \frac{Q'\theta}{Q(\theta)} \\
&= \frac{\int h(x)\exp(\theta^\intercal T(x))T(x)dx}{\int h(x)\exp(\theta^\intercal T(x))dx} \\
&= \frac{\int h(x)\exp(\theta^\intercal T(x))T(x)dx}{\exp A(\theta)} \\
&= \int h(x)\exp(\theta^\intercal T(x) - A(\theta))T(x)dx \\
&= \int p_\theta(x)T(x)dx \\
&= \mathbb{E}_\theta[T(x)] \qquad (23) \\
&\Rightarrow \mathbb{E}_\theta[T(x)] = \frac{-d\log c(\theta)}{d\theta}
\end{aligned}$$

The second derivative of $A(\theta)$ is the following:

$$\begin{aligned}
\frac{d^2A(\theta)}{d\theta^2} &= \frac{d}{d\theta}\left[\frac{Q'(\theta)}{Q(\theta)}\right] = \frac{d}{d\theta}\left[Q'(\theta)\frac{1}{Q(\theta)}\right] = \frac{Q''(\theta)}{Q(\theta)} - \frac{(Q')^2}{(Q(\theta))^2} \\
&= \frac{\int h(x)\exp(\theta^\intercal T(x))T^2(x)dx}{\int h(x)\exp(\theta^\intercal T(x))dx} - \left(\frac{Q'(\theta)}{Q(\theta)}\right)^2 && \text{from Eq. (23)} \\
&= \frac{\int h(x)\exp(\theta^\intercal T(x))T^2(x)dx}{\int h(x)\exp(\theta^\intercal T(x))dx} - (\mathbb{E}_\theta[T(x)])^2 \\
&= \mathbb{E}_\theta[T^2(x)] - (\mathbb{E}_\theta[T(x)])^2 = \operatorname{cov}_{p_\theta}[T(\theta)]
\end{aligned}$$

The log-likelihood of $p_\theta(x)$ is the following:

$$l(x,\theta) = \log c(\theta) + \theta^\intercal T(x) + \text{const} \tag{24}$$

We then have Y as observed and Z as unobserved variables such as $X = (Y, Z)$. The expectation of the log-likelihood is then:

$$\begin{aligned}
Q(\theta|\theta_{(i)}) &= \mathbb{E}[l(x,\theta|\theta_{(i)})] \\
&= \mathbb{E}[\log c(\theta) + \theta^\intercal T(x) + \text{const}] \\
&= \mathbb{E}[\log c(\theta)] + \mathbb{E}[\theta^\intercal T(x)] \\
&= \log c(\theta) + \theta^\intercal \mathbb{E}[T(x)]
\end{aligned}$$

where $T(y) = \mathbb{E}[T(y,z|y)]$. So we have the expectation as follows:

$$Q(\theta|\theta_{(i)}) = \log c(\theta) + \theta^\intercal T(y) \tag{25}$$

We maximize expectation in Eq. (25) with differentiation w.r.t θ and then we equalize it to 0. We also use from Eq. (23) that $\mathbb{E}_\theta[T(x)] = \frac{-d\log c(\theta)}{d\theta}$ and therefore we have:

$$\begin{aligned}
\frac{dQ(\theta|\theta_{(i)})}{d\theta} &= \frac{d\log c(\theta)}{d\theta} + T(y) \\
&= -\mathbb{E}_\theta[T(x)] + \mathbb{E}[T(y,z|y)] \\
&= 0 \\
&\Rightarrow \mathbb{E}[T(y,z|y)] = -\mathbb{E}_{\theta-(i)}[T(x)]
\end{aligned}$$

We note that $\mathbb{E}[T(y,z|y)]$ is the expectation of the incomplete data and $\mathbb{E}_\theta[T(x)]$ is the expectation of the complete data.

5 Expectation Maximization in Bayesian estimation

The EM algorithm can also be used to find a maximum a posteriori (MAP) estimate in a Bayesian setup. Let $f(\theta)$ be a prior density function, $f(\theta|y)$ be the posterior function of the observed data y and $f(\theta|x)$ be the posterior function of the complete data x. The goal is to find the $f(\theta|y)$ for the unobserved data. Then using Bayes rule we have that $f(\theta|y) \propto f(y|\theta)f(\theta)$. The log of the posterior is then $\log f(\theta|y) = \log f(y|\theta) + \log f(\theta)$. So we find the MAP value as follows:

$$\theta_{MAP} = \arg\max_\theta(\log f(y|\theta) + \log f(\theta)) \tag{26}$$

Then we define the EM algorithm.

E-Step. Compute the expectation of the posterior:

$$\mathbb{E}_\theta[\log f(\theta|x)|y] = \mathbb{E}_\theta[\log f(x|\theta)|y] + \log f(\theta) \tag{27}$$

M-step. We maximize Eq. (27) by maximizing $\mathbb{E}[\log f(x|\theta)|y]$ as we normally do in the EM algorithm. We note here again that $\mathbb{E}[\log f(x|\theta)|y]$ is the Expectation of the log-likelihood that we need to maximize.

Basically in the EM algorithm in bayesian setups, we try find the posterior by maximizing the expectation of the log-likelihood (normal EM algorithm) plus the log of the prior.

6 Monte Carlo Expectation Maximization algorithm

Sometimes the expectation in the E-step is difficult to compute. So another way is to approximate the expectation with Monte Carlo methods. This method is called Monte Carlo Expectation Minimization (MCEM) (Levine and Casella, 2001; Wei and Tanner, 1990). We can rewrite the expectation $Q(\theta|\theta_{(i)})$ with $X = (Y, Z)$, where Z are the latent variables (unobserved) and Y the observed as follows:

$$\begin{aligned} Q(\theta|\theta_{(i)}) &= \mathbb{E}_{\theta_{(i)}}[\log f_{X|\Theta}(X|\theta)|Y] && (28) \\ &= \mathbb{E}_{\theta_{(i)}}[\log f_{X|\Theta}((Y,Z)|\theta)|Y] && (29) \\ & && (30) \end{aligned}$$

Then the MCEM algorithm is the following:

E-step. We draw z samples from $f_{Z|Y,\Theta}(z|y,\theta_{(i)})$ and plug in the samples for the latent variable. After that we use the same process as in EM. The corresponding expectation is the following:

$$Q(\theta|\theta_{(i)}) = \frac{1}{M}\sum_{m=1}^{M} \log f_{X|\Theta}((y, z_m)|\theta) \tag{31}$$

M-step. We maximize $Q(\theta|\theta_{(i)})$ in Eq. (31).

[NEED TO ADD MORE DETAILS ON THE MCEM. READ THE MAIN PAPER]

6.1 Stochastic Expectation Maximization algorithm

Stochastic Expectation Maximization is a special case of MCEM in which we use only one sample, $M = 1$.

References

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

Richard A Levine and George Casella. 2001. Implementations of the monte carlo em algorithm. *Journal of Computational and Graphical Statistics* 10(3):422–439.

Greg CG Wei and Martin A Tanner. 1990. A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association* 85(411):699–704.