EM Algorithm for Latent Variable Models

David Rosenberg

New York University

May 2, 2017
Gaussian Mixture Models (Review)
Example: Old Faithful Geyser
Let’s consider a generative model for the data.

Suppose

1. There are \( k \) clusters.
2. We have a probability density for each cluster.

Generate a point \( x \) as follows

1. Choose a random cluster \( z \in \{1, 2, \ldots, k\} \).
2. Choose a point \( x \) from the distribution for cluster \( z \).
Gaussian Mixture Model \((k = 3)\)

1. Choose \(z \in \{1, 2, 3\}\) with \(p(1) = p(2) = p(3) = \frac{1}{3}\).
2. Choose \(x \mid z \sim \mathcal{N}(X \mid \mu_z, \Sigma_z)\).
Gaussian Mixture Model Parameters ($k$ Components)

Cluster probabilities: \( \pi = (\pi_1, \ldots, \pi_k) \)

Cluster means: \( \mu = (\mu_1, \ldots, \mu_k) \)

Cluster covariance matrices: \( \Sigma = (\Sigma_1, \ldots, \Sigma_k) \)

For now, suppose all these parameters are known. We’ll discuss how to learn or estimate them later.
The GMM “Inference” Problem

- *Suppose* we know all the model parameters $\pi, \mu, \Sigma$, and thus $p(x, z)$.

- The **inference problem**: We observe $x$. We want to know its cluster $z$.

- We can get a **soft cluster assignment** from the conditional distribution:

  $$p(z | x) = \frac{p(x, z)}{p(x)}$$

- A **hard cluster assignment** is given by

  $$z^* = \arg \max_{z \in \{1, \ldots, k\}} p(z | x).$$

- So if we know the model parameters, we can compute $p(z | x)$, and clustering is trivial.
The GMM “Learning” Problem

- Given data $x_1, \ldots, x_n$ drawn i.i.d. from a GMM,
- Estimate the parameters:
  
  Cluster probabilities: $\pi = (\pi_1, \ldots, \pi_k)$
  
  Cluster means: $\mu = (\mu_1, \ldots, \mu_k)$
  
  Cluster covariance matrices: $\Sigma = (\Sigma_1, \ldots, \Sigma_k)$

- Traditional approach is maximum [marginal] likelihood:

  $$(\pi, \mu, \Sigma) = \arg\max_{\pi, \mu, \Sigma} p(x_1, \ldots, x_n)$$

- Unfortunately, this is very difficult.

- Note that the fully observed problem is easy. That is:

  $$(\pi, \mu, \Sigma) = \arg\max_{\pi, \mu, \Sigma} p(x_1, \ldots, x_n, z_1, \ldots, z_n)$$
EM Algorithm for Latent Variable Models
EM Algorithm for Latent Variable Models

General Latent Variable Model

- Two sets of random variables: \( z \) and \( x \).
- \( z \) consists of unobserved **hidden variables**.
- \( x \) consists of **observed variables**.
- Joint probability model parameterized by \( \theta \in \Theta \):

\[
p(x, z \mid \theta)
\]

Definition

A **latent variable model** is a probability model for which certain variables are never observed.

e.g. The Gaussian mixture model is a latent variable model.
Suppose we have a data set $\mathcal{D} = (x_1, \ldots, x_n)$. To simplify notation, take $x$ to represent the entire dataset

$$x = (x_1, \ldots, x_n),$$

and $z$ to represent the corresponding unobserved variables

$$z = (z_1, \ldots, z_n).$$

An observation of $x$ is called an **incomplete data set**. An observation $(x, z)$ is called a **complete data set**.
Our Objectives

- **Learning problem**: Given incomplete dataset $\mathcal{D} = x = (x_1, \ldots, x_n)$, find MLE

$$\hat{\theta} = \arg \max_{\theta} p(\mathcal{D} | \theta).$$

- **Inference problem**: Given $x$, find conditional distribution over $z$:

$$p(z_i | x_i, \theta).$$

- For Gaussian mixture model, learning is hard, inference is easy.
- For more complicated models, inference can also be hard. (See DSGA-1005)
Log-Likelihood and Terminology

Note that

$$\arg\max_{\theta} p(x \mid \theta) = \arg\max_{\theta} \log p(x \mid \theta).$$

Often easier to work with this “log-likelihood”.

We often call $p(x)$ the **marginal likelihood**, because it is $p(x, z)$ with $z$ “marginalized out”:

$$p(x) = \sum_{z} p(x, z)$$

We often call $p(x, y)$ the **joint**. (for “joint distribution”)

Similarly, $\log p(x)$ is the **marginal log-likelihood**.
The EM Algorithm **Key Idea**

- Marginal log-likelihood is hard to optimize:
  \[
  \max_\theta \log p(x \mid \theta)
  \]

- **Typically** the complete data log-likelihood is easy to optimize:
  \[
  \max_\theta \log p(x, z \mid \theta)
  \]

- What if we had a **distribution** \( q(z) \) for the latent variables \( z \)?
- Then maximize the **expected complete data log-likelihood**:
  \[
  \max_\theta \sum_z q(z) \log p(x, z \mid \theta)
  \]

- **EM assumes** this maximization is relatively easy.
Let $q(z)$ be any PMF on $\mathcal{Z}$, the support of $z$:

$$
\log p(x \mid \theta) = \log \left[ \sum_z p(x, z \mid \theta) \right] \\
= \log \left[ \sum_z q(z) \left( \frac{p(x, z \mid \theta)}{q(z)} \right) \right] \quad \text{(log of an expectation)} \\
\geq \sum_z q(z) \log \left( \frac{p(x, z \mid \theta)}{q(z)} \right) \quad \text{(expectation of log)}
$$

Inequality is by Jensen’s, by concavity of the log.

This inequality is the basis for “variational methods”, of which EM is a basic example.
The ELBO

- For any PMF $q(z)$, we have a lower bound on the marginal log-likelihood

$$\log p(x | \theta) \geq \sum_z q(z) \log \left( \frac{p(x, z | \theta)}{q(z)} \right)$$

- Marginal log likelihood $\log p(x | \theta)$ also called the evidence.

- $\mathcal{L}(q, \theta)$ is the evidence lower bound, or “ELBO”.

In EM algorithm (and variational methods more generally), we maximize $\mathcal{L}(q, \theta)$ over $q$ and $\theta$. 

David Rosenberg (New York University)
MLE, EM, and the ELBO

- For any PMF \( q(z) \), we have a lower bound on the marginal log-likelihood

\[
\log p(x \mid \theta) \geq \mathcal{L}(q, \theta).
\]

- The MLE is defined as a maximum over \( \theta \):

\[
\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} \log p(x \mid \theta).
\]

- In EM algorithm, we maximize the lower bound (ELBO) over \( \theta \) and \( q \):

\[
\hat{\theta}_{\text{EM}} = \arg \max_{\theta} \left[ \max_{q} \mathcal{L}(q, \theta) \right]
\]
A Family of Lower Bounds

- For each $q$, we get a lower bound function: $\log p(x \mid \theta) \geq \mathcal{L}(q, \theta) \forall \theta$.
- Two lower bounds (blue and green curves), as functions of $\theta$:

![Graph showing two curves, one red and one green, representing lower bounds.]

- Ideally, we’d find the maximum of the red curve. Maximum of green is close.

From Bishop’s *Pattern recognition and machine learning*, Figure 9.14.
Choose sequence of $q$’s and $\theta$’s by “coordinate ascent”.

EM Algorithm (high level):

1. Choose initial $\theta^{\text{old}}$.
2. Let $q^* = \arg\max_q \mathcal{L}(q, \theta^{\text{old}})$
3. Let $\theta^{\text{new}} = \arg\max_{\theta} \mathcal{L}(q^*, \theta^{\text{old}})$.
4. Go to step 2, until converged.

Will show: $p(x \mid \theta^{\text{new}}) \geq p(x \mid \theta^{\text{old}})$

Get sequence of $\theta$’s with monotonically increasing likelihood.
EM Algorithm for Latent Variable Models

EM: Coordinate Ascent on Lower Bound

1. Start at $\theta^{\text{old}}$.
2. Find $q$ giving best lower bound at $\theta^{\text{old}} \Rightarrow \mathcal{L}(q, \theta)$.
3. $\theta^{\text{new}} = \arg\max_{\theta} \mathcal{L}(q, \theta)$.

From Bishop’s *Pattern recognition and machine learning*, Figure 9.14.
We now give 2 different re-expressions of \( \mathcal{L}(q, \theta) \) that make it easy to compute

- \( \text{arg max}_q \mathcal{L}(q, \theta) \), for a given \( \theta \), and
- \( \text{arg max}_\theta \mathcal{L}(q, \theta) \), for a given \( q \).
Let’s investigate the lower bound:

\[ \mathcal{L}(q, \theta) = \sum_z q(z) \log \left( \frac{p(x, z | \theta)}{q(z)} \right) \]

\[ = \sum_z q(z) \log \left( \frac{p(z | x, \theta) p(x | \theta)}{q(z)} \right) \]

\[ = \sum_z q(z) \log \left( \frac{p(z | x, \theta)}{q(z)} \right) + \sum_z q(z) \log p(x | \theta) \]

\[ = -\text{KL}[q(z), p(z | x, \theta)] + \log p(x | \theta) \]

Amazing! We get back an equality for the marginal likelihood:

\[ \log p(x | \theta) = \mathcal{L}(q, \theta) + \text{KL}[q(z), p(z | x, \theta)] \]
Maxizing over $q$ for fixed $\theta = \theta^{old}$.

- Find $q$ maximizing
  \[
  \mathcal{L}(q, \theta^{old}) = -\text{KL}[q(z), p(z \mid x, \theta^{old})] + \log p(x \mid \theta^{old})
  \]

- Recall $\text{KL}(p\|q) \geq 0$, and $\text{KL}(p\|p) = 0$.
- Best $q$ is $q^*(z) = p(z \mid x, \theta^{old})$ and
  \[
  \mathcal{L}(q^*, \theta^{old}) = -\text{KL}[p(z \mid x, \theta^{old}), p(z \mid x, \theta^{old})] + \log p(x \mid \theta^{old}) = 0
  \]

- Summary:
  \[
  \log p(x \mid \theta^{old}) = \mathcal{L}(q^*, \theta^{old}) \quad \text{(tangent at } \theta^{old}).
  \]
  \[
  \log p(x \mid \theta) \geq \mathcal{L}(q^*, \theta) \quad \forall \theta
  \]
Tight lower bound for any chosen $\theta$

For $\theta^{\text{old}}$, take $q(z) = p(z \mid x, \theta^{\text{old}})$. Then

1. $\log p(x \mid \theta) \geq \mathcal{L}(q, \theta) \forall \theta$. [Global lower bound].
2. $\log p(x \mid \theta^{\text{old}}) = \mathcal{L}(q, \theta^{\text{old}})$. [Lower bound is tight at $\theta^{\text{old}}$.]

From Bishop’s *Pattern recognition and machine learning*, Figure 9.14.
Maximizing over $\theta$ for fixed $q$

- Consider maximizing the lower bound $\mathcal{L}(q, \theta)$:

$$
\mathcal{L}(q, \theta) = \sum_z q(z) \log \left( \frac{p(x, z | \theta)}{q(z)} \right)
= \sum_z q(z) \log p(x, z | \theta) - \sum_z q(z) \log q(z)
$$

- Maximizing $\mathcal{L}(q, \theta)$ equivalent to maximizing $\mathbb{E}[\text{complete data log-likelihood}]$ (for fixed $q$).
General EM Algorithm

1. Choose initial $\theta^{\text{old}}$.

2. **Expectation Step**
   - Let $q^*(z) = p(z \mid x, \theta^{\text{old}})$. [$q^*$ gives best lower bound at $\theta^{\text{old}}$]
   - Let
     \[
     J(\theta) := \mathcal{L}(q^*, \theta) = \sum_z q^*(z) \log \left( \frac{p(x, z \mid \theta)}{q^*(z)} \right)
     \]
     expectation w.r.t. $z \sim q^*(z)$

3. **Maximization Step**
   \[
   \theta^{\text{new}} = \arg \max_{\theta} J(\theta).
   \]
   [Equivalent to maximizing expected complete log-likelihood.]

4. Go to step 2, until converged.
Does EM Work?
Does EM Work?

EM Gives Monotonically Increasing Likelihood: By Picture

From Bishop’s *Pattern recognition and machine learning*, Figure 9.14.
EM Gives Monotonically Increasing Likelihood: By Math

1. Start at $\theta^{\text{old}}$.
2. Choose $q^*(z) = \arg \max_q \mathcal{L}(q, \theta^{\text{old}})$. We’ve shown

$$\log p(x | \theta^{\text{old}}) = \mathcal{L}(q^*, \theta^{\text{old}})$$

3. Choose $\theta^{\text{new}} = \arg \max_\theta \mathcal{L}(q^*, \theta^{\text{old}})$. So

$$\mathcal{L}(q^*, \theta^{\text{new}}) \geq \mathcal{L}(q^*, \theta^{\text{old}}).$$

Putting it together, we get

$$\log p(x | \theta^{\text{new}}) \geq \mathcal{L}(q^*, \theta^{\text{new}}) \quad \mathcal{L} \text{ is a lower bound}$$

$$\geq \mathcal{L}(q^*, \theta^{\text{old}}) \quad \text{By definition of } \theta^{\text{new}}$$

$$= \log p(x | \theta^{\text{old}}) \quad \text{Bound is tight at } \theta^{\text{old}}.$$
Suppose We Maximize the ELBO...

- Suppose we have found a **global maximum** of $\mathcal{L}(q, \theta)$:

$$L(q^*, \theta^*) \geq L(q, \theta) \forall q, \theta,$$

where of course

$$q^*(z) = p(z \mid x, \theta^*).$$

- Claim: $\theta^*$ is a global maximum of $\log p(x \mid \theta^*)$.

- Proof: For any $\theta'$, we showed that for $q'(z) = p(z \mid x, \theta')$ we have

$$\log p(x \mid \theta') = \mathcal{L}(q', \theta') + \text{KL}[q', p(z \mid x, \theta')]$$

$$= \mathcal{L}(q', \theta')$$

$$\leq \mathcal{L}(q^*, \theta^*)$$

$$= \log p(x \mid \theta^*)$$
Convergence of EM

- Let $\theta_n$ be value of EM algorithm after $n$ steps.
- Define “transition function” $M(\cdot)$ such that $\theta_{n+1} = M(\theta_n)$.
- Suppose log-likelihood function $\ell(\theta) = \log p(x | \theta)$ is differentiable.
- Let $S$ be the set of stationary points of $\ell(\theta)$. (i.e. $\nabla_\theta \ell(\theta) = 0$)

**Theorem**

*Under mild regularity conditions*, for any starting point $\theta_0$,

- $\lim_{n \to \infty} \theta_n = \theta^*$ for some stationary point $\theta^* \in S$ and
- $\theta^*$ is a fixed point of the EM algorithm, i.e. $M(\theta^*) = \theta^*$. Moreover,
- $\ell(\theta_n)$ strictly increases to $\ell(\theta^*)$ as $n \to \infty$, unless $\theta_n \equiv \theta^*$.

---

Variations on EM
EM Gives Us Two New Problems

• The “E” Step: Computing

\[ J(\theta) := \mathcal{L}(q^*, \theta) = \sum_z q^*(z) \log \left( \frac{p(x, z | \theta)}{q^*(z)} \right) \]

• The “M” Step: Computing

\[ \theta^{\text{new}} = \arg \max_\theta J(\theta). \]

• Either of these can be too hard to do in practice.
Generalized EM (GEM)

- Addresses the problem of a difficult “M” step.
- Rather than finding
  \[ \theta^{\text{new}} = \arg\max_{\theta} J(\theta), \]
  find any \( \theta^{\text{new}} \) for which
  \[ J(\theta^{\text{new}}) > J(\theta^{\text{old}}). \]
- Can use a standard nonlinear optimization strategy
  - e.g. take a gradient step on \( J \).
- We still get monotonically increasing likelihood.
Suppose “E” step is difficult:
- Hard to take expectation w.r.t. \( q^*(z) = p(z \mid x, \theta^{old}) \).

Solution: Restrict to distributions \( Q \) that are easy to work with.

Lower bound now looser:

\[
q^* = \arg \min_{q \in Q} \text{KL}[q(z), p(z \mid x, \theta^{old})]
\]
EM in Bayesian Setting

- Suppose we have a prior $p(\theta)$.
- Want to find MAP estimate: $\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} p(\theta | x)$:

\[
p(\theta | x) = \frac{p(x | \theta)p(\theta)}{p(x)}
\]

\[
\log p(\theta | x) = \log p(x | \theta) + \log p(\theta) - \log p(x)
\]

- Still can use our lower bound on $\log p(x, \theta)$.

\[
J(\theta) := \mathcal{L}(q^*, \theta) = \sum_z q^*(z) \log \left( \frac{p(x, z | \theta)}{q^*(z)} \right)
\]

- Maximization step becomes

\[
\theta^{\text{new}} = \arg\max_{\theta} [J(\theta) + \log p(\theta)]
\]

- Homework: Convince yourself our lower bound is still tight at $\theta$. 
Summer Homework: Gaussian Mixture Model (Hints)
Homework: Derive EM for GMM from General EM Algorithm

- Subsequent slides may help set things up.
- Key skills:
  - MLE for multivariate Gaussian distributions.
  - Lagrange multipliers
Gaussian Mixture Model (k Components)

- GMM Parameters
  - Cluster probabilities: \( \pi = (\pi_1, \ldots, \pi_k) \)
  - Cluster means: \( \mu = (\mu_1, \ldots, \mu_k) \)
  - Cluster covariance matrices: \( \Sigma = (\Sigma_1, \ldots, \Sigma_k) \)

- Let \( \theta = (\pi, \mu, \Sigma) \).

- Marginal log-likelihood
  \[
  \log p(x \mid \theta) = \log \left\{ \sum_{z=1}^{k} \pi_z N(x \mid \mu_z, \Sigma_z) \right\}
  \]
$q^*(z)$ are “Soft Assignments”

- Suppose we observe $n$ points: $X = (x_1, \ldots, x_n) \in \mathbb{R}^{n \times d}$.
- Let $z_1, \ldots, z_n \in \{1, \ldots, k\}$ be corresponding hidden variables.
- Optimal distribution $q^*$ is:

$$q^*(z) = p(z \mid x, \theta).$$

- Convenient to define the conditional distribution for $z_i$ given $x_i$ as

$$\gamma_i^j := p(z = j \mid x_i) = \frac{\pi_j \mathcal{N}(x_i \mid \mu_j, \Sigma_j)}{\sum_{c=1}^{k} \pi_c \mathcal{N}(x_i \mid \mu_c, \Sigma_c)}.$$
**Expectation Step**

- The complete log-likelihood is

\[
\log p(x, z | \theta) = \sum_{i=1}^{n} \log \left[ \pi_{z} \mathcal{N}(x_i | \mu_{z}, \Sigma_{z}) \right]
\]

- Take the expected complete log-likelihood w.r.t. \( q^* \):

\[
J(\theta) = \sum_{z} q^*(z) \log p(x, z | \theta)
\]

\[
= \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{i}^{j} \left[ \log \pi_{j} + \log \mathcal{N}(x_i | \mu_{j}, \Sigma_{j}) \right]
\]
Maximization Step

Find $\theta^*$ maximizing $J(\theta)$:

\[
\mu_c^{\text{new}} = \frac{1}{n_c} \sum_{i=1}^{n} \gamma_{ci} x_i
\]
\[
\Sigma_c^{\text{new}} = \frac{1}{n_c} \sum_{i=1}^{n} \gamma_{ci} (x_i - \mu_{\text{MLE}})(x_i - \mu_{\text{MLE}})^T
\]
\[
\pi_c^{\text{new}} = \frac{n_c}{n},
\]

for each $c = 1, \ldots, k$. 