EM (more formally)

Semi-optional technical material alert!

The formal treatment helps us use EM correctly in more complex situations. However, EM algorithms for many problems can "guessed at" using intuition.

The more formal treatment is not needed for the homework.

EM (more formally)

- Assume K clusters. Index over clusters by k, over points by n.
- New notation for cluster membership:

  For each point, n, z_n is a vector of K values where exactly one $z_{n,k} = 1$, and all others are 0. Note that $\sum_k z_{n,k} = 1$.

EM (more formally)

- Denote cluster priors by:

$$\pi_k \equiv p\left(z_k = 1\right)$$

- Denote the responsibilities that each cluster has for each point by:

$$\gamma\left(z_{n,k}\right) \equiv p\left(z_{n,k} = 1 \middle| x_n, \theta^{(s)}\right) = \frac{\pi_k \; p\left(x_n \middle| z_{n,k} = 1, \theta_k^{(s)}\right)}{\sum_{k'} \pi_{k'} \; p\left(x_n \middle| z_{n,k'} = 1, \theta_{k'}^{(s)}\right)}$$

EM (more formally)

Represent the entire data set of N points, $\mathbf{x}_n$,
as a matrix X with rows $\mathbf{x}_n^T$.

Represent the latent variable assignments with a matrix Z.
(For the true assignment, each row is zero except for a single element that is 1.)

We call $\{Z, X\}$ the *complete* data set (everthing is known).
The observed data, $\{X\}$, is called the *incomplete* data set.

EM (more formally)

We assume that computing the MLE of parameters,

$$\arg\max_{\theta}\left\{\log\left\{p(Z,X|\theta)\right\}\right\}$$

with complete data is relatively easy.

Recall our intuitive treatment of EM for GMM. If we knew the cluster membership, we would know how to compute the means.

Since we did not know the cluster membership we did a weighted computation.

EM (more formally)

Notice the complexity of the incomplete log likelihood:

$$\log\left(p(X|\theta)\right)=\sum_{n}\log\underbrace{\left(\sum_{k}\pi_k\ p(x_n|\theta)\right)}_{\text{nasty sum in log}}$$

By constrast, for complete log likelihood we can incorporate the assignment by:

$$p(X,Z|\theta)=\prod_{n}\prod_{k}\pi_k^{z_{n,k}}\left\{p(x_n|\theta)\right\}^{z_{n,k}}$$

So

$$\log\left(p(X,Z|\theta)\right)=\sum_{n}\sum_{k}\left\{z_{n,k}\ \left(\log(\pi_k)+\log\left(p(x_n|\theta)\right)\right)\right\}$$

(No nasty sum in log; well suited for the expectation calculation).

EM (more formally)

For the E step, we compute the responsibilities which is straightforward.

Next, define $\mathrm{Q}\left(\theta^{(s+1)},\theta^{(s)}\right)=\sum_{Z}p\left(Z|X,\theta^{(s)}\right)\log\left(p\left(X,Z|\theta^{(s+1)}\right)\right)$

(Expectation of $\log\left(p\left(X,Z|\theta^{(s+1)}\right)\right)$ over $p\left(Z|X,\theta^{(s)}\right)$).

The M step then computes $\theta^{(s+1)}=\arg\max_{\theta}\left\{\mathrm{Q}\left(\theta^{(s+1)},\theta^{(s)}\right)\right\}$

Maximizing Q is generally feasible and corresponds to the intuitive development.

General EM algorithm

1. Choose initial values for $\theta^{(s=1)}$
   (can also do assignments, but then jump to M step).
2. E step: Evalute $p\left(Z|X,\theta^{(s)}\right)$
3. M step: Evalute $\theta^{(s+1)}=\arg\max_{\theta}\left\{\mathrm{Q}\left(\theta^{(s+1)},\theta^{(s)}\right)\right\}$
   where $\mathrm{Q}\left(\theta^{(s+1)},\theta^{(s)}\right)=\sum_{Z}p\left(Z|X,\theta^{(s)}\right)\log\left(p\left(X,Z|\theta^{(s+1)}\right)\right)$
4. Check for convergence; If not done, goto 2.

★ At each step, our objective function increases unless it is at a local maximum. It is important to check this is happening for debugging!

General EM algorithm

★ At each step, our objective function (conditioned on the current model) increases unless it is at a local maximum. It is important to check this is happening for debugging!

Recall our objective function for the case of a mixture model:

$$p(X|\theta)=\prod_n \sum_k p(k)p(x_n|\theta_k)$$

or

$$\log\left(p(X|\theta)\right)=\sum_n \log\left(\sum_k p(k)p(x_n|\theta_k)\right)$$

Implementation tip. This is conveniently available from the computation of responsibilities (before normalization).

Deriving the GMM M-step

Evalute $\theta^{(s+1)} = \arg\max_\theta \left\{Q\left(\theta^{(s+1)},\theta^{(s)}\right)\right\}$

where $Q\left(\theta^{(s+1)},\theta^{(s)}\right)=\sum_Z p\left(Z|X,\theta^{(s)}\right)\log\left(p\left(X,Z|\theta^{(s)}\right)\right)$

Recall that $\log\left(p(X,Z|\theta)\right)=\sum_n\sum_k\left\{z_{n,k}\left(\log(\pi_k)+\log\left(p(x_n|\theta_k)\right)\right)\right\}$

So $Q\left(\theta^{(s+1)},\theta^{(s)}\right)=\sum_Z p\left(Z|X,\theta^{(s)}\right)\sum_n\sum_k\left\{z_{n,k}\left(\log(\pi_k)+\log\left(p(x_n|\theta_k)\right)\right)\right\}$

Deriving the GMM M-step

$$Q\left(\theta^{(s+1)},\theta^{(s)}\right)=\sum_Z p\left(Z|X,\theta^{(s)}\right)\sum_n\sum_k\left\{z_{n,k}\left(\log(\pi_k)+\log\left(p(x_n|\theta_k)\right)\right)\right\}$$
$$=\sum_n\sum_k\underbrace{\sum_Z p\left(Z|X,\theta^{(s)}\right)\left\{z_{n,k}\left(\log(\pi_k)+\log\left(p(x_n|\theta_k)\right)\right)\right\}}_{\text{inner sum}}$$

This exchanging of summing order says that **instead** of summing over points and clusters for all correspondences Z, we sum over all correspondences for a **given** point and cluster.

We will focus on the inner sum.

Z is all possible correspondences. To generate them all more explicitly, we can consider the first point. For each possible assignment of the first point, we then need all possible combinations of the other points. To get that, we consider all possible assignments of the second point, together with all possible assignments of the remaining points. This shows:

$$\sum_Z(\ ) \equiv \sum_{\mathbf{z}_1}\sum_{\mathbf{z}_2}\sum_{\mathbf{z}_3}\cdots\sum_{\mathbf{z}_N}(\ )$$

Note that a sum over $\mathbf{z}_n$ is short hand for a sum over cluster assignments for point n. Hence each of the sums on the right are over clusters.

$$\sum_{\mathbf{z}_n}(\bullet) \equiv \sum_{k_n=1}^{K}(\bullet)$$

Further, because the points are independent, we have:

$$p(Z|\bullet)=\prod_{\mathbf{z}_i} p(\mathbf{z}_i|\bullet)$$

$$\underbrace{\sum_{Z} p\left(Z \middle| X,\theta^{(s)}\right) \bullet z_{n,k} \bullet \left(\log\left(\pi_k\right) + \log\left(p\left(x_n \middle| \theta_k\right)\right)\right)}_{\text{inner sum from previous}}$$

$$= \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \cdots \sum_{k_N=1}^{K} \prod_{i} p\left(z_{i,k_i} \middle| x_i,\theta^{(s)}\right) \bullet z_{n,k} \bullet \left(\log\left(\pi_k\right) + \log\left(p\left(x_n \middle| \theta_k\right)\right)\right)$$

$$= \sum_{k_n=1}^{K} p\left(z_{n,k_n} \middle| x_i,\theta^{(s)}\right) \bullet z_{n,k} \bullet \left(\log\left(\pi_k\right) + \log\left(p\left(x_n \middle| \theta_k\right)\right)\right) \bullet \underbrace{\sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \cdots \sum_{k_{n-1}=1}^{K} \sum_{k_{n+1}=1}^{K} \cdots \sum_{k_N=1}^{K} \prod_{i \neq n} p\left(z_{i,k_i} \middle| x_i,\theta^{(s)}\right)}_{\text{All possibilities without point n. Since it covers all cases, this evaluates to unity!}}$$

pedantically,

$$= \sum_{k_n=1}^{K} p\left(z_{n,k_n} \middle| x_i,\theta^{(s)}\right) \bullet z_{n,k} \bullet \left(\log\left(\pi_k\right) + \log\left(p\left(x_n \middle| \theta_k\right)\right)\right) \bullet \sum_{k_1=1}^{K} p\left(z_{1,k_1} \middle| x_1,\theta^{(s)}\right) \sum_{k_2=1}^{K} p\left(z_{2,k_2} \middle| x_2,\theta^{(s)}\right) \cdots \sum_{k_{n-1}=1}^{K} p\left(z_{n-1,k_{n-1}} \middle| x_{n-1},\theta^{(s)}\right) \sum_{k_{n+1}=1}^{K} p\left(z_{2,k_2} \middle| x_i,\theta^{(s)}\right) \cdots \sum_{k_N=1}^{K} p\left(z_{N,k_N} \middle| x_N,\theta^{(s)}\right)$$

$$= \sum_{k_n=1}^{K} p\left(z_{n,k_n} \middle| x_i,\theta^{(s)}\right) \bullet z_{n,k} \bullet \left(\log\left(\pi_k\right) + \log\left(p\left(x_n \middle| \theta_k\right)\right)\right)$$

$$\underbrace{\sum_{Z} p\left(Z \middle| X,\theta^{(s)}\right) \bullet z_{n,k} \bullet \left(\log\left(\pi_k\right) + \log\left(p\left(x_n \middle| \theta_k\right)\right)\right)}_{\text{inner sum from previous}}$$

$$= \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \cdots \sum_{k_N=1}^{K} \prod_{i} p\left(z_{i,k_i} \middle| x_i,\theta^{(s)}\right) \bullet z_{n,k} \bullet \left(\log\left(\pi_k\right) + \log\left(p\left(x_n \middle| \theta_k\right)\right)\right)$$

$$= \sum_{k_n=1}^{K} p\left(z_{n,k_n} \middle| x_i,\theta^{(s)}\right) \bullet z_{n,k} \bullet \left(\log\left(\pi_k\right) + \log\left(p\left(x_n \middle| \theta_k\right)\right)\right) \bullet \underbrace{\sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \cdots \sum_{k_{n-1}=1}^{K} \sum_{k_{n+1}=1}^{K} \cdots \sum_{k_N=1}^{K} \prod_{i \neq n} p\left(z_{i,k_i} \middle| x_i,\theta^{(s)}\right)}_{\text{All possibilities without point n. Since it covers all cases, this evaluates to unity!}}$$

$$= \sum_{k_n=1}^{K} p\left(z_{n,k_n} \middle| x_i,\theta^{(s)}\right) \bullet z_{n,k} \bullet \left(\log\left(\pi_k\right) + \log\left(p\left(x_n \middle| \theta_k\right)\right)\right)$$

$$= p\left(z_{n,k} = 1 \middle| X,\theta^{(s)}\right)\left(\log\left(\pi_k\right) + \log\left(p\left(x_n \middle| \theta_k\right)\right)\right)$$

$$= \gamma\left(z_{n,k}\right)\left(\log\left(\pi_k\right) + \log\left(p\left(x_n \middle| \theta_k\right)\right)\right)$$ (definition of $\gamma\left(z_{n,k}\right)$, the responsibility)

Deriving the M-step

$$\mathrm{Q}\left(\theta^{(s+1)},\theta^{(s)}\right) = \sum_{Z} p\left(Z \middle| X,\theta^{(s)}\right) \sum_{n} \sum_{k} \left\{ z_{n,k} \left(\log\left(\pi_k\right) + \log\left(p\left(x_n \middle| \theta_k\right)\right)\right)\right\}$$

$$= \sum_{n} \sum_{k} \left\{ \gamma\left(z_{n,k}\right) \left(\log\left(\pi_k\right) + \log\left(p\left(x_n \middle| \theta_k\right)\right)\right)\right\}$$

We need to maximize this with respect to the parameters for each cluster, *k*. Notice that:

$$\frac{\delta}{\delta\theta_{k^*}} \mathrm{Q}\left(\theta^{(s+1)},\theta^{(s)}\right) = \sum_{n} \left\{ \gamma\left(z_{n,k^*}\right) \frac{\delta}{\delta\theta_{k^*}} \left(\log\left(\pi_{k^*}\right) + \log\left(p\left(x_n \middle| \theta_{k^*}\right)\right)\right)\right\}$$

(The values of *k* not of current interest, i.e., not *k**, die)

Example—deriving the GMM M-step

$$\frac{\delta}{\delta\mu_k} \mathrm{Q}\left(\theta^{(s+1)},\theta^{(s)}\right) = \sum_{n} \left\{ \gamma\left(z_{n,k}\right) \frac{\delta}{\delta\mu_k} \left(\log\left(\pi_k\right) + \log\left(p\left(x_n \middle| \theta_k\right)\right)\right)\right\}$$

$$= \sum_{n} \left\{ \gamma\left(z_{n,k}\right) \frac{\delta}{\delta\mu_k} \left(\log\left(p\left(x_n \middle| \theta_k\right)\right)\right)\right\}$$

$$\sum_{n} \left\{ \gamma\left(z_{n,k}\right) \frac{\delta}{\delta\mu_k} \left(\log\left(\mathrm{N}\left(x_n \middle| \mu_k, \Sigma_k\right)\right)\right)\right\}$$

Example—deriving the GMM M-step

$$\mathrm{N}\left(\mathbf{x}_n \middle| \mu_k, \Sigma_k\right) = \frac{1}{(2\pi)^{D/2}\left|\Sigma_k\right|^{1/2}} \exp\left(-\frac{1}{2}\left(\mathbf{x}_n - \mu_k\right)^T \Sigma_k^{-1}\left(\mathbf{x}_n - \mu_k\right)\right)$$

$$\log\left(\mathrm{N}\left(\mathbf{x}_n \middle| \mu_k, \Sigma_k\right)\right) = \log\left(\frac{1}{(2\pi)^{D/2}\left|\Sigma_k\right|^{1/2}}\right) - \frac{1}{2}\left(\mathbf{x}_n - \mu_k\right)^T \Sigma_k^{-1}\left(\mathbf{x}_n - \mu_k\right)$$

$$\frac{\delta}{\delta\mu_k}\log\left(\mathrm{N}\left(\mathbf{x}_n \middle| \mu_k, \Sigma_k\right)\right) = \Sigma_k^{-1}\left(\mathbf{x}_n - \mu_k\right)$$

(exercise for the interested)

Example—deriving the GMM M-step

$$\frac{\delta}{\delta\mu_k}\mathrm{Q}\left(\theta^{(s+1)}, \theta^{(s)}\right) = \sum_n \left\{\gamma\left(z_{n,k}\right) \frac{\delta}{\delta\mu_k}\left(\log\left(\mathrm{N}\left(x_n \middle| \mu_k, \Sigma_k\right)\right)\right)\right\}$$

$\frac{\delta}{\delta\mu_k}\mathrm{Q}\left(\theta^{(s+1)}, \theta^{(s)}\right) = 0$ means that

$\sum_n \left\{\gamma\left(z_{n,k}\right) \Sigma_k^{-1}\left(\mathbf{x}_n - \mu_k\right)\right\} = 0$ (Inner sigma is precision matrix, not a sum

$\sum_n \left\{\gamma\left(z_{n,k}\right)\left(\mathbf{x}_n - \mu_k\right)\right\} = 0$ (Multiply by Σ_k)

Example—deriving the GMM M-step

So, $\sum_n \left\{\gamma\left(z_{n,k}\right)\left(\mathbf{x}_n - \mu_k\right)\right\} = 0$

and $\mu_k \sum_n \left\{\gamma\left(z_{n,k}\right)\right\} = \sum_n \left\{\gamma\left(z_{n,k}\right)\left(\mathbf{x}_n\right)\right\}$

and $\mu_k = \dfrac{\sum_n \left\{\gamma\left(z_{n,k}\right)\left(\mathbf{x}_n\right)\right\}}{\sum_n \left\{\gamma\left(z_{n,k}\right)\right\}}$ (same as before)

Example—deriving the GMM M-step

Finding variances/covariances is similar.

Finding the mixing coefficients is also similar, except we also need to enforce that they sum to one.

(Here the equations for the *k*'s are coupled).

So we use Lagrange Multipliers.