EM for statistical clustering

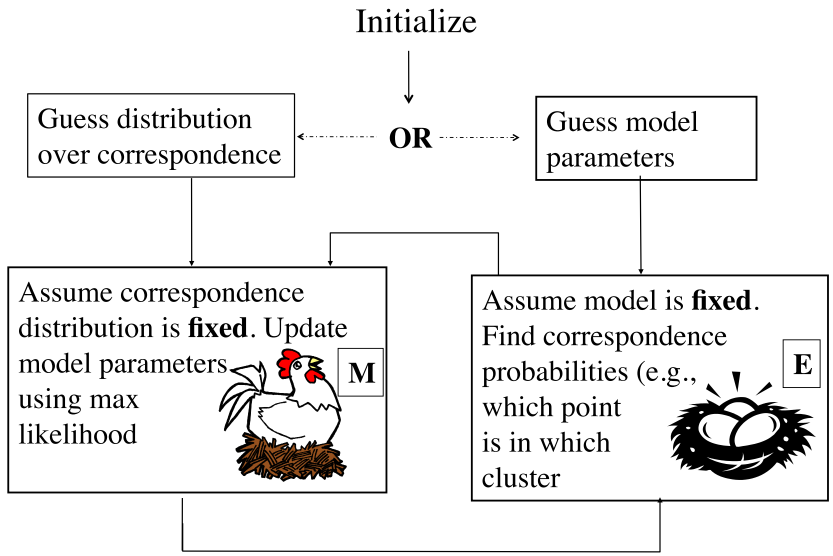
Given the number of clusters and some data, we fit the cluster parameters by maximizing the objective function:

$$\log\left(p\left(\left\{\mathbf{x}_i\right\}\right)\right)=\sum_i \log\left\{\sum_k p\left(k\right)p\left(\mathbf{x}_i|k\right)\right\}$$

This is generally intractable. (We can only find a local maximum).

We do this with Expectation Maximization (EM).

EM flow chart



Initialize

Guess distribution over correspondence

OR

Guess model parameters

Assume correspondence distribution is **fixed**. Update model parameters using max likelihood

M

Assume model is **fixed**. Find correspondence probabilities (e.g., which point is in which cluster

E

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 $p\left(Z\middle|X,\theta^{(s)}\right)$
3. M step: Evalute $\theta^{(s+1)}=\arg\max_{\theta}\left\{Q\left(\theta^{(s+1)},\theta^{(s)}\right)\right\}$

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

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

or

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

GMM M-step

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

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

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

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

GMM M-step

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

(Intuitive?)

GMM M-step

$$\begin{aligned} &\sum_{Z} p\left(Z \middle| X,\theta^{(s)}\right) f(n,k) \\ &= \sum_{Z}\prod_{n'} p\left(z_{n'} \middle| X,\theta^{(s)}\right) f(n,k) \qquad \text{(independence!)} \\ &= \sum_{z_n} p\left(z_n \middle| X,\theta^{(s)}\right) f(n,k) \underbrace{\sum_{z_1}\ldots\sum_{z_{n-1}}\sum_{z_{n+1}}\ldots.\sum_{N}\prod_{n'\neq n} p\left(z_{n'} \middle| X,\theta^{(s)}\right)}_{\text{all possibilities without point n.}} \\ &= \sum_{z_n} p\left(z_n \middle| X,\theta^{(s)}\right) f(n,k) \end{aligned}$$

GMM M-step

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

Here we use the special nature of the indicator variables. They pick the value in the sum corresponding to the index.

GMM 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:

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

GMM M-step

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

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

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

GMM M-step

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

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

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

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

$$-\frac{1}{2}\Sigma_k\bullet\sum_{n}\left\{\gamma\left(z_{n,k}\right)\left(\mathbf{x}_n-\mu_k\right)\right\}=0$$

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

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)

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.

Using Lagrange Multipliers

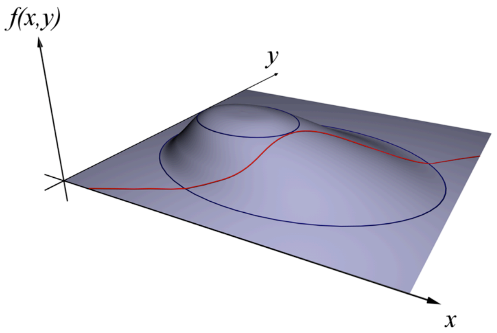
Now we find stationary points with respect to $\left\{\pi_k, \lambda\right\}$ of

$$\mathrm{Q}\left(\theta^{(s+1)}, \theta^{(s)}\right) + \lambda\left(\sum_{k} \pi_k - 1\right)$$

Note that differentiating with respect to λ, and setting the result to zero puts the constraint into the equations.

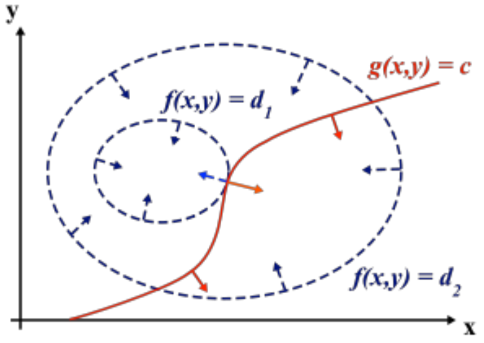
But the real problem is doing the optimization under the constraint.

Using Lagrange Multipliers



From WikiPedia

Using Lagrange Multipliers



From WikiPedia

Using Lagrange Multipliers

$\nabla f \parallel \nabla g$

$\nabla f = \lambda \nabla g$

So, $\nabla(f - \lambda g) = 0$

or, $\nabla(f + \lambda g) = 0$ (negate λ)

Using Lagrange Multipliers

Now we find stationary points with respect to $\{\pi_k, \lambda\}$ of

$$\mathrm{Q}\left(\theta^{(s+1)}, \theta^{(s)}\right) + \lambda\left(\sum_k \pi_k - 1\right)$$

$$\frac{\delta}{\delta \pi_k}\left\{\mathrm{Q}\left(\theta^{(s+1)}, \theta^{(s)}\right) + \lambda\left(\sum_k \pi_k - 1\right)\right\}$$

$$= \sum_n \left\{\gamma\left(z_{n,k}\right) \frac{\delta}{\delta \pi_k}\left(\log\left(\pi_k\right) \log\left(\mathrm{N}\left(x_n \middle| \mu_k, \Sigma_k\right)\right)\right)\right\} + \lambda$$

$$= \sum_n \left\{\gamma\left(z_{n,k}\right) \frac{1}{\pi_k}\right\} + \lambda$$

Setting the result to zero, $\sum_n \left\{\gamma\left(z_{n,k}\right) \frac{1}{\pi_k}\right\} + \lambda = 0$

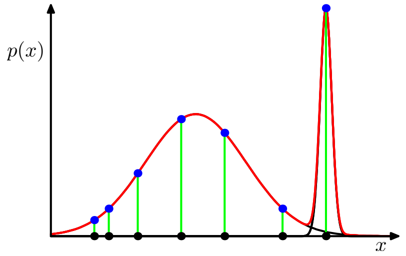
So $\pi_k = \frac{\sum_n \{\gamma(z_{n,k})\}}{-\lambda}$

Summing over k gives, $1 = \frac{\sum_k \sum_n \{\gamma(z_{n,k})\}}{-\lambda} = \frac{N}{-\lambda}$

So, $\lambda = -N$, and $\pi_k = \frac{\sum_n \{\gamma(z_{n,k})\}}{N}$ as before.

EM in practice

- For GMM we need to consider clusters that have only one point:



- Easily fixed by adding a prior on the variance.

EM in practice

- Tying parameters (using GMM as an example)
  - We can improve stability by assuming the variances (or covariances) for all clusters are the same.
  - Updates work as you expect. Instead of multiple weighted sums, you just use one big on.
  - But note that one advantage of GMM over K-means is that the scale is naturally taken of, and the clusters **can** have different variances.

EM in practice

- You must check that the log likelihood increases!
- A simple way to compute it during an iteration:

Recall our objective function:

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

Consider how we might compute the responsibilities

$\gamma(n,k) \propto p(k)p(x_n|k)$
(Then normalize once you have them all).

So, make a running sum of the unormalized values

EM in practice

- Precision problems --> must work with logs

- But we need to exponentiate to normalize --> rescaling tricks

Let $P=\{p_i\}$.

Suppose we want $Q=\frac{1}{\sum_i p_i}\{p_i\}$

Where we need to use $V=\{\log(p_i)\}$

and $\exp(p_i)$ is too small, and the sum of them might be zero.

Let $M = \max\{\log(p_i)\}$

Observe that working with $V' = \{\log(p_i)-M\}$ does the trick.

EM in practice

- Memory problems ---> we can compute means, etc., as running totals so that we do not need to store responsibilities for all points over all clusters.