Statistical clustering

Basic, very general, generative model

$$p(\mathbf{x}) = \sum_k p(k)\, p(\mathbf{x}|k)$$

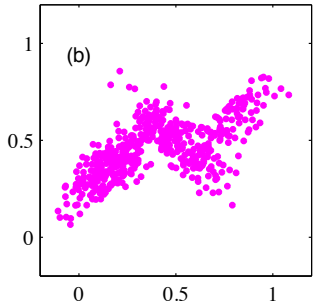
Data density, given the above model.

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

Example: Three bivariate Gaussians

Points sampled according to:

$$p(\mathbf{x}) = \sum_k p(k)\, p(\mathbf{x}|k)$$



Even if we know the exact model, we cannot be sure from the data which point comes from which cluster. We only have the distribution for this.

EM for statistical clustering

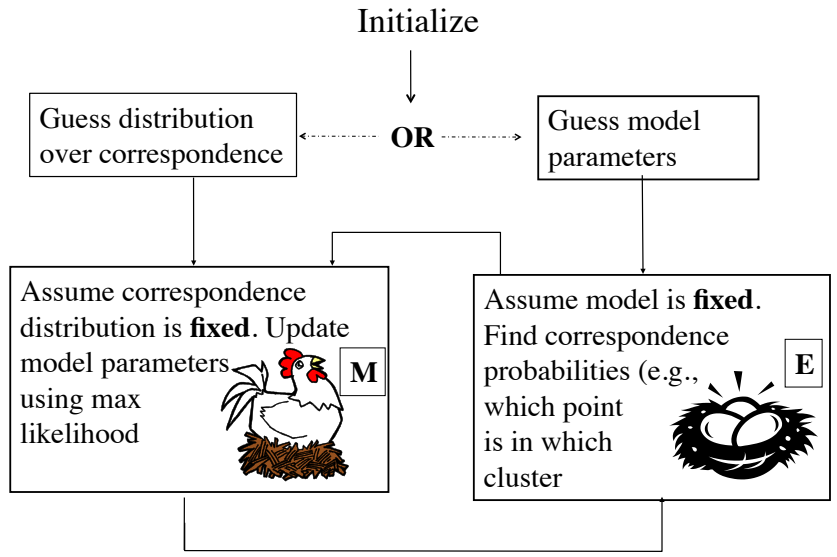
Given the number of clusters and some data, we fit the cluster parameters by maximizing the objective function:

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

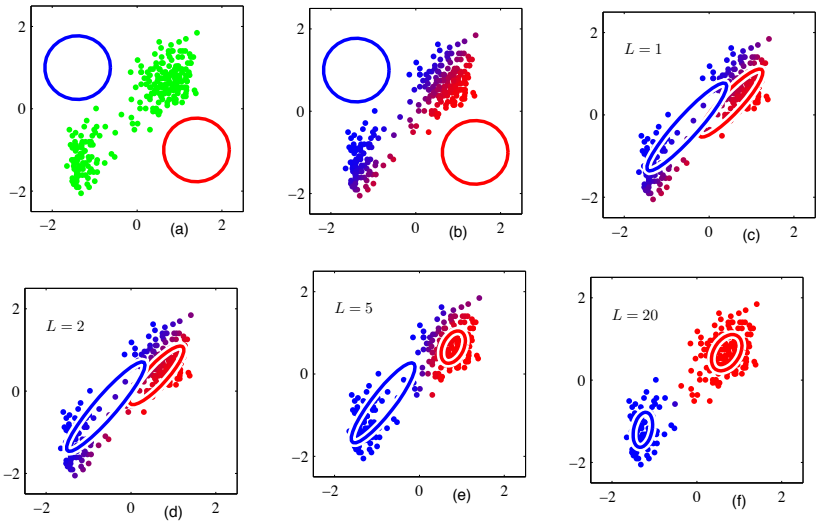
This is generally intractable. (We can only find a local maximum).

We do this with Expectation Maximization (EM).

EM flow chart



EM illustrated



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 $z_{n,k} = 1$, all others are 0. Note that $\sum_k z_{n,k} = 1$.

EM (more formally)

- Denote cluster priors by:

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

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

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

EM (more formally)

- Responsibilities For GMM:

$$\gamma(z_{n,k}) = \frac{\pi_k \mathrm{N}(x_n | u_k^{(s)}, \Sigma_k^{(s)})}{\sum_{k'} \pi_{k'} \mathrm{N}(x_n | u_{k'}^{(s)}, \Sigma_{k'}^{(s)})}$$

EM (more formally)

Represent the entire data set of N points, $\mathbf{x}_n$,
with 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, which happens to be an expectation of the complete log likelihood, over the assignment (responsibility) distribution.

EM (more formally)

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

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 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.

EM (more formally)

Notice the complexity of the incomplete log likelihood:

$$\log\left(p(X|\theta)\right) = \sum_n \log\underbrace{\left(\sum_k \pi_k \; p(x_n|\theta)\right)}_{\text{nasty sum in log}}$$

The 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

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

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

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

   where $\mathrm{Q}\left(\theta^{(s+1)},\theta^{(s)}\right)=\sum_{Z}p\left(Z\middle|X,\theta^{(s)}\right)\log\left(p\left(X,Z\middle|\theta^{(s+1)}\right)\right)$
4. Check for convergence; If not done, goto 2.

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

GMM M-step

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

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

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

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

GMM M-step

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

(Intuitive?)

GMM M-step

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

GMM M-step

$$Q\left(\theta^{(s+1)},\theta^{(s)}\right)$$

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

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

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

(Here we use the special nature of the indicator variables.

All possible values of z_n are the K cluster assignments for point n.)

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

(We know that $z_{n,k}$ is zero for $k' \neq k$)

GMM M-step

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

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

We need to maximize this with respect to the parameters for each cluster, k.