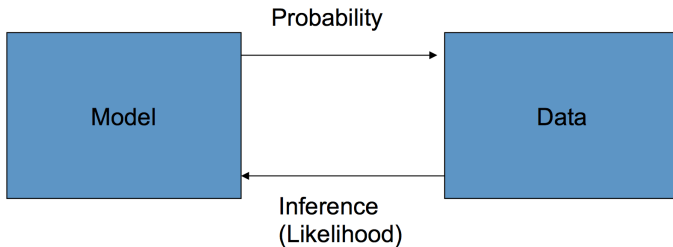


Maximum Likelihood Estimation, Expectation Maximization

March 14, 2017

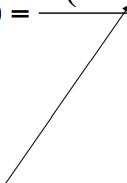


A model of the data generating process gives rise to data.
Model estimation from data is most commonly through
Likelihood estimation

Likelihood Function

$$P(Model | Data) = \frac{P(Data | Model)P(Model)}{P(Data)}$$

Likelihood Function



Find the “best” model which has generated the data. In a likelihood function the data is considered fixed and one searches for the best model over the different choices available.

Estimating Parameters

- Let Y be a random variable with a distribution of known type but unknown parameter value θ .
Bernoulli or geometric with unknown p .
Poisson with unknown mean λ .
- Denote the pdf of Y by $p_Y(y; \theta)$ to emphasize that there is a parameter θ .
- Do n independent trials to get data $y_1, y_2, y_3, \dots, y_n$. The joint pdf is

$$P_{Y_1, \dots, Y_n}(y_1, \dots, y_n; \theta) = P_Y(y_1; \theta) \cdots P_Y(y_n; \theta)$$

- Goal: Use the data to estimate θ .

Likelihood Function

- Previously, we knew the parameter θ and regarded the y 's as unknowns (occurring with certain probabilities).
- Define the likelihood of θ given data y_1, \dots, y_n to be

$$L(\theta; y_1, \dots, y_n) = P_{Y_1, \dots, Y_n}(y_1, \dots, y_n; \theta) = P_Y(y_1; \theta) \cdots P_Y(y_n; \theta)$$

- It's the exact same formula as the joint pdf; the difference is the interpretation. Now we consider the data y_1, \dots, y_n to be given and θ to be an unknown.

Maximum Likelihood Estimate (MLE)

- MLE definition: The value $\theta = \hat{\theta}$ that maximizes the likelihood is the Maximum Likelihood Estimate.
- We have reduced the problem of selecting the best model to that of selecting the best parameter.
- We want to select the parameter θ which will maximize the probability that the data was generated from the model with the parameter θ plugged-in.
- The parameter θ is called the maximum likelihood estimator.
- The maximum of the function can be obtained by setting the derivative of the function $=0$ and solving for θ .

- What is MLE?

Given

A sample $X = \{X_1, \dots, X_n\}$

A vector of parameters θ .

- We define

Likelihood of the data: $P(X|\theta)$

Log-likelihood of the data: $L(\theta) = \log P(X|\theta)$.

- Given X , find $\theta_{ML} = \arg \max L(\Theta), \theta \in \Omega$.

- Often we assume that X_i 's are independently identically distributed (i.i.d.)

$$\begin{aligned}\theta_{ML} &= \arg \max_{\theta \in \Omega} L(\theta) \\ &= \arg \max_{\theta \in \Omega} \log P(X|\theta) \\ &= \arg \max_{\theta \in \Omega} \log P(X_1, \dots, X_n|\theta) \\ &= \arg \max_{\theta \in \Omega} \log \prod_i P(X_i|\theta) \\ &= \arg \max_{\theta \in \Omega} \sum_i \log P(X_i|\theta)\end{aligned}$$

- Depending on the form of $p(x|\theta)$, solving optimization problem can be easy or hard.

Desireable properties of an estimator $\hat{\theta}$

- $\hat{\theta}$ should be narrowly distributed around the correct value of θ .
- Increasing n should improve the estimate.
- The distribution of $\hat{\theta}$ should be known.

MLE for Poisson distribution

- Y has a Poisson distribution with unknown parameter $\lambda \geq 0$.
- Collect data from independent trials:

$$Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n$$

- Likelihood:

$$L(\lambda; y_1, \dots, y_n) = \prod_{i=1}^n e^{-\lambda} \frac{\lambda^{y_i}}{y_i!} = \frac{e^{-n\lambda} \lambda^{y_1+y_2+\dots+y_n}}{y_1! y_2! \dots y_n!}$$

- Log likelihood is maximized at the same λ and is easier to use:

$$\ln L(\lambda; y_1, \dots, y_n) = -n\lambda + (y_1 + \dots + y_n) \ln \lambda - \ln(y_1! \dots y_n!)$$

- Critical point: Solve $d(\ln L)/d\lambda = 0$:

$$d(\ln L)/d\lambda = -n + \frac{y_1 + \dots + y_n}{\lambda} = 0$$

So

$$\lambda = \frac{y_1 + \dots + y_n}{n}.$$

- Check the second derivative is negative:

$$\frac{d^2 \ln L}{d\lambda^2} = -\frac{y_1 + \cdots + y_n}{\lambda^2} = -\frac{n^2}{y_1 + \cdots + y_n} < 0$$

provided $y_1 + \cdots + y_n > 0$. So it's a max unless $y_1 + \cdots + y_n = 0$.

- The exceptional case on the previous slide was $y_1 + \cdots + y_n = 0$, giving $y_1 = \cdots = y_n = 0$ (since all $y_i \geq 0$).
- In this case,

$$\begin{aligned} \ln L(\lambda; y_1, \dots, y_n) &= -n\lambda + (y_1 + \cdots + y_n) \ln \lambda - \ln(y_1! \cdots y_n!) \\ &= -n\lambda + 0 \ln \lambda - \ln(0! \cdots 0!) \\ &= -n\lambda \end{aligned}$$

- On the range $\lambda \geq 0$, this is maximized at $\hat{\lambda} = 0$, which agrees with $\hat{\lambda} = y_1 + \cdots + y_n = 0 + \cdots + 0 = 0$.

MLE for univariate Gaussian

- Suppose we have $Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n \sim (i.i.d)N(\mu, \sigma^2)$. But we don't know μ and σ^2 , $-\infty < \mu < \infty, \sigma^2 > 0$.
- For which μ, σ^2 is Y_1, Y_2, \dots, Y_n most likely?
- Likelihood:

$$L(\mu, \sigma^2; y_1, y_2, \dots, y_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2\right\}$$

- Log likelihood function is:

$$\ln L = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2$$

- Set $\frac{\partial \ln L}{\partial \mu} = 0, \frac{\partial \ln L}{\partial \sigma^2} = 0$, we have $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n y_i = \bar{Y}$,
 $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{Y})^2$.

A Multinomial Example

- Multinomial distribution:

$$P(X_1 = n_1, \dots, X_k = n_k) = \frac{n!}{n_1! n_2! \dots n_k!} p_1^{n_1} \dots p_k^{n_k}, \quad \sum_{i=1}^k n_i = n$$

- The observed data vector of frequencies:

$$y = (y_1, y_2, y_3, y_4)^T.$$

- Suppose the data are generated from a multinomial distribution with probabilities:

$$\frac{1}{2} + \frac{1}{4}\phi, \frac{1}{4}(1 - \phi), \frac{1}{4}(1 - \phi), \frac{1}{4}\phi$$

- Use MLE to estimate ϕ .

- The probability function is:

$$L(\phi; y) = \frac{n!}{y_1!y_2!y_3!y_4!} \left(\frac{1}{2} + \frac{1}{4}\phi\right)^{y_1} \left(\frac{1}{4}(1 - \phi)\right)^{y_2} \left(\frac{1}{4}(1 - \phi)\right)^{y_3} \left(\frac{1}{4}\phi\right)^{y_4}$$

- The log likelihood function apart from an additive term not involving ϕ is:

$$\log L(\phi) = y_1 \log(2 + \phi) + (y_2 + y_3) \log(1 - \phi) + y_4 \log \phi$$

•

$$\frac{\partial \log L(\phi)}{\partial \phi} = \frac{y_1}{2 + \phi} - \frac{y_2 + y_3}{1 - \phi} + \frac{y_4}{\phi}$$

$$l(\phi; y) = -\partial^2 \log L(\phi) / \partial \phi^2 = \frac{y_1}{(2 + \phi)^2} + \frac{y_2 + y_3}{(1 - \phi)^2} + \frac{y_4}{\phi^2}$$

- Suppose $y_1 = y_{11} + y_{12}$, where y_{11} and y_{12} have probabilities $\frac{1}{2}$ and $\frac{1}{4}\phi$.
- Suppose y_{11}, y_{12} are unobservable, we only observe their sum y_1 . Then the observed vector of frequencies y is viewed as being incomplete and the complete-data vector is taken to be

$$x = (y_{11}, y_{12}, y_2, y_3, y_4)^T.$$

They are assumed to arise from a multinomial distribution with probabilities

$$\frac{1}{2}, \frac{1}{4}\phi, \frac{1}{4}(1 - \phi), \frac{1}{4}(1 - \phi), \frac{1}{4}\phi$$

The log likelihood for the complete-data is:

$$\log L_c(\phi) = (y_{12} + y_4) \log \phi + (y_2 + y_3) \log(1 - \phi)$$

- Use MLE, we will get $\phi = \frac{y_{12}+y_4}{y_{12}+y_2+y_3+y_4}$.
- Since the frequency y_{12} is unobservable, we are unable to estimate ϕ .
- We can use an iterative method to estimate ϕ .
- For unobserved data, we fill in by averaging the complete-data log likelihood over its conditional distribution given the observed data y .

- Given a specified value of ϕ^0 , the conditional expectation of $\log L_c(\phi)$ can be written as:

$$Q(\phi; \phi^0) = E_{\phi^0}\{\log L_c(\phi)|y\}$$

- As $\log L_c(\phi)$ is a linear function of y_{11} , y_{12} , we can replace y_{12} by its current conditional expectation given the observed data y .
- The random variable Y_{11} corresponding to y_{11} has a binomial distribution with sample size y_1 and probability parameter $\frac{1}{2}/(\frac{1}{2} + \frac{1}{2}\phi^0)$.

$$E_{\phi^0}(Y_{11}|y_1) = y_{11}^0 = \frac{1}{2}y_1/(\frac{1}{2} + \frac{1}{4}\phi^0)$$

$$y_{12}^0 = y_1 - y_{11}^0 = \frac{1}{4}y_1\phi^0/(\frac{1}{2} + \frac{1}{4}\phi^0)$$

M-step: Maximization Q , we get

$$\phi^1 = \frac{y_{12}^0 + y_4}{y_{12}^0 + y_2 + y_3 + y_4}$$

Iteration steps:

$$\phi^{k+1} = (y_{12}^k + y_4)/(n - y_{11}^k)$$

where

$$y_{11}^k = \frac{1}{2}y_1/(\frac{1}{2} + \frac{1}{2}\phi^k)$$

$$y_{12}^k = y_1 - y_{11}^k$$

Expectation Maximization (EM)

- A parameter estimation method: it falls into the general framework of maximum-likelihood estimation (MLE).
- The general form was given in (Dempster, Laird, and Rubin, 1977), although essence of the algorithm appeared previously in various forms.
- The EM algorithm is used to find the maximum likelihood parameters of a statistical model in cases where the equations cannot be solved directly.
- Typically these models involve latent variables in addition to unknown parameters and known data observations. That is, either there are missing values among the data, or the model can be formulated more simply by assuming the existence of additional unobserved data points.

- Given a statistical model which generates a set \mathbf{X} of observed data, a set of unobserved latent data or missing values \mathbf{Z} , and a vector of unknown parameters θ , along with a likelihood function $L(\theta; \mathbf{X}, \mathbf{Z}) = p(\mathbf{X}, \mathbf{Z}|\theta)$.
- The maximum likelihood estimate (MLE) of the unknown parameters is determined by the marginal likelihood of the observed data.

$$L(\theta; \mathbf{X}) = p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta)$$

- The EM algorithm seeks to find the MLE of the marginal likelihood by iteratively applying the following two steps:
Expectation step (E-step): Calculate the expected value of the log likelihood function, with respect to the conditional distribution of \mathbf{Z} given \mathbf{X} under the current estimate of the parameters $\theta^{(t)}$:

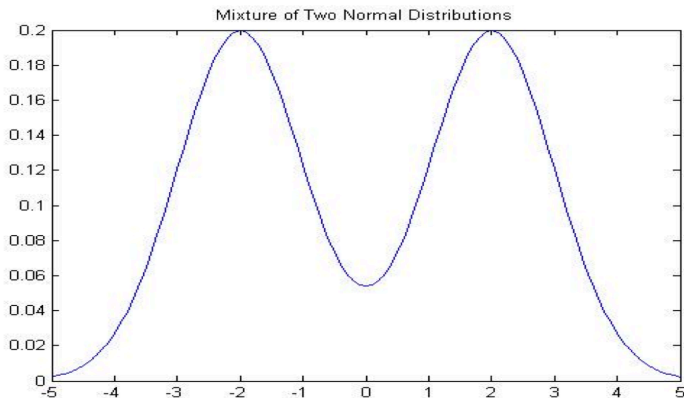
$$Q(\theta|\theta^{(t)}) = E_{\mathbf{Z}|\mathbf{X},\theta^{(t)}} [\log L(\theta; \mathbf{X}, \mathbf{Z})]$$

Maximization step (M-step): Find the parameter that maximizes this quantity:

$$\theta^{(t+1)} = \arg \max_{\theta} Q(\theta|\theta^{(t)})$$

Example: Gaussian mixture distribution

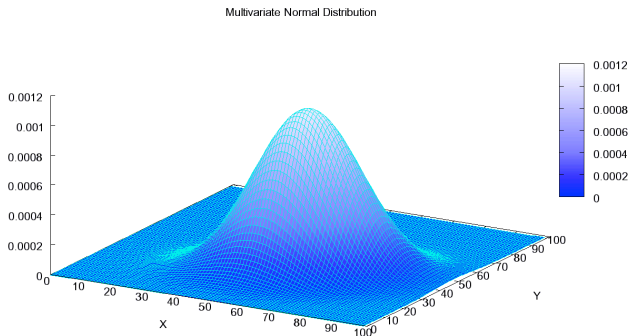
$$f(\pi_1, \pi_2, \mu_1, \mu_2, \sigma_1, \sigma_2) = \pi_1 N(\mu_1, \sigma_1^2) + \pi_2 N(\mu_2, \sigma_2^2)$$



Example: Gaussian mixture distribution

The pdf of the multivariate normal distribution:

$$f_{\mathbf{x}}(x_1, \dots, x_k) = \frac{1}{\sqrt{(2\pi)^k |\mathbf{\Sigma}|}} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$



Example: Gaussian mixture distribution

- When we proceed to calculate the MLE for a mixture, the presence of the sum of the distributions prevents a “neat” factorization using the log function.
- A completely new rethink is required to estimate the parameter.
- The new rethink also provides a solution to the clustering problem, which can be seen as a problem of estimating missing data. The missing data are the cluster labels.
- Clustering is only one example of a missing data problem. Several other problems can be formulated as missing data problems.

- Let $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ be a sample of n independent observations from a mixture of two multivariate normal distributions of dimension d , and let $\mathbf{z} = (z_1, z_2, \dots, z_n)$ be the latent variables that determine the component from which the observation originates.

$$\mathbf{x}_i | (z_i = 1) \sim N_d(\boldsymbol{\mu}_1, \Sigma_1) \text{ and } \mathbf{x}_i | (z_i = 2) \sim N_d(\boldsymbol{\mu}_2, \Sigma_2)$$

where

$$P(z_i = 1) = \tau_1 \text{ and } P(z_i = 2) = \tau_2 = 1 - \tau_1$$

- Let $f(\mathbf{x}; \boldsymbol{\mu}_j, \Sigma_j)$ be the pdf of $N_d(\boldsymbol{\mu}_1, \Sigma_1)$, $N_d(\boldsymbol{\mu}_2, \Sigma_2)$.

$$f(\mathbf{x}) = \tau_1 f(\mathbf{x}; \boldsymbol{\mu}_1, \Sigma_1) + \tau_2 f(\mathbf{x}; \boldsymbol{\mu}_2, \Sigma_2)$$

- The aim is to estimate the unknown parameters:

$$\theta = (\boldsymbol{\tau}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2)$$

- The incomplete-data likelihood function is

$$L(\theta; \mathbf{x}) = P(\mathbf{x}|\theta) = \prod_{i=1}^n \sum_{j=1}^2 \tau_j f(\mathbf{x}_i; \boldsymbol{\mu}_j, \Sigma_j),$$

Its log-likelihood function is:

$$\log L(\theta; \mathbf{x}) = \sum_{i=1}^n \log \sum_{j=1}^2 \tau_j f(\mathbf{x}_i; \boldsymbol{\mu}_j, \Sigma_j).$$

- We consider the unobserved latent variables $z_i = 1$ or 0 . We use $z_{ij} = 1$ to denote the i th sample belongs to the j distribution, $z_{ij} = 0$ to denote the i th sample does not belong to the j distribution, $j = 1, 2$.
- The complete-data likelihood function is

$$L(\theta; \mathbf{x}, \mathbf{z}) = P(\mathbf{x}, \mathbf{z} | \theta) = \prod_{i=1}^n \prod_{j=1}^2 (\tau_j f(\mathbf{x}_i; \boldsymbol{\mu}_j, \Sigma_j))^{z_{ij}}$$

Its log-likelihood function is:

$$\log L(\theta; \mathbf{x}, \mathbf{z}) = \sum_{i=1}^n \sum_{j=1}^2 z_{ij} \left[\log \tau_j - \frac{1}{2} \log |\Sigma_j| - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_j)^\top \Sigma_j^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_j) - \frac{d}{2} \log(2\pi) \right].$$

where \mathbb{I} is an indicator function and f is the probability density function of a multivariate normal.

E-step:

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= E(\log L(\theta; \mathbf{x}, \mathbf{z})) \\ &= E\left(\sum_{i=1}^n \sum_{j=1}^2 z_{ij} \left[\log \tau_j - \frac{1}{2} \log |\Sigma_j| - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_j)^\top \Sigma_j^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_j) - \frac{d}{2} \log(2\pi) \right]\right) \\ &= \sum_{i=1}^n \sum_{j=1}^2 T_{j,i}^{(t)} \left[\log \tau_j - \frac{1}{2} \log |\Sigma_j| - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_j)^\top \Sigma_j^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_j) - \frac{d}{2} \log(2\pi) \right] \end{aligned}$$

Given our current estimate of the parameters $\theta^{(t)}$, the conditional distribution of the Z_i is determined by Bayes theorem to be the proportional height of the normal density weighted by τ :

$$T_{j,i}^{(t)} := P(z_{ij} = 1 | X_i = \mathbf{x}_i; \theta^{(t)}) = \frac{\tau_j^{(t)} f(\mathbf{x}_i; \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})}{\tau_1^{(t)} f(\mathbf{x}_i; \boldsymbol{\mu}_1^{(t)}, \boldsymbol{\Sigma}_1^{(t)}) + \tau_2^{(t)} f(\mathbf{x}_i; \boldsymbol{\mu}_2^{(t)}, \boldsymbol{\Sigma}_2^{(t)})}.$$

These are called the “membership probabilities” which are normally considered the output of the E-step.

M-step:

The fact that $Q(\theta|\theta(t))$ is quadratic in form means that determining the maximizing values of θ is relatively straightforward.

To begin, consider τ , which has the constraint $\tau_1 + \tau_2 = 1$:

$$\begin{aligned}\tau^{(t+1)} &= \arg \max_{\tau} Q(\theta|\theta^{(t)}) \\ &= \arg \max_{\tau} \left\{ \left[\sum_{i=1}^n T_{1,i}^{(t)} \right] \log \tau_1 + \left[\sum_{i=1}^n T_{2,i}^{(t)} \right] \log \tau_2 \right\}\end{aligned}$$

This has the same form as the MLE for the binomial distribution, so

$$\tau_j^{(t+1)} = \frac{\sum_{i=1}^n T_{j,i}^{(t)}}{\sum_{i=1}^n (T_{1,i}^{(t)} + T_{2,i}^{(t)})} = \frac{1}{n} \sum_{i=1}^n T_{j,i}^{(t)}.$$

For the estimates of (μ_1, σ_1) :

$$\begin{aligned}(\mu_1^{(t+1)}, \Sigma_1^{(t+1)}) &= \arg \max_{\mu_1, \Sigma_1} Q(\theta | \theta^{(t)}) \\&= \arg \max_{\mu_1, \Sigma_1} \sum_{i=1}^n T_{1,i}^{(t)} \left\{ -\frac{1}{2} \log |\Sigma_1| - \frac{1}{2} (\mathbf{x}_i - \mu_1)^\top \Sigma_1^{-1} (\mathbf{x}_i - \mu_1) \right\}\end{aligned}$$

This has the same form as a weighted MLE for a normal distribution, so

$$\mu_1^{(t+1)} = \frac{\sum_{i=1}^n T_{1,i}^{(t)} \mathbf{x}_i}{\sum_{i=1}^n T_{1,i}^{(t)}} \text{ and } \Sigma_1^{(t+1)} = \frac{\sum_{i=1}^n T_{1,i}^{(t)} (\mathbf{x}_i - \mu_1^{(t+1)}) (\mathbf{x}_i - \mu_1^{(t+1)})^\top}{\sum_{i=1}^n T_{1,i}^{(t)}}$$

and, by symmetry

$$\mu_2^{(t+1)} = \frac{\sum_{i=1}^n T_{2,i}^{(t)} \mathbf{x}_i}{\sum_{i=1}^n T_{2,i}^{(t)}} \text{ and } \Sigma_2^{(t+1)} = \frac{\sum_{i=1}^n T_{2,i}^{(t)} (\mathbf{x}_i - \mu_2^{(t+1)}) (\mathbf{x}_i - \mu_2^{(t+1)})^\top}{\sum_{i=1}^n T_{2,i}^{(t)}}.$$

Example: Multinomial with complex cell structure

- Suppose there are four cells (O,A,B,AB) with probability $r^2, p^2 + 2pr, q^2 + 2qr, 2pq$.
- The observed data are: $y = (n_O, n_A, n_B, n_{AB})^T$.
- The unknown parameter vector is: $\theta = (p, q)^T, r = 1 - p - q$.
- The log likelihood function for θ , apart from an additive constant, is

$$\log L(\theta) = 2n_O \log r + n_A \log(p^2 + 2pr) + n_B \log(q^2 + 2qr) + n_{AB} \log(2pq).$$

- The complete data vector is $x = (n_O, n_{AB}, z^T)^T$, where $z = (n_{AA}, n_{AO}, n_{BB}, n_{BO})^T$.
- The corresponding probabilities can be set to be $r^2, 2pq, p^2, 2pr, q^2, 2qr$.
- The complete data log likelihood function can be written as (apart from an additive constant):

$$\log L_c(\theta) = 2n_A^+ \log p + 2n_B^+ \log q + 2n_O^+ \log r,$$

where

$$n_A^+ = n_{AA} + \frac{1}{2}n_{AO} + \frac{1}{2}n_{AB}$$

$$n_B^+ = n_{BB} + \frac{1}{2}n_{BO} + \frac{1}{2}n_{AB}$$

$$n_O^+ = n_O + \frac{1}{2}n_{AO} + \frac{1}{2}n_{BO}.$$

- The complete data MLE gives:

$$\hat{p} = \frac{n_A^+}{n}; \hat{q} = \frac{n_B^+}{n}.$$

- Conditional on y , n_{AA} has a binomial distribution with sample size n_A and probability parameter $p^{(k)^2} / (p^{(k)^2} + 2p^{(k)}r^{(k)})$.
- E-step:

$$E_{\theta^{(k)}}(n_{AA}) = n_A p^{(k)^2} / (p^{(k)^2} + 2p^{(k)}r^{(k)})$$

Similarly, we get the conditional expectations of n_{AO} , n_{BB} , n_{BO} and n_{BB} .

- M-step:

$$p^{(k+1)} = (n_{AA}^{(k)} + \frac{1}{2}n_{AO}^{(k)} + \frac{1}{2}n_{AB})/n$$

$$q^{(k+1)} = (n_{BB}^{(k)} + \frac{1}{2}n_{BO}^{(k)} + \frac{1}{2}n_{AB})/n$$

Derivation of the EM algorithm

- The EM algorithm is an iterative procedure for maximizing $L(\theta; X)$, where X is the observed data.
- Assume that after the k th iteration the current estimate for θ is θ_k .
- We wish to compute an updated estimate θ , such that $L(\theta) > L(\theta_k)$.
- Equivalently we want to maximize the difference: $L(\theta) - L(\theta_k)$.
- Hidden variables may be introduced purely as an artifice for making the maximum likelihood estimation of θ tractable.
- In this case, it is assumed that knowledge of the hidden variables will make the maximization of the likelihood function easier.

- Denote the hidden random vector by Z and a given realization by z .
- The total probability $P(X|\theta)$ may be written in terms of the hidden variables z as:

$$P(X|\theta) = \sum_z P(X|z, \theta)P(z|\theta)$$

- We directly take the log of the likelihood function and still use L to denote the function.

$$L(\theta) - L(\theta_k) = \log\left(\sum_z P(X|z, \theta)P(z|\theta)\right) - \log P(X|\theta_k)$$

- (Jensen's inequality) Let f be a convex function defined on an interval I . If $x_1, x_2, \dots, x_n \in I$ and $\lambda_1, \lambda_2, \dots, \lambda_n \geq 0$ with $\sum_i \lambda_i = 1$, we have

$$\sum_i \lambda_i f(x_i) \geq f\left(\sum_i \lambda_i x_i\right)$$

- Using Jensen's inequality, let $f = \log$, we have

$$\sum_i \lambda_i \log(x_i) \geq \log\left(\sum_i \lambda_i x_i\right)$$

$$\begin{aligned}
& L(\theta) - L(\theta_k) \\
= & \log\left(\sum_z P(X|z, \theta)P(z|\theta)\right) - \log P(X|\theta_k) \\
= & \log\left(\sum_z P(X|z, \theta)P(z|\theta) \cdot \frac{P(z|X, \theta_k)}{P(z|X, \theta_k)}\right) - \log P(X|\theta_k) \\
= & \log\left(\sum_z P(z|X, \theta_k) \cdot \frac{P(X|z, \theta)P(z|\theta)}{P(z|X, \theta_k)}\right) - \log P(X|\theta_k) \\
\geq & \sum_z P(z|X, \theta_k) \log\left(\frac{P(X|z, \theta)P(z|\theta)}{P(z|X, \theta_k)}\right) - \log P(X|\theta_k) \\
= & \sum_z P(z|X, \theta_k) \log\left(\frac{P(X|z, \theta)P(z|\theta)}{P(z|X, \theta_k)P(X|\theta_k)}\right) \\
= & \Delta(\theta|\theta_k)
\end{aligned}$$

- Define $l(\theta|\theta_k) = L(\theta_k) + \Delta(\theta|\theta_k)$, we have $L(\theta) > l(\theta|\theta_k)$.

$$\begin{aligned} l(\theta_k|\theta_k) &= L(\theta_k) + \sum_z P(z|X, \theta_k) \log\left(\frac{P(X|z, \theta_k)P(z|\theta_k)}{P(z|X, \theta_k)P(X|\theta_k)}\right) \\ &= L(\theta_k) \end{aligned}$$

- Therefore, any θ which increases $l(\theta|\theta_k)$ will also increase $L(\theta)$.
- In order to achieve the greatest possible increase in the value of $L(\theta)$, the EM algorithm calls for selecting θ such that $l(\theta|\theta_k)$ is maximized.

$$\begin{aligned}
\theta_{k+1} &= \arg \max_{\theta} l(\theta|\theta_k) \\
&= \arg \max_{\theta} L(\theta_k) + \sum_z P(z|X, \theta_k) \log\left(\frac{P(X|z, \theta)P(z|\theta)}{P(z|X, \theta_k)P(X|\theta_k)}\right) \\
&= \arg \max_{\theta} \sum_z P(z|X, \theta_k) \log(P(X|z, \theta)P(z|\theta)) \\
&= \arg \max_{\theta} \sum_z P(z|X, \theta_k) \log \frac{P(X, z, \theta)P(z, \theta)}{P(z, \theta)P(\theta)} \\
&= \arg \max_{\theta} \sum_z P(z|X, \theta_k) \log \frac{P(X, z, \theta)}{P(\theta)} \\
&= \arg \max_{\theta} \sum_z P(z|X, \theta_k) \log P(X, z|\theta) \\
&= \arg \max_{\theta} E_{Z|X, \theta} \log P(X, z|\theta)
\end{aligned}$$

