09 Finding MLE.

(1) EM algorithm.

When do we use EM?

- when we have missing values or assume hidden (latent) parameter for problem simplification
- when maximising incomplete-data log likelihood directly is hard.

Algorithm:

let $L(\theta|y,z) = f(y,z|\theta)$ → complete-data likelihood.
(y: observation: known and fixed; z: latent variables: unknown r.v.; θ: unknown fixed)

So $L(\theta|y,z)$ is a function of z: $h_{\theta,y}(z)$.

(1) Start with $\theta^{(0)}$

(2) Given the current $\theta^{(r)}$, $r=1,2,\dots$

E-step:

$$Q(\theta,\theta^{(r)}) = E[\ln f(y,z|\theta)\,|\,y,\theta^{(r)}] \rightarrow \text{expected value of complete-data likelihood}$$

$$= \int_z \ln f(y,z|\theta)\, k(z|y,\theta^{(r)})\,dz$$

(k: conditional pdf of latent data depending on current value of $\theta^{(r)}$ and observed data.)

recall: $E[h(Y_2)|X=x] = \int_y h(y) f(y|x)\,dy$.

M-step:

$\theta^{(r+1)} = \arg\max_\theta Q(\theta,\theta^{(r)})$ [if $\theta^{(r+1)}$ s.t. $Q(\theta^{(r+1)},\theta^{(r)}) > Q(\theta^{(r)},\theta^{(r)})$ ⇒ it's called GEM].

(3) The iteration continues until $\|\theta^{(r+1)} - \theta^{(r)}\|$ or $|Q(\theta^{(r+1)},\theta^{(r)}) - Q(\theta^{(r)},\theta^{(r)})|$ is smaller than tolerance level $\varepsilon > 0$, e.g. 10^{-6}.

Why does it work?

- Each iteration increases $\log L$
  i.e. $\ln g(y|\theta^{(r+1)}) \geq \ln g(y|\theta^{(r)})$
- and the EM algorithm increases the incomplete-log likelihood.
  i.e. if θ'' is found s.t. $Q(\theta'',\theta') = \max_\theta Q(\theta,\theta')$ ⇒ $\ln L(\theta''|y) \geq \ln L(\theta'|y)$

Example:

$Y \sim$ multinomial(P_1,P_2,P_3,P_4) with $(P_1,P_2,P_3,P_4) = (\frac{1}{2}+\frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4})$ with $0\leq\theta\leq 1$.
and $y = (y_1,y_2,y_3,y_4) = (125, 18, 20, 34)$.

$X \sim$ multinomial(P_1,P_2,P_3,P_4,P_5) with $(P_1,P_2,P_3,P_4,P_5) = (\frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4})$
and $x = (x_1,x_2,x_3,x_4,x_5)$ ⇒ $x_1+x_2 = y_1$, $x_3 = y_2$, $x_4 = y_3$, $x_5 = y_4$.

$$\Rightarrow L(\theta|x) = \frac{(x_1+x_2+\dots+x_5)!}{x_1!\,x_2!\cdots x_5!}\left(\frac{1}{2}\right)^{x_1}\left(\frac{\theta}{4}\right)^{x_2}\left(\frac{1-\theta}{4}\right)^{x_3+x_4}\left(\frac{\theta}{4}\right)^{x_5}$$

$$\Rightarrow l(\theta|x) = C + \underline{x_1 \ln\tfrac{1}{2}} + (x_2+x_5)\ln\frac{\theta}{4} + (x_3+x_4)\ln\left(\frac{1-\theta}{4}\right) = C + (x_2+x_5)\ln\theta + (x_3+x_4)\ln(1-\theta).$$

$$\Rightarrow Q(\theta,\theta^{(r)}) = E[l(\theta|x)\,|\,y,\theta^{(r)}] = E[(x_2+x_5)\ln\theta + (x_3+x_4)\ln(1-\theta)\,|\,\theta^{(r)},y].$$

$$= E[x_2\ln\theta\,|\,y,\theta^{(r)}] + y_4\ln\theta + (y_2+y_3)\ln(1-\theta)$$

$$= \frac{\frac{\theta^{(r)}}{4}}{\frac{1}{2}+\frac{\theta^{(r)}}{4}}\,y_1\ln\theta + y_4\ln\theta + (y_2+y_3)\ln(1-\theta)$$

$$= \left[\frac{\theta^{(r)}}{2+\theta^{(r)}}\,y_1 + y_4\right]\ln\theta + (y_2+y_3)\ln(1-\theta)$$

To maximise $Q(\theta,\theta^{(r)})$:

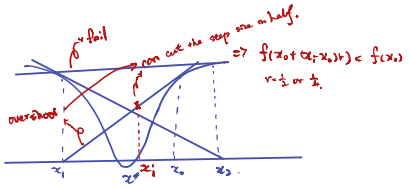
$$\text{let } \frac{\partial}{\partial\theta}Q(\theta,\theta^{(r)}) = \frac{1}{\theta}\left[\frac{\theta^{(r)}}{2+\theta^{(r)}}\,y_1 + y_4\right] - \frac{1}{1-\theta}(y_2+y_3) = 0 \quad \Rightarrow (1-\theta)\left(\frac{\theta^{(r)}}{2+\theta^{(r)}}\,y_1 + y_4\right) = \theta(y_2+y_3).$$

$$\Rightarrow \theta^{(r+1)} = \frac{y_1\theta^{(r)}/(2+\theta^{(r)}) + y_4}{y_1\theta^{(r)}/(2+\theta^{(r)}) + y_4 + y_2 + y_3}.$$

(2) Newton-Raphson (faster but less stable numerically and less tractable analytically). when will it fail?

formular:

for 1-dimension: $x^{(r+1)} = x^{(r)} - \frac{f(x^{(r)})}{f'(x^{(r)})}$ to solve $f(x)=0$



$\Rightarrow f(x_0 + \alpha_r x_r) < f(x_0)$ $r = \frac{1}{2}$ or $\frac{1}{4}$

$|f'(x)|^{-1} \le s$ for all $x^* - d < x < x^* + d$.

for higher dimension:

let $x_n = (x_1, \dots x_n)^T$

$\theta = (\theta_1, \dots \theta_q)^T$ for q parameters.

$l(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln f(x_i|\theta)$.

$U(\theta) = \frac{\partial}{\partial \theta} l(\theta)$ → score function, is a $q \times 1$ vector.

$H(\theta) = \frac{\partial}{\partial \theta} U(\theta) = \frac{\partial^2}{\partial \theta^2} l(\theta)$ → Hessian function, is a $q \times q$ vector. $\begin{pmatrix} \frac{\partial^2}{\partial \theta_1^2} l(\theta) & \cdots & \frac{\partial^2}{\partial \theta_1 \partial \theta_q} l(\theta) \\ \vdots & & \\ \frac{\partial^2}{\partial \theta_q \partial \theta_1} & \cdots & \frac{\partial^2}{\partial \theta_q^2} l(\theta) \end{pmatrix}$

$J(\theta) = -H(\theta)$ → observed information function.

$I(\theta) = E(J(\theta)) = E[-H(\theta)]$ → fisher information function.

(1) Then start with $\theta^{(0)}$,

(2) for $k = 0, 1, 2, \dots$

$\theta^{(k+1)} = \theta^{(k)} - [H(\theta^{(k)})]^{-1} U(\theta^{(k)})$ if $\theta^{(k+1)} = \theta^{(k)} + [I(\theta^{(k)})]^{-1} U(\theta^{(k)})$ => Fisher-scoring

$\cong \theta^{(k)} + [J(\theta^{(k)})]^{-1} U(\theta^{(k)})$ to solve $Q(\theta) = 0$

(3) The iteration continues until $\|\theta^{(r+1)} - \theta^{(r)}\|$ or $|Q(\theta^{(r+1)}, \theta^{(r)}) - Q(\theta^{(r)}, \theta^{(r)})|$ is smaller than tolerance level $\varepsilon > 0$, e.g. 10^{-6}.

note:

- need to check $H(\theta') < 0$ for θ' s.t. $U(\theta') = 0$.
- NR can be extended to BFGS, BHHH, Nelder-Mead and simulated annealing etc.

Example continued:

The observed data log-likelihood is:

$l_Y(\theta) = \ln C_1 + y_1 \ln(2+\theta) + (y_2 + y_3) \ln(1-\theta) + y_4 \ln \theta$.

and

$\frac{\partial}{\partial \theta} l_Y(\theta) = \frac{y_1}{2+\theta} - \frac{y_2 + y_3}{1-\theta} + \frac{y_4}{\theta}$

let $U(\theta) = \frac{\partial}{\partial \theta} l_Y(\theta) = \frac{y_1}{2+\theta} - \frac{y_2 + y_3}{1-\theta} + \frac{y_4}{\theta}$

$\Rightarrow H(\theta) = \frac{\partial^2}{\partial \theta^2} l_Y(\theta) = \frac{-y_1}{(2+\theta)^2} - \frac{y_2 + y_3}{(1-\theta)^2} - \frac{y_4}{\theta^2}$

$\Rightarrow I(\theta) = E[-H(\theta)] = E\left[\frac{y_1}{(2+\theta)^2} + \frac{y_2 + y_3}{(1-\theta)^2} + \frac{y_4}{\theta^2}\right] = \frac{\frac{(2+\theta)}{4} \cdot n}{(2+\theta)^2} + \frac{\frac{1-\theta}{4} \cdot 2 \cdot n}{(1-\theta)^2} + \frac{\frac{\theta}{4} n}{\theta^2}$

$= n\left(\frac{1}{4(2+\theta)} + \frac{1}{2(1-\theta)} + \frac{1}{4\theta}\right)$

$\Rightarrow \theta^{(r+1)} = \theta^{(r)} - \frac{U(\theta^{(r)})}{H(\theta^{(r)})}$ or $\theta^{(r+1)} = \theta^{(r)} + \frac{U(\theta^{(r)})}{I(\theta^{(r)})}$ of fisher scoring.

(3) maxLik() in R.

(4) Estimation of variance of MLE.

(1) Using asymptotic efficiency of MLE,

- $Var(\hat{\theta}) \approx (J(\hat{\theta}))^{-1} = \left[-\frac{\partial^2}{\partial \theta \partial \theta^T} l(\hat{\theta})\right]^{-1}$ or $\approx (I(\hat{\theta}))^{-1} = E\left[-\frac{\partial^2}{\partial \theta \partial \theta^T} l(\hat{\theta})\right]^{-1}$

  s.e.$(\hat{\theta}_i) = \sqrt{Var(\hat{\theta})_{ii}}$, $i = 1, \dots q$ i.e. the i^{th} diagonal of $Var(\hat{\theta})$.

(2) Using Louis method,

- $Var(\hat{\theta}) \approx (J(\hat{\theta}))^{-1}$ where $J(\hat{\theta}) = -H(\hat{\theta}) = -\frac{\partial^2 Q(\theta, \phi)}{\partial \theta \partial \theta^T}\Big|_{\phi=\hat{\theta}}$ (→ $\frac{\partial^2}{\partial \theta \partial \theta^T}$ expected complete data log-likelihood.) $- Var\left[\frac{\partial l(\theta|y,z)}{\partial \theta} | y, \phi\right]\Big|_{\phi=\hat{\theta}}$ (→ $\frac{\partial}{\partial \theta}$ complete data log-likelihood.)

example continued:

$$-\frac{\partial^2 Q(\theta,\hat{\theta})}{\partial\theta^2}\Big|_{\theta=\hat{\theta}} = -\frac{\partial^2}{\partial\theta^2}\left[\frac{\hat{\theta}}{2+\hat{\theta}}y_1 + y_4\right]\ln\theta + (y_2+y_3)\ln(1-\theta)$$

$$= \frac{1}{\hat{\theta}^2}\left(\frac{\hat{\theta}}{2+\hat{\theta}}y_1 + y_4\right) + \frac{1}{(1-\hat{\theta})^2}(y_2+y_3). \quad \cdots (1)$$

(replace $\theta^{(r)}$ with $\hat{\theta}$ in $Q(\theta,\theta^{(m)})$; and then differentiate with respect to θ; then replace θ with $\hat{\theta}$).

$$\frac{\partial}{\partial\theta}\ell(\theta|y,z) = \frac{\partial}{\partial\theta}\left[C + (x_2+x_5)\ln\theta + (x_3+x_4)\ln(1-\theta)\right] = \frac{x_2+x_5}{\theta} - \frac{x_3+x_4}{1-\theta}.$$

$$\frac{\partial^2}{\partial\theta^2}\ell(\theta|y,z) = -\frac{x_2+x_5}{\theta^2} - \frac{x_3+x_4}{(1-\theta)^2}$$

$$\Rightarrow -\mathrm{Var}\left[\frac{\partial\ell(\theta|y,z)}{\partial\theta}\,|\,y,\phi\right]\Big|_{\phi=\hat{\theta}} = -\mathrm{Var}\left[\frac{x_2+x_5}{\theta^2} + \frac{x_3+x_4}{(1-\theta)^2}\,|\,y,\hat{\theta}\right] = -\mathrm{Var}\left(\frac{x_2}{\theta^2}\,|\,y,\hat{\theta}\right) = -\frac{y_1\frac{\hat{\theta}}{\hat{\theta}+2}\left(1-\frac{\hat{\theta}}{\hat{\theta}+2}\right)}{\hat{\theta}^2}$$

$$= \frac{-2y_1}{\hat{\theta}(\hat{\theta}+2)^2} \quad \cdots (2)$$

$$\Rightarrow \mathrm{Var}(\hat{\theta}) \approx [(1)+(2)]^{-1}$$