EM Algorithm - Exponential Families

Exponential Family $\exp\left(A(\theta) + h(x) + t(\theta)s(x)\right)$

$A(\theta)$: normalizing constant

$s(x)$: suff. stat.

$t(\theta)$: corresponding parameter for suff. stat.

- The computation of these two steps simplify a great deal when it can be shown that the log-likelihood is linear in the sufficient statistic for θ (Exponential Family).

For EM algorithm, if your dist belongs to exponential family, derivation of EM algorithm is much easier.

- The E-step reduces to computing the expectation of the complete-data sufficient statistic given the observed data.

E-step: Calculate the expectation of complete data suff. stat given the observed data

- In M-step, the conditional expectations of the sufficient statistics computed in the E-step can be directly substituted for the sufficient statistics that occur in the expressions obtained for the complete-data maximum likelihood estimators of θ, to obtain the next iterate.

Exponential family : MLE is a function of suff. stat.

We plug-in the result of E-step into suff. stat.

If density cannot find MLE analytically, we have to use optimization

=> density belongs to exponential family. (we don't need)

Example: Univariate Normal Distribution

Complete data $y = (y_{obs}, y_{miss})$ $\quad y \sim N(\mu, \sigma^2)$

$$L(\mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\left(\sum y_i^2 - 2\mu\sum y_i + n\mu^2\right)\right)$$

suff. stat

Let the complete-data vector $y = (y_1, \cdots, y_n)$ be a random sample from $N(\mu, \sigma^2)$. Then,

$$f(y|\mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum y_i^2 - 2\mu\sum y_i + n\mu^2\right)\right\}.$$

which implies that $\left(\sum y_i, \sum y_i^2\right)$ are sufficient statistics for $\theta = (\mu, \sigma^2)$. The complete-data log-likelihood function is

$$l(\mu.\sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum y_i^2 + \frac{\mu}{\sigma^2}\sum y_i - \frac{n\mu^2}{\sigma^2} + \text{constant}.$$

Example: Univariate Normal Distribution

Suppose $y_i, i = 1, \cdots, m$ are observed and $y_i, i = m+1, \cdots, n$ are missing.

Denote the observed data by $y_{obs} = (y_1, \cdots, y_m)$.

- The E-step requires the computation of

$$E_\theta\left(\sum y_i | y_{obs}\right) \quad \text{and} \quad E_\theta\left(\sum y_i^2 | y_{obs}\right).$$

- At the t-th iteration of the E-step, compute

$$s_1^{(t)} = E_\theta\left(\sum y_i | y_{obs}\right) = \sum_{i=1}^{m} y_i + (n-m)\mu^{(t)}$$

$$s_2^{(t)} = E_\theta\left(\sum y_i^2 | y_{obs}\right) = \sum_{i=1}^{m} y_i^2 + (n-m)\left[\sigma^{2(t)} + \mu^{(t)2}\right]$$

Example: Univariate Normal Distribution

M-Step : Normal ~ Exponential family.

$\hat{\mu} = \frac{1}{n}\sum y_i$ ↳ $s_1^{(t)}$ $\quad \hat{\sigma}^2 = \frac{1}{n}\sum y_i^2 - \left(\frac{1}{n}\sum y_i\right)^2 = \frac{1}{n}\sum y_i^2 - \hat{\mu}^2$ ↳ $s_2^{(t)}$

- The complete-data maximum likelihood estimation of μ and σ^2 are

$$\hat{\mu} = \frac{\sum_{i=1}^n y_i}{n} \quad \text{and} \quad \hat{\sigma^2} = \frac{\sum_{i=1}^n y_i^2}{n} - \left(\frac{\sum_{i=1}^n y_i}{n}\right)^2.$$

- At the t-th iteration of the M-step, compute

$$\mu^{(t+1)} = \frac{s_1^{(t)}}{n}$$

$$\sigma^{2(t+1)} = \frac{s_2^{(t)}}{n} - \mu^{(t+1)2}.$$

$\mu^{(t+1)} = \frac{1}{n} s_1^{(t)}$.

$\sigma^{2(t+1)} = \frac{1}{n} s_2^{(t)} - \left(\mu^{(t+1)}\right)^2$.

Example: Multinomial Distribution

EM (1977)

$y = (y_1, y_2, y_3, y_4)$

$(p_1, p_2, p_3, p_4) = \left(\frac{1}{2} - \frac{1}{2}\theta, \frac{1}{4}\theta, \frac{1}{4}\theta, \frac{1}{2}\right)$

We want to estimate θ.

- Define $y = (y_1, y_2, y_3, y_4)$ with multinomial probability $\left(\frac{1}{2} - \frac{1}{2}\theta, \frac{1}{4}\theta, \frac{1}{4}\theta, \frac{1}{2}\right) = (p_1, p_2, p_3, p_4)$.
- Let $y_{obs} = (y_1, y_2, y_3 + y_4) = (38, 34, 125)$ be observed counts from a multinomial population, and then, observed data vector is a function of the complete-data vector.
- Since only $y_3 + y_4$ is observed and y_3 and y_4 are not, the observed data is considered incomplete.
- The complete-data log-likelihood is

Complete-data log likelihood not completely observed.

What we observed; $(y_1, y_2, y_3+y_4) = (38, 34, 125)$

genotype
phenotype

$l(\theta) = y_1 \log p_1 + y_2 \log p_2 + y_3 \log p_3 + y_4 \log p_4 + \text{constant}$

$p_1 + p_2 + p_3 + p_4 = 1$

$$l(\theta) = y_1 \log p_1 + y_2 \log p_2 + y_3 \log p_3 + y_4 \log p_4 + \text{constant}.$$

$= y_1 \log\left(\frac{1}{2} - \frac{1}{2}\theta\right) + y_2 \log \frac{1}{4}\theta + y_3 \log \frac{1}{4}\theta + y_4 \log \frac{1}{2} + \text{constant}$

where y_1, y_2, y_3, y_4 are sufficient statistics.

$\frac{\partial l(\theta)}{\partial \theta} = -\frac{y_1}{1-\theta} + \frac{y_2}{\theta} + \frac{y_3}{\theta} = 0 \Rightarrow \frac{y_1}{1-\theta} = \frac{y_2+y_3}{\theta} \Rightarrow (y_2+y_3)(1-\theta) = y_1\theta$

$(y_1+y_2+y_3)\theta = y_2+y_3$

Example: Multinomial Distribution

- E-step:

$$E_\theta(y_1|y_{obs}) = y_1 = 38, \quad \text{and} \quad E_\theta(y_2|y_{obs}) = y_2 = 34,$$

$$E_\theta(y_3|y_{obs}) = E_\theta(y_3|y_3+y_4) = 125\frac{\frac{1}{4}\theta}{\frac{1}{2}+\frac{1}{4}\theta}, \quad y_3^{(t)} = 125\frac{\frac{1}{4}\theta^{(t)}}{\frac{1}{2}+\frac{1}{4}\theta^{(t)}}$$

- M-step: The complete-data maximum likelihood estimate of θ is

$$\hat{\theta} = \frac{y_2+y_3}{y_1+y_2+y_3}.$$

Then,

$$\theta^{(t+1)} = \frac{34+y_3^{(t)}}{72+y_3^{(t)}}.$$

Example: Peppered Moths

- Three possible alleles: C, I, and T.
- C is dominant to I and T is recessive to I
  - Genotype CC, CI, and CT $\Rightarrow$ carbonaria phenotype.
  - Genotype TT $\Rightarrow$ typica phenotype.
  - Genotype II and TI $\Rightarrow$ intermediate phenotype called insularia.
- If the allele frequencies in the population are p_C, p_I, and p_T, then the genotype frequencies should be p_C^2, $2p_Cp_1$, $2p_Cp_T$, p_I^2, $2p_Ip_T$, and p_T^2 for genotypes CC, CI, CT, II, IT, and TT, respectively. Note that $p_C + p_I + p_T = 1$.
- Capture n moths, of which there are n_C, n_I, and n_T of the carbonaria, insularia, and typica phenotypes, respectively. Thus, $n = n_C + n_I + n_T$.

Example: Peppered Moths

- The observed data are $x = (n_C, n_I, n_T)$ and the complete data are $y = (n_{CC}, n_{CI}, n_{CT}, n_{II}, n_{IT}, n_{TT})$. The mapping from the complete data to the observed data is $x = M(y) = (n_{CC} + n_{CI} + n_{CT}, n_{II} + n_{IT}, n_{TT})$.
- Wish to estimate the allele probabilities, p_C, p_I, and p_T.
- The complete data log likelihood function is multinomial

$$\begin{aligned}\log L(p|y) = {} & n_{CC} \log \left\{p_C^2\right\} + n_{CI} \log \{2p_C p_I\} + n_{CT} \log \{2p_C p_T\} \\ & + n_{II} \log \left\{p_I^2\right\} + n_{IT} \log \{2p_I p_T\} + n_{TT} \log \left\{p_T^2\right\} \\ & + \log \binom{n}{n_{CC}\ n_{CI}\ n_{CT}\ n_{II}\ n_{IT}\ n_{TT}}.\end{aligned}$$

Example: Peppered Moths

- E-Step

$$E\left\{N_{CC}|n_C, n_I, n_T, p^{(t)}\right\} = n_{CC}^{(t)} = \frac{n_C\left(p_C^{(t)}\right)^2}{\left(p_C^{(t)}\right)^2 + 2p_C^{(t)}p_I^{(t)} + 2p_C^{(t)}p_T^{(t)}}$$

$$E\left\{N_{CI}|n_C, n_I, n_T, p^{(t)}\right\} = n_{CI}^{(t)} = \frac{2n_C p_C^{(t)} p_I^{(t)}}{\left(p_C^{(t)}\right)^2 + 2p_C^{(t)}p_I^{(t)} + 2p_C^{(t)}p_T^{(t)}}$$

$$E\left\{N_{CT}|n_C, n_I, n_T, p^{(t)}\right\} = n_{CT}^{(t)} = \frac{2n_C p_C^{(t)} p_T^{(t)}}{\left(p_C^{(t)}\right)^2 + 2p_C^{(t)}p_I^{(t)} + 2p_C^{(t)}p_T^{(t)}}$$

Example: Peppered Moths

- E-Step

$$E\left\{N_{II}|n_C, n_I, n_T, p^{(t)}\right\} = n_{II}^{(t)} = \frac{n_I\left(p_I^{(t)}\right)^2}{\left(p_I^{(t)}\right)^2 + 2p_I^{(t)}p_T^{(t)}}$$

$$E\left\{N_{IT}|n_C, n_I, n_T, p^{(t)}\right\} = n_{IT}^{(t)} = \frac{2n_I p_I^{(t)} p_T^{(t)}}{\left(p_I^{(t)}\right)^2 + 2p_I^{(t)}p_T^{(t)}}$$

Then,

$$\begin{aligned} Q(p|p^{(t)}) = {} & n_{CC}^{(t)}\log\{p_C^2\} + n_{CI}^{(t)}\log\{2p_Cp_I\} + n_{CT}^{(t)}\log\{2p_Cp_T\} \\ & + n_{II}^{(t)}\log\{p_I^2\} + n_{IT}^{(t)}\log\{2p_Ip_T\} + n_{TT}\log\{p_T^2\} + \kappa(n_C, n_I, n_T, p^{(t)}). \end{aligned}$$

Example: Peppered Moths

- M-Step: Recall $p_T = 1 - p_C - p_I$ and differentiate w.r.t p_C and p_I.

$$\frac{dQ(p|p^{(t)})}{dp_C} = \frac{2n_{CC}^{(t)} + n_{CI}^{(t)} + n_{CT}^{(t)}}{p_C} + \frac{2n_{TT}^{(t)} + n_{CT}^{(t)} + n_{IT}^{(t)}}{1 - p_C - p_I}$$
$$\frac{dQ(p|p^{(t)})}{dp_I} = \frac{2n_{II}^{(t)} + n_{IT}^{(t)} + n_{CI}^{(t)}}{p_C} + \frac{2n_{TT}^{(t)} + n_{CT}^{(t)} + n_{IT}^{(t)}}{1 - p_C - p_I}$$

  Then,

$$p_C^{(t+1)} = \frac{2n_{CC}^{(t)} + n_{CI}^{(t)} + n_{CT}^{(t)}}{2n}, \qquad p_I^{(t+1)} = \frac{2n_{II}^{(t)} + n_{IT}^{(t)} + n_{CI}^{(t)}}{2n}$$

- Suppose the observed phenotpye counts are $n_C = 85$, $n_I = 196$, and $n_T = 341$.

Observed (n_C, n_I, n_T)

we want to estimate (p_C, p_I, p_T).

Direct estimation of p_C, p_I, p_T is not possible.

$\Rightarrow$ $C = (CC, CI, CT)$ $I = (II, IT)$

Complete data : $\sim$ Multinomial dist $(p_{CC}, p_{CI}, p_{CT}, p_{II}, p_{IT}, p_{TT})$

$y = (n_{CC}, n_{CI}, n_{CT}, n_{II}, n_{IT}, n_{TT})$

Observed data

$x = (n_{CC} + n_{CI} + n_{CT}, n_{II} + n_{IT}, n_{TT})$

$\vdots \qquad \vdots \qquad \vdots$

$p_C \qquad p_I \qquad p_T.$

$(p_C^2, 2p_Cp_I, 2p_Cp_T, p_I^2, 2p_Ip_T, p_T^2)$

Complete-data log likelihood

$$l(p) = n_{CC} \log p_C^2 + n_{CI} \log 2p_Cp_I + n_{CT} \log 2p_Cp_T + n_{II} \log p_I^2 + n_{IT} \log 2p_Ip_T + n_{TT} \log p_T^2 + \text{Constant}.$$

E-Step $E\{N_{CC} \mid n_C, n_I, n_T, p^{(t)}\} = n_{CC}^{(t)} = n_C \dfrac{p_C^{(t)2}}{p_C^{(t)2} + 2p_C^{(t)}p_I^{(t)} + 2p_C^{(t)}p_T^{(t)}}$

$E\{N_{CI} \mid n_C, n_I, n_T, p^{(t)}\} = n_{CI}^{(t)} = n_C \dfrac{2p_C^{(t)}p_I^{(t)}}{p_C^{(t)2} + 2p_C^{(t)}p_I^{(t)} + 2p_C^{(t)}p_T^{(t)}}$

$E\{N_{CT} \mid n_C, n_I, n_T, p^{(t)}\} = n_{CT}^{(t)} = n_C \dfrac{2p_C^{(t)}p_T^{(t)}}{p_C^{(t)2} + 2p_C^{(t)}p_I^{(t)} + 2p_C^{(t)}p_T^{(t)}}$

$$E\{N_{II} \mid n_c, n_I, n_T, p^{(t)}\} = n_{II}^{(t)} = n_I \times \frac{p_I^{(t)2}}{p_I^{(t)2} + 2p_I^{(t)}p_T^{(t)}}$$

$$E\{N_{IT} \mid n_c, n_I, n_T, p^{(t)}\} = n_{IT}^{(t)} = n_I \frac{2p_I^{(t)}p_T^{(t)}}{p_I^{(t)2} + 2p_I^{(t)}p_T^{(t)}}$$

$$E\{N_{TT} \mid n_c, n_I, n_T, p^{(t)}\} = ? \quad X.$$

$C \rightarrow I \rightarrow T$

estimation : p

$$Q(p \mid p^{(t)}) = \underbrace{n_{cc}^{(t)}}_{\text{Expectation } p^{(t)}} \log p_c^2 + n_{cI}^{(t)} \log 2p_c p_I + n_{cT}^{(t)} \log 2p_c p_T$$

use observed value.

$$+ n_{II}^{(t)} \log p_I^2 + n_{IT}^{(t)} \log 2p_I p_T + n_{TT} \log p_T^2 + \text{constant}.$$

$p_T = (1 - p_c - p_I)$

M-Step

$$\frac{dQ(p \mid p^{(t)})}{dp_c} = \frac{2n_{cc}^{(t)} + n_{cI}^{(t)} + n_{cT}^{(t)}}{p_c} + \frac{2n_T + n_{cT}^{(t)} + n_{IT}^{(t)}}{1 - p_c - p_I} = 0$$

$$\frac{dQ(p \mid p^{(t)})}{dp_I} = \frac{2n_{II}^{(t)} + n_{cI}^{(t)} + n_{IT}^{(t)}}{p_I} + \frac{2n_T + n_{cT}^{(t)} + n_{IT}^{(t)}}{1 - p_c - p_I} = 0$$

$$p_c^{(t+1)} = \frac{2n_{cc}^{(t)} + n_{cI}^{(t)} + n_{cT}^{(t)}}{2n}$$

$$p_I^{(t+1)} = \frac{2n_{II}^{(t)} + n_{cI}^{(t)} + n_{IT}^{(t)}}{2n}$$

(Ex) $n_c = 85$, $n_I = 196$, $n_T = 341$. $\Rightarrow$ Implement EM algorithm. (HW)

EM algorithm HW3 (Total HW : 5)

Due : Nov. 11

Class Material
- Optimization
- Combinatorial
- EM.
- Numerical Integration
- Bootstrap.

(Ex) Normal Mixture Model

(J Group)

$x_1, \cdots, x_n \sim \sum_{j=1}^{J} p_j N(\mu_j, \sigma_j^2)$ ... (Observed Data)

(p_j: prop. of dist^n. μ_j: mean. σ_j^2: variance.)

Define $(y_{i1}, \cdots, y_{iJ}) \sim$ Multinomial $(1; p_1, \cdots, p_J)$ where $\sum_j y_{ij} = 1$.

↳ Indicator variable (which distribution x_i belongs to)

Given $y_{ij^*} = 1$ and $y_{ij} = 0$ for $j \neq j^*$,

we assume $x_i \sim N(\mu_{j^*}, \sigma_{j^*}^2)$

Complete data : $\{x_i, y_{i1}, \cdots, y_{iJ}\}_i$

Observed-data log-likelihood

$$l(\mu, \sigma^2, p \mid x) = \sum_i \log \left\{ \sum_{j=1}^{J} p_j N(x_i; \mu_j, \sigma_j^2) \right\}$$

$$L(x_i; y_{i1}, \cdots y_{ip}) = \prod_{j=1}^{J} \left\{ p_j N(x_i; \mu_j, \sigma_j^2) \right\}^{y_{ij}}$$

$$L(x; y_1, \cdots, y_p) = \prod_{i=1}^{n} \prod_{j=1}^{J} \left\{ p_j N(x_i; \mu_j, \sigma_j^2) \right\}^{y_{ij}}$$

Complete-data log-likelihood

$$l(\mu, \sigma^2, p \mid x, y) = \sum_i \sum_j y_{ij} \left\{ \log p_j + \log N(x_i; \mu_j, \sigma_j^2) \right\}$$

$$\propto \sum_{ij} y_{ij} \left\{ \log p_j - \frac{1}{2\sigma_j^2}(x_i - \mu_j)^2 - \log \sigma_j \right\}$$

E-Step : Need to calculate the expected proportion of mixture component

$x_i \rightarrow \begin{matrix} p_{i1} \\ \vdots \\ p_{ij} \end{matrix}$ Need to estimate ... $w_{ij}^{(t)}$

$w_{ij}^{(t)} = E(y_{ij} \mid x_i, \mu^{(t)}, \sigma^{2(t)}, p^{(t)})$ ← indicator variable.

weight for individual

$= P(y_{ij} = 1 \mid x_i, \mu^{(t)}, \sigma^{2(t)}, p^{(t)})$

$= \dfrac{p_j^{(t)} N(x_i; \mu_j^{(t)}, \sigma_j^{2(t)})}{\sum_{j=1}^{J} p_j^{(t)} N(x_i; \mu_j^{(t)}, \sigma_j^{2(t)})}$

M-Step :

$p_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} w_{ij}^{(t)}$

$\mu_j^{(t+1)} = \sum_i w_{ij}^{(t)} \cdot x_i \Big/ \sum_i w_{ij}^{(t)}$

$\sigma_j^{2(t+1)} = \sum_i w_{ij}^{(t)} (x_i - \mu_j^{(t+1)})^2 \Big/ \sum_i w_{ij}^{(t)}$

(EX2) Mixed-Effect Model

$Y_i = X_i \beta + Z_i b_i + \varepsilon_i$

β: fixed effect

b_i: random effect (it has a different value by individual)

$b_i \sim N_q(0, D)$ ← covariance matrix for random effect.

$\varepsilon_i \sim N_{n_i}(0, \sigma^2 I_{n_i})$

Our Object: Estimate β, D, σ^2.

$\Sigma_i = Z_i D Z_i^T + \sigma^2 I$

Observed-data log-likelihood

$l(\beta, D, \sigma^2 \mid Y_1, \cdots, Y_n) = \sum_{i=1}^{n} \left\{ -\frac{1}{2} (Y_i - X_i\beta)^T \Sigma_i^{-1} (Y_i - X_i\beta) - \frac{1}{2} \log |\Sigma_i| \right\}$

IWLS

We can directly maximize (β, D, σ^2) using Newton-Raphson method!