Allele Frequency Estimation

Example: ABO blood types

- ABO genetic locus exhibits three alleles: A, B, and O
- Four phenotypes: A, B, AB, and O

Genotype	A/A	A/O	A/B	B/B	B/O	O/O
Phenotype	A	A	AB	B	B	O

- *Data:* Observed counts of four phenotypes A, B, AB, and O

n_A	n_B	n_{AB}	n_O	n
186	38	13	284	521

- *Aim:* Estimate frequencies p_A, p_B, and p_O of alleles A, B, and O

Modelling:

- Observed data: N_A, N_B, N_{AB}, N_O
- Complete data: N_{AA}, N_{AO}, N_{BB}, N_{BO}, N_{AB}, N_O
- According to the Hardy-Weinberg law, the genotype frequencies are

Genotype	A/A	A/O	A/B	B/B	B/O	O/O
Frequency	p_A^2	$2p_Ap_O$	$2p_Ap_B$	p_B^2	$2p_Bp_O$	p_O^2

- Genotype counts $N = (N_{AA}, N_{AO}, N_{AB}, N_{BB}, N_{BO}, N_O)$ are jointly multinomially distributed.

Allele Frequency Estimation

Complete-data log-likelihood function

$$\begin{aligned} l_n(p|N) &= N_{AA}\log(p_A^2) + N_{BB}\log(p_B^2) + N_O\log(p_O^2) \\ &\quad + N_{AB}\log(2p_Ap_B) + N_{AO}\log(2p_Ap_O) + N_{BO}\log(2p_Bp_O) \\ &\quad + \log\left(\frac{n!}{N_{AA}!\,N_{AO}!\,N_{AB}!\,N_{BB}!\,N_{BO}!\,N_O!}\right) \end{aligned}$$

Application of EM algorithm

- Let $N_{\text{obs}} = (N_A, N_B, N_{AB}, N_O)$.
- ***E-step:*** Since $N_{AA} + N_{AO} = N_A$ we have

  $$N_{AA}|N_A \sim \text{Bin}\left(N_A, \frac{p_A^2}{p_A^2 + 2p_Ap_O}\right)$$

  which yields the expectations

  $$\begin{aligned} N_{AA}^{(k)} &= \mathbb{E}(N_{AA}|N_{\text{obs}}, p^{(k)}) = N_A \cdot \frac{p_A^{(k)^2}}{p_A^{(k)^2} + 2p_A^{(k)}p_O^{(k)}} \\ N_{AO}^{(k)} &= \mathbb{E}(N_{AO}|N_{\text{obs}}, p^{(k)}) = N_A \cdot \frac{2p_A^{(k)}p_O^{(k)}}{p_A^{(k)^2} + 2p_A^{(k)}p_O^{(k)}} \end{aligned}$$

  and similarly

  $$\begin{aligned} N_{BB}^{(k)} &= \mathbb{E}(N_{BB}|N_{\text{obs}}, p^{(k)}) = N_B \cdot \frac{p_B^{(k)^2}}{p_B^{(k)^2} + 2p_B^{(k)}p_O^{(k)}} \\ N_{BO}^{(k)} &= \mathbb{E}(N_{BO}|N_{\text{obs}}, p^{(k)}) = N_B \cdot \frac{2p_B^{(k)}p_O^{(k)}}{p_B^{(k)^2} + 2p_B^{(k)}p_O^{(k)}} \end{aligned}$$

  while obviously

  $$\mathbb{E}(N_{AB}|N_{\text{obs}}, p^{(k)}) = N_{AB} \qquad \text{and} \qquad \mathbb{E}(N_O|N_{\text{obs}}, p^{(k)}) = N_O.$$

Allele Frequency Estimation

- ***M-step:*** Maximize $Q(p|p^{(k)})$ under the restriction $p_A + p_B + p_O = 1$. Introduce Lagrange multiplier (Rice, p. 259) and maximize

  $$Q_L(p, \lambda|p^{(k)}) = Q(p|p^{(k)}) + \lambda(p_A + p_B + p_O - 1)$$

  with respect to p and λ.

  $$\begin{aligned} \frac{\partial Q_L(p,\lambda|p^{(k)})}{\partial p_A} &= \frac{2N_{AA}^{(k)}}{p_A} + \frac{N_{AO}^{(k)}}{p_A} + \frac{N_{AB}}{p_A} + \lambda \\ \frac{\partial Q_L(p,\lambda|p^{(k)})}{\partial p_B} &= \frac{2N_{BB}^{(k)}}{p_B} + \frac{N_{BO}^{(k)}}{p_B} + \frac{N_{AB}}{p_B} + \lambda \\ \frac{\partial Q_L(p,\lambda|p^{(k)})}{\partial p_O} &= \frac{N_{AO}^{(k)}}{p_O} + \frac{N_{BO}^{(k)}}{p_O} + \frac{2N_O}{p_O} + \lambda \\ \frac{\partial Q_L(p,\lambda|p^{(k)})}{\partial \lambda} &= p_A + p_B + p_O - 1 \end{aligned}$$

  Taking the sum of the three equations, we get (using $p_A + p_B + p_O = 1$)

  $$\lambda = -2\,n$$

  which yields for the first three equations the solutions

  $$\begin{aligned} p_A^{(k+1)} &= \frac{2N_{AA}^{(k)} + N_{AO}^{(k)} + N_{AB}}{2n} \\ p_B^{(k+1)} &= \frac{2N_{BB}^{(k)} + N_{BO}^{(k)} + N_{AB}}{2n} \\ p_O^{(k+1)} &= \frac{N_{AO}^{(k)} + N_{BO}^{(k)} + 2N_O}{2n} \end{aligned}$$

Allele Frequency Estimation

Starting values:

$$p_A = p_B = p_O = \frac{1}{3}$$

Iterations:

k	p_A	p_B	p_O
1	0.2505	0.0611	0.6884
2	0.2185	0.0505	0.7311
3	0.2142	0.0502	0.7357
4	0.2137	0.0501	0.7362
5	0.2136	0.0501	0.7363
6	0.2136	0.0501	0.7363

Starting values:

$$p_A = p_O = 0.01, p_B = 0.98$$

Iterations:

k	p_A	p_B	p_O
1	0.2505	0.0847	0.6648
2	0.2193	0.0511	0.7296
3	0.2143	0.0502	0.7355
4	0.2137	0.0501	0.7362
5	0.2136	0.0501	0.7363
6	0.2136	0.0501	0.7363

Note: Results do not change for different starting values.

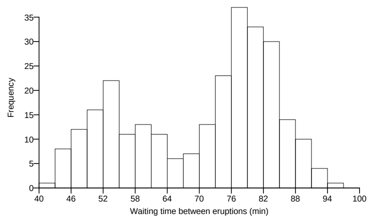
Implementation in R

```
#EM iteration
#  Arguments:
#  N=(Na,Nb,Nab,No)
#  p=(pa,pb,po)
emstep<-function(N,p) {
  #E-step
  Naa<-N[1]*p[1]^2/(p[1]^2+2*p[1]*p[3])
  Nao<-N[1]*2*p[1]*p[3]/(p[1]^2+2*p[1]*p[3])
  Nbb<-N[2]*p[2]^2/(p[2]^2+2*p[2]*p[3])
  Nbo<-N[2]*2*p[2]*p[3]/(p[2]^2+2*p[2]*p[3])
  #M-step
  n<-sum(N)
  p[1]=(2*Naa+Nao+N[3])/(2*n)
  p[2]=(2*Nbb+Nbo+N[3])/(2*n)
  p[3]=(2*Nao+Nbo+N[4])/(2*n)
  p
}
#Data
N<-c(186,38,13,284)
#Starting value
p<-c(1,1,1)/3
#First iteration
p<-emstep(N,p)
p
#Second iteration
p<-emstep(N,p)
p
#Repeat until convergence
```

Mixtures

Example: Old Faithful

Data: 272 waiting times between eruptions for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA



Model: Mixture of two Gaussian populations (short/long waiting times):

$$f_Y(y|\theta) = p\,\frac{1}{\sigma_1}\,\varphi\Big(\frac{x-\mu_1}{\sigma_1}\Big) + (1-p)\,\frac{1}{\sigma_2}\,\varphi\Big(\frac{x-\mu_2}{\sigma_2}\Big)$$

Parameters: $\theta = (p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)^\mathsf{T}$

Idea: If we knew the group which each observation belongs to, we could simply fit a normal distribution to each group.

Missing data: Group indicator

$$Z_i = \begin{cases} 1 & Y_i \text{ belongs to group of long waiting times} \\ 0 & Y_i \text{ belongs to group of short waiting times} \end{cases}$$

Z_i is Bernoulli distributed with parameter p: $Z_i \overset{\text{iid}}{\sim} \text{Bin}(1,p)$

Complete-data likelihood:

$$L_n(\theta|Y,Z) = \prod_{i=1}^n p^{Z_i}(1-p)^{1-Z_i} \cdot \frac{1}{\sigma_1^{Z_i}}\,\varphi\Big(\frac{Y_i-\mu_1}{\sigma_1}\Big)^{Z_i} \frac{1}{\sigma_2^{1-Z_i}}\,\varphi\Big(\frac{Y_i-\mu_2}{\sigma_2}\Big)^{1-Z_i}$$

Mixtures

Log-likelihood function

$$\begin{aligned} l_n(\theta|Y,Z) = {} & \sum_{i=1}^n Z_i \cdot \log(p) + \sum_{i=1}^n (1-Z_i)\cdot\log(1-p) \\ & - \frac{1}{2}\sum_{i=1}^n Z_i \cdot \log(2\pi\sigma_1^2) - \frac{1}{2\sigma_1^2}\sum_{i=1}^n Z_i(Y_i-\mu_1)^2 \\ & - \frac{1}{2}\sum_{i=1}^n (1-Z_i)\cdot\log(2\pi\sigma_2^2) - \frac{1}{2\sigma_2^2}\sum_{i=1}^n (1-Z_i)(Y_i-\mu_2)^2 \end{aligned}$$

Application of EM algorithm

- ***E-step:***
  $l_n(\theta|Y,Z)$ is linear in Z_i. It therefore suffices to find the conditional mean $\mathbb{E}(Z_i|Y_i,\theta^{(k)})$.

  The conditional distribution of Z_i given Y is

  $$Z_i|Y_i,\theta^{(k)} \sim \text{Bin}(1, p_i^{(k)})$$

  with

  $$p_i^{(k)} = \frac{p^{(k)}\,\frac{1}{\sigma_1^{(k)}}\,\varphi\Big(\frac{x-\mu_1^{(k)}}{\sigma_1^{(k)}}\Big)}{p^{(k)}\,\frac{1}{\sigma_1^{(k)}}\,\varphi\Big(\frac{Y_i-\mu_1^{(k)}}{\sigma_1^{(k)}}\Big) + (1-p^{(k)})\,\frac{1}{\sigma_2^{(k)}}\,\varphi\Big(\frac{Y_i-\mu_2^{(k)}}{\sigma_2^{(k)}}\Big)}.$$

  Thus the conditional mean is

  $$\mathbb{E}(Z_i|Y_i,\theta^{(k)}) = p_i^{(k)}.$$

Mixtures

- ***M-step:*** Substituting $p_i^{(k)}$ for Z_i we obtain

$$\begin{aligned} Q(\theta|\theta^{(k)}) = {} & \sum_{i=1}^n p_i^{(k)} \cdot \log(p) + \sum_{i=1}^n q_i^{(k)}\cdot\log(1-p) \\ & - \frac{1}{2}\sum_{i=1}^n p_i^{(k)} \cdot \log(2\pi\sigma_1^2) - \frac{1}{2\sigma_1^2}\sum_{i=1}^n p_i^{(k)}(Y_i-\mu_1)^2 \\ & - \frac{1}{2}\sum_{i=1}^n q_i^{(k)}\cdot\log(2\pi\sigma_2^2) - \frac{1}{2\sigma_2^2}\sum_{i=1}^n q_i^{(k)}(Y_i-\mu_2)^2 \end{aligned}$$

where $q_i^{(k)} = 1 - p_i^{(k)}$.

Setting the first derivatives of $Q(\theta|\theta^{(k)})$ equal to zero we obtain

$$p^{(k+1)} = \frac{1}{n}\sum_{i=1}^n p_i^{(k)}$$

$$\mu_1^{(k+1)} = \frac{\sum_{i=1}^n p_i^{(k)}\,Y_i}{\sum_{i=1}^n p_i^{(k)}}$$

$$\mu_2^{(k+1)} = \frac{\sum_{i=1}^n q_i^{(k)}\,Y_i}{\sum_{i=1}^n q_i^{(k)}}$$

$$(\sigma_1^{(k+1)})^2 = \frac{\sum_{i=1}^n p_i^{(k)}\,(Y_i-\mu_1^{(k+1)})^2}{\sum_{i=1}^n p_i^{(k)}}$$

$$(\sigma_2^{(k+1)})^2 = \frac{\sum_{i=1}^n q_i^{(k)}\,(Y_i-\mu_2^{(k+1)})^2}{\sum_{i=1}^n q_i^{(k)}}$$

Mixtures

Starting values:

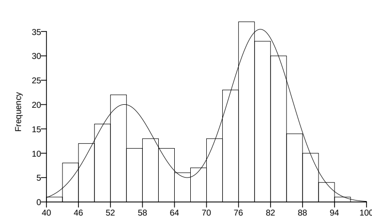
$$p^{(0)} = 0.4 \quad \mu_1^{(0)} = 40 \quad \sigma_1^{(0)} = 4$$
$$\mu_2^{(0)} = 90 \quad \sigma_2^{(0)} = 4$$

Iterations:

k	$p^{(k)}$	$\mu_1^{(k)}$	$\mu_2^{(k)}$	$\sigma_1^{(k)}$	$\sigma_2^{(k)}$
1	0.3508	54.22	79.91	5.465	5.999
2	0.3539	54.38	79.94	5.671	6.013
3	0.3562	54.46	79.99	5.744	5.969
4	0.3578	54.51	80.02	5.787	5.935
5	0.3588	54.55	80.05	5.815	5.912
6	0.3595	54.57	80.06	5.834	5.897
7	0.3600	54.59	80.07	5.846	5.887
8	0.3603	54.60	80.08	5.855	5.880
9	0.3605	54.60	80.08	5.860	5.876
10	0.3606	54.61	80.09	5.864	5.873
11	0.3607	54.61	80.09	5.866	5.871
12	0.3608	54.61	80.09	5.868	5.870
13	0.3608	54.61	80.09	5.869	5.869
14	0.3608	54.61	80.09	5.870	5.869
15	0.3609	54.61	80.09	5.870	5.868
20	0.3609	54.61	80.09	5.871	5.868
25	0.3609	54.61	80.09	5.871	5.868

Implementation in R

```
p<-c(0.5,40,90,20,20)
emstep<-function(Y,p) {
  EZ<-p[1]*dnorm(Y,p[2],sqrt(p[4]))/
    (p[1]*dnorm(Y,p[2],sqrt(p[4]))
    +(1-p[1])*dnorm(Y,p[3],sqrt(p[5])))
  p[1]<-mean(EZ)
  p[2]<-sum(EZ*Y)/sum(EZ)
  p[3]<-sum((1-EZ)*Y)/sum(1-EZ)
  p[4]<-sum(EZ*(Y-p[2])^2)/sum(EZ)
  p[5]<-sum((1-EZ)*(Y-p[3])^2)/sum(1-EZ)
  p
}
emiteration<-function(Y,p,n=10) {
  for (i in (1:n)) {
    p<-emstep(Y,p)
  }
  p
}
p<-c(0.5,40,90,20,20)
p<-emiteration(Y,p,20)
p
p<-emstep(Y,p)
p
```

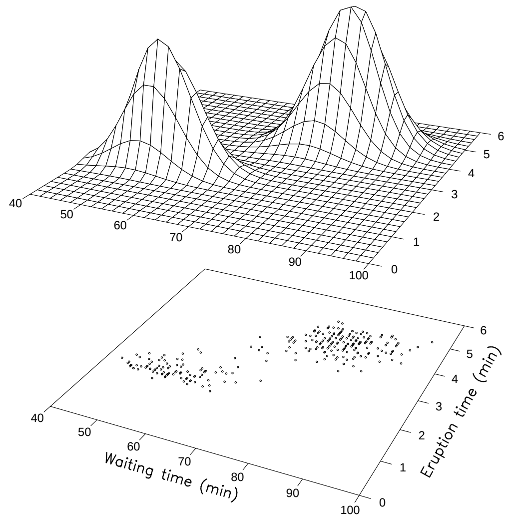


Mixtures

Example: Bivariate distribution

Data:

- Waiting times between eruptions (in min) for the Old Faithful geyser
- Eruption times (in min) for the Old Faithful geyser



Example: EM algorithm for bivariate Gaussian mixtures (JAVA applet)

`http://dowww.epfl.ch/mantra/tutorial/english/gaussian/html`

Convergence of the EM algorithm

Example: Bivariate t-distribution

Suppose that $Y_i = (Y_{i1}, Y_{i2})^{\mathsf{T}}$, $i = 1, \ldots, 5+m$ are independently sampled from a bivariate t distribution with likelihood function

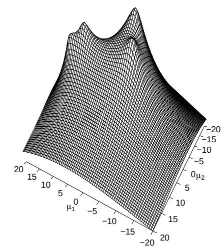
$$L_n(\mu|Y) = \prod_{i=1}^{n} \big(1 - (Y_{i1} - \mu_1)^2 + (Y_{i2} - \mu_2)^2\big)^{-\frac{3}{2}}.$$

Furthermore, suppose that only the first 5 values are observed.

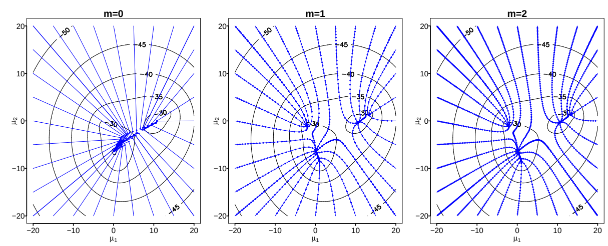
- Convergence of the EM algorithm depends on the amount of missing data.
- The more data are missing and have to be estimated, the slower the EM algorithm converges.
- Here

$$\mu^{(k+1)} = \frac{\sum_{i=1}^{5} x_i^{(k)} y_i + m \cdot \mu^{(k)}}{\sum_{i=1}^{5} x_i^{(k)} + m}$$

is a weighted mean with strong weight on the previous $\mu^{(k)}$ if the proportion of missing data is large.



Log-likelihood function $l_n(\mu|y)$



Convergence of the EM algorithm for $m = 0, 1, 2$.