Shuyang_HW2

Assignment2

Question 1

(a)Full data log-likelihood function:

l(pA, pB, pO) = nAAlog(pA^2) + nAOlog(2pApO) + nBBlog(pB^2) + nBOlog(2pBpO) + nABlog(2pApB) + nOOlog(pO^2)

(b)E-step:

nAA(t) = 186 X $pA^{2/(pA}$2 + 2pApO)

nAO(t) = 186 X 2pApO/(pA^2 + 2pApO)

nBB(t) = 38 X $pB^{2/(pB}$2 + 2pBpO)

nBO(t) = 38 X 2pBpO/(pB^2 + 2pBpO)

nAB(t) = 13

nOO(t) = 284

(c) pA(t) = (2nAA(t) + nAB(t) + nAO(t))/2n = (186 + 13 + nAA(t))/2 * 521

pB(t) = (2nBB(t) + nAB(t) + nBO(t))/2n = (38 + 13 + nBB(t))/2 * 521

pO(t) = (2nOO(t) + nAO(t) + nBO(t))/2n = (284 + 284 + nAO(t) + nBO(t))/2*521

(d)

```
p.old = c(186/521, 38/521, 284/521)
p.diff=1
count=0

while(p.diff>0.001){
  #E-step
  nAA = 186 * p.old[1]^2/(p.old[1]^2 + 2*p.old[1]*p.old[3])
  nAO = 186 * 2*p.old[1]*p.old[3]/(p.old[1]^2 + 2*p.old[1]*p.old[3])
  nBB = 38 * p.old[2]^2/(p.old[2]^2 + 2*p.old[2]*p.old[3])
  nBO = 38 * 2*p.old[2]*p.old[3]/(p.old[2]^2 + 2*p.old[2]*p.old[3])
  #M-step
  pA = (186 + 13 + nAA)/1042
  pB = (38 + 13 + nBB)/1042
  pO = (284 + 284 + nAO + nBO)/1042
  p.new=c(pA,pB,pO)
  p.diff = max(abs(p.new - p.old))
  p.old=p.new
  count=count+1
}

pA
```

```
## [1] 0.2136286
```

```
pB
```

```
## [1] 0.05014597
```

```
p0
```

```
## [1] 0.7362254
```

Question 2

(a)Full data log-likelihood function:

l(p11, p10, p01, p00) = (40 + y11)log(p11) + (25 + y10)log(p10) + (150 + y01)log(p01) + (100 + y00)log(p00)

(b)E-step:

y11(t) = 15 X p11/(p11 + p10)

y10(t) = 15 X p10/(p11 + p10)

y01(t) = 70 X p01/(p01 + p00)

y00(t) = 70 X p00/(p01 + p00)

(c)M-step: p11(t) = 40 + y11(t)/(315 + y11(t) + y10(t) + y01(t) + y00(t))

p10(t) = 25 + y10(t)/(315 + y11(t) + y10(t) + y01(t) + y00(t))

p01(t) = 150 + y01(t)/(315 + y11(t) + y10(t) + y01(t) + y00(t))

p01(t) = 100 + y00(t)/(315 + y11(t) + y10(t) + y01(t) + y00(t))

(d)

```
p.old=c(40/315,25/315,150/315,100/315)
p.diff=1
count=0

while(p.diff>0.001){
  #E-step
  y11=15*p.old[1]/(p.old[1]+p.old[2])
  y10=15*p.old[2]/(p.old[1]+p.old[2])
  y01=70*p.old[3]/(p.old[3]+p.old[4])
  y00=70*p.old[4]/(p.old[3]+p.old[4])
  sumy=y11+y10+y01+y00
  #M-step
  p11=(40+y11)/(315+sumy)
  p10=(25+y10)/(315+sumy)
  p01=(150+y01)/(315+sumy)
  p00=(100+y00)/(315+sumy)
  p.new=c(p11,p10,p01,p00)
  p.diff=max(abs(p.new-p.old))
  p.old=p.new
  count=count+1
}

b0 = log(p10/p00)
b1 = log(p11*p00/p10*p01)

b0
```

```
## [1] -1.425515
```

```
b1
```

```
## [1] -1.4034
```

Question 3

(a)Primary scientific question of interest: find the stratification of the population for more accurate association mapping results.

Analysis model: Expectation Maximization Maximun Likelihood Estimation

Interpretation of parameters: Means and variances of the normal distributions representing the principle components of the genotypic values for each individual

(b)Missing variables: stratification of the population

(c)The modeling assumptions include: There are M independet markers out of all genotypes. The stratification only depends on genotypes/markers. The stratification/priciple components follow approximately a mixture of K normal distributions.

(d)Purpose of EM algorithms: E step estimates the probability each individual originated from each principle component. M step estimates the mean and variation of the normal distributions of the principle components.

(e) If the population substructure is not taken into account, the association mapping is prone to false-positive errors. For example, population stratification across different subgroups can lead to spurious associations as a result of confounding. The most likely source of confounding is ethnicity, in which both disease prevalence and allele frequencies are different by ethnicity, and cases and controls are not well-matched in terms of ethnic origin.