STA 6106 Homework 7: EM

Tyler Grimes

March 29, 2016

1. 10 observations are selected from a Bivariate Normal distribution. Find the MLE for $\theta = (\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22})$.

w1:	8	11	16	18	6	4	20	25	9	13
w2:	10	14	16	15	20	4	18	22	NA	NA

The density of the bivariate normal is

$$f(w;\theta) = \frac{1}{2\pi}|\Sigma|^{-1} exp\left(-\frac{1}{2}(w-\mu)'\Sigma^{-1}(w-\mu)\right)$$

where $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$. Let $T_1 = \sum_{i=1}^n w_{1i}$, $T_2 = \sum_{i=1}^n w_{2i}$, $T_{11} = \sum_{i=1}^n w_{1i}^2$, $T_{22} = \sum_{i=1}^n w_{2i}^2$, and $T_{12} = \sum_{i=1}^n w_{1i}w_{2i}$. Without missing values, the MLE estimates for the parameters θ are,

$$\begin{aligned}
\hat{\mu}_1 &= \frac{T_1}{n} \\
\hat{\mu}_2 &= \frac{T_2}{n} \\
\hat{\sigma}_{11} &= \frac{T_{11} - T_1^2/n}{n} \\
\hat{\sigma}_{22} &= \frac{T_{22} - T_2^2/n}{n} \\
\hat{\sigma}_{12} &= \frac{T_{12} - T_1 T_2/n}{n}
\end{aligned}$$

For the E-step, we will need both $E(w_{2i}|w_{1i},\theta)$ and $E(w_{2i}^2|w_{1i},\theta)$.

$$\begin{aligned}
E(w_{2i}|w_{1i},\theta) &= \mu_2 + \frac{\sigma_{12}}{\sigma_{11}}(w_{1i} - \mu_1) \\
E(w_{2i}^2|w_{1i},\theta) &= \sigma_{22}(1-\rho^2) + (E(w_{2i}|w_{1i},\theta))^2
\end{aligned}$$

where $\rho^2 = \frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}}$. We use these conditional expectations to estimate the missing values.

$$\begin{aligned}
w_{2i}^{(k+1)} &= \hat{\mu}_2 + \frac{\hat{\sigma}_{12}}{\hat{\sigma}_{11}}(w_{1i} - \hat{\mu}_1) \\
w_{2i}^{2(k+1)} &= \hat{\sigma}_{22}(1-r^2) + (w_{2i}^{(k)})^2
\end{aligned}$$

where $r^2 = \frac{\hat{\sigma}_{12}^2}{\hat{\sigma}_{11}\hat{\sigma}_{22}}$. To start the EM algorithm, an initial guess for the missing values is needed. We will use 0's for those missing.

```r
#Use 0's for initial values of missing data.
data[9, 2] = 0
data[10, 2] = 0

#Find MLE estimates of parameters.
mu1 = sum(data[, 1])/n
mu2 = sum(data[, 2])/n
sigma11 = (sum(data[, 1]^2) - sum(data[, 1])^2/n)/n
sigma22 = (sum(data[, 2]^2) - sum(data[, 2])^2/n)/n
sigma12 = (sum(data[, 1]*data[, 2]) - sum(data[, 1])*sum(data[, 2])/n)/n
```

The initial MLE estimates of the parameters are shown below, along with the initial guess for $w_{2,9}, w_{2,10}, w_{2,9}^2$, and $w_{2,10}^2$.

k	mu1	sigma11	mu2	sigma22	sigma12	w2,9	w2,9^2	s2,10	w2,10^2
0	13.00	40.20	11.90	58.49	25.90	0.00	0.00	0.00	0.00

Since μ_1 and σ_{11} do not depend on the missing values, they will remain constant. The other values will be updated on each iteration of the EM algorithm. The results from each iteration are shown in the following table.

```r
STOP = FALSE
iter = 1

while(!STOP && iter < MAX_ITERATIONS) {
  #Use the conditional expectation to get estimates of missing data.
  w29 = mu2 + sigma12/sigma11*(data[9, 1] - mu2)
  w210 = mu2 + sigma12/sigma11*(data[10, 1] - mu2)

  w29.sq = sigma22*(1 - sigma12^2/(sigma11*sigma22)) + w29^2
  w210.sq = sigma22*(1 - sigma12^2/(sigma11*sigma22)) + w210^2

  if(abs(w29 - data[9, 2]) < EPSILON && abs(w210 - data[10, 2]) < EPSILON) {
    STOP = TRUE
  }

  #Update the data.
  data[9, 2] = w29
  data[10, 2] = w210

  #Find new MLE estimates using the updated data.
  mu2 = sum(data[, 2])/n
  sigma22 = (sum(c(data[1:8, 2]^2, w29.sq, w210.sq)) - sum(data[, 2])^2/n)/n
  sigma12 = (sum(data[, 1]*data[, 2]) - sum(data[, 1])*sum(data[, 2])/n)/n

  tab <- rbind(tab, (c(iter, format(c(mu1, sigma11, mu2, sigma22, sigma12,
                                      w29, w29.sq, w210, w210.sq),
                       digits = 8))))

  iter <- iter + 1
}
```

k	mu2	sigma22	sigma12	w2,9	w2,9^2	s2,10	w2,10^2
0	11.90	58.49	25.90	0.00	0.00	0.00	0.00
1	14.164030	33.802127	21.887363	10.031592	142.436023	12.608706	200.782663
2	14.388267	28.649345	21.359035	11.352413	150.762576	13.530260	204.953233
3	14.417603	27.610363	21.289849	11.525377	150.135194	13.650654	203.641236
4	14.421529	27.401481	21.280620	11.548450	149.701995	13.666843	203.117894
5	14.422056	27.359556	21.279383	11.551542	149.574302	13.669016	202.978193
6	14.422126	27.351151	21.279218	11.551956	149.543264	13.669308	202.945547
7	14.422136	27.349468	21.279195	11.552012	149.536319	13.669347	202.938387
8	14.422137	27.349131	21.279192	11.552019	149.534832	13.669352	202.936870
9	14.422137	27.349063	21.279192	11.552020	149.534521	13.669353	202.936555
10	14.422137	27.349050	21.279192	11.552020	149.534457	13.669353	202.936491

Note, the values for μ_1 and σ_{11} are dropped from the table since they are constant. The algorithm stops once the estimated values for $w_{2,9}$ and $w_{2,10}$ converge. For this run, we used $\epsilon = 10^{-6}$ and stop once $|w_{2,i}^{(k)} - w_{2,i}^{(k-1)}| < \epsilon$ for $i = 9$ and 10.

We find that the MLE estimates are

$$\begin{aligned}
\hat{\mu}_1 &= 13.000000 \\
\hat{\mu}_2 &= 14.422137 \\
\hat{\sigma}_{11} &= 40.200000 \\
\hat{\sigma}_{12} &= 21.279192 \\
\hat{\sigma}_{22} &= 27.349050
\end{aligned}$$