Homework 5

Ziyang (Bob) Ding

24 February, 2020

Problem 1

Suppose we observe independent Bernoulli variables $X_1, \ldots, X_n$, which depend on unobservable variables Z_i distributed independently as $N\left(\alpha, \sigma^2\right)$, where

$$X_i = \begin{cases} 0 & \text{if } Z_i \leq u \\ 1 & \text{if } Z_i > u \end{cases}$$

Assuming that u is known, we are interested in obtaining MLEs of α and σ^2

- (a) **Show that the likelihood function is**

$$p^S(1-p)^{n-S}$$

where $S = \sum x_i$ and

$$p = \operatorname{Pr}\left(Z_i > u\right) = \Phi\left(\frac{\alpha - u}{\sigma}\right)$$

- (b) **If we consider $z_1, \ldots, z_n$ to be missing data, show that the expected complete-data loglikelihood is**

$$-\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[E\left(Z_i^2|x_i\right) - 2\alpha E\left(Z_i|x_i\right) + \alpha^2\right]$$

- (c) **Show that the EM sequence is given by**

$$\hat{\alpha}_{(j+1)} = \frac{1}{n}\sum_{i=1}^{n} t_i\left(\hat{\alpha}_{(\hat{j})}, \hat{\sigma}_{(j)}^2\right)$$
$$\hat{\sigma}_{(j+1)}^2 = \frac{1}{n}\left[\sum_{i=1}^{n} v_i\left(\hat{\alpha}_{(j)}, \hat{\sigma}_{(j)}^2\right) - \frac{1}{n}\sum_{i=1}^{n}\left(t_i(\hat{\alpha}_{(j)}, \hat{\sigma}_{(j)}^2)\right)^2\right]$$

where $t_i\left(\alpha, \sigma^2\right) = E\left(Z_i|x_i, \alpha, \sigma^2\right)$ and $v_i\left(\alpha, \sigma^2\right) = E\left(Z_i^2|x_i, \alpha, \sigma^2\right)$

- (d) **Show that**

$$E\left(Z_i|x_i, \alpha, \sigma^2\right) = \alpha + \sigma H_i\left(\frac{u-\alpha}{\sigma}\right)$$
$$E\left(Z_i^2|x_i, \alpha, \sigma^2\right) = \alpha^2 + \sigma^2 + \sigma(u-\alpha)H_i\left(\frac{u-\alpha}{\sigma}\right)$$

where

$$H_i(t) = \begin{cases} \frac{\phi(t)}{1-\phi(t)} & \text{if } X_i = 1 \\ -\frac{\phi(t)}{\Phi(t)} & \text{if } X_i = 0 \end{cases}$$

Proof

- **(a)**

Because X_i follows bernouli distribution, we know that the sum follows binomial distribution. This is the likelihood function.

- **(b)**

$$
\begin{aligned}
\mathbb{E}_{Z|X,\theta=(\alpha,\sigma)}[\ell(X,Z|\theta)] &= \mathbb{E}_{Z|X,\theta}[\log \mathbb{P}(X|\theta,Z)\mathbb{P}(Z|\theta)] \\
&= \mathbb{E}_{Z|X,\theta}[\log \mathbb{P}(X|\theta,Z) + \log \mathbb{P}(Z|\theta)] \\
&= \mathbb{E}_{Z|X,\theta}[\log \mathbb{P}(X|\theta,Z)] + \mathbb{E}_{Z|X,\theta}[\log \mathbb{P}(Z|\theta)] \\
&= \sum_{i=1}^{n} \mathbb{E}_{z_i|X_i,\theta}[\log \mathbb{P}(X_i|\theta,z_i)] + \sum_{i=1}^{n} \mathbb{E}_{z_i|X_i,\theta}[\log \mathbb{P}(z_i|\theta)] \\
&= \sum_{i=1}^{n} \mathbb{E}_{z_i|X_i,\theta}[\log \mathbb{P}(X_i|\theta,z_i)] + \sum_{i=1}^{n} \mathbb{E}_{z_i|X_i,\theta}[-\log\sqrt{2\pi}\sigma + \frac{(z_i-\alpha)^2}{2\sigma^2}] \\
&= \sum_{i=1}^{n} \mathbb{E}_{z_i|X_i,\theta}[\log \mathbb{P}(X_i|\theta,z_i)] - n\log\sqrt{2\pi}\sigma + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\mathbb{E}[Z_i^2|x_i] - 2\alpha\mathbb{E}[Z_i|x_i] + \alpha^2) \\
&= -\frac{n}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\mathbb{E}[Z_i^2|x_i] - 2\alpha\mathbb{E}[Z_i|x_i] + \alpha^2)
\end{aligned}
$$

- **(c)**

$$
\begin{aligned}
Q(\hat{\alpha}_{(j)}, \hat{\sigma^2}_{(j)}) &:= -\frac{n}{2}\log 2\pi\hat{\sigma^2}_{(j)} + \frac{1}{2\hat{\sigma^2}_{(j)}}\sum_{i=1}^{n}(\mathbb{E}[Z_i^2|x_i] - 2\hat{\alpha}_{(j)}\mathbb{E}[Z_i|x_i] + \hat{\alpha}^2_{(j)}) \\
\hat{\alpha}_{(j+1)} &= arg\,max_{\hat{\alpha}_{(j)}} Q(\hat{\alpha}_{(j)}, \hat{\sigma^2}_{(j)}) \\
\hat{\sigma^2}_{(j+1)} &= arg\,max_{\hat{\sigma^2}_{(j)}} Q(\hat{\alpha}_{(j)}, \hat{\sigma^2}_{(j)}) \\
\frac{\partial Q(\hat{\alpha}_{(j)}, \hat{\sigma^2}_{(j)})}{\partial \hat{\alpha}_{(j)}} &= \frac{1}{2\hat{\sigma^2}_{(j)}}\sum_{i=1}^{n}[2\hat{\alpha}_{(j)} - 2\mathbb{E}[Z_i^2|x_i]] = 0 \\
\hat{\alpha}_{(j+1)} &= \frac{1}{n}\sum_{i=1}^{n} t_i\left(\hat{\alpha}_{(j)}, \hat{\sigma}^2_{(j)}\right) \\
\frac{\partial Q(\hat{\alpha}_{(j)}, \hat{\sigma^2}_{(j)})}{\partial \hat{\alpha}_{(j)}} &= -\frac{2\pi n}{2\times 2\pi\hat{\sigma^2}_{(j)}} - \frac{1}{2(\hat{\sigma^2}_{(j)})^2}\sum_{i=1}^{n}(\mathbb{E}[Z_i^2|x_i] - 2\hat{\alpha}_{(j)}\mathbb{E}[Z_i|x_i] + \hat{\alpha}^2_{(j)}) \\
&= -\frac{n}{2\hat{\sigma^2}_{(j)}} - \frac{1}{2(\hat{\sigma^2}_{(j)})^2}\sum_{i=1}^{n}(\mathbb{E}[Z_i^2|x_i] - \hat{\alpha}_{(j)}\mathbb{E}[Z_i|x_i] + \hat{\alpha}^2_{(j)}) \\
\hat{\alpha}_{(j+1)} &= \frac{1}{n}\left[\sum_{i=1}^{n} v_i\left(\hat{\alpha}_{(j)}, \hat{\sigma}^2_{(j)}\right) - \frac{1}{n}\left(\sum_{i=1}^{n} t_i\left(\hat{\alpha}_{(j)}, \hat{\sigma}^2_{(j)}\right)\right)^2\right]
\end{aligned}
$$

- **(d)**

Given the information of X_i, we can directly deduce that Z_i must be a truncated-normal distribution. Therefore, we just need to integrate Z_i from a normal distribution on the corresponding to the non-degenerated support.

First, suppose that $X_i = 0$:

$$\begin{aligned} E\left(Z_i | x_i = 0, \alpha, \sigma^2\right) &= \int_{-\infty}^{u} z \cdot \phi(z)\left(z, \alpha, \sigma^2\right) / \Phi\left(\frac{u-\alpha}{\sigma}\right) dz \\ &= \frac{1}{\Phi\left(\frac{u-\alpha}{\sigma}\right)} \int_{-\infty}^{\frac{u-\alpha}{\sigma}} (\sigma y + \alpha)\phi(y) dy \\ &= \alpha + \sigma \int_{-\infty}^{\frac{u-\alpha}{\sigma}} y\phi(y) dy / \phi\left(\frac{u-\alpha}{\sigma}\right) \end{aligned}$$

Because we know that

$$\int_{-\infty}^{t} y\phi(y)dy = \int_{-\infty}^{t} y \frac{1}{2x} e^{-\frac{y^2}{2}} dy = \frac{1}{2\pi} \int_{+\infty}^{\frac{t^2}{2}} e^{-\frac{y^2}{2}} d\frac{y^2}{2} = -\frac{1}{2\pi} e^{-\frac{t^2}{2}} = -\phi(t)$$

This enables us to further simplify and get that

$$E\left(Z_i | x_i = 0, \alpha, \sigma^2\right) = \alpha + \sigma \int_{-\infty}^{\frac{u-\alpha}{\sigma}} y\phi(y) dy / \phi\left(\frac{u-\alpha}{\sigma}\right) = \alpha - \sigma \cdot \frac{\phi\left(\frac{n-\alpha}{\sigma}\right)}{\Phi\left(\frac{u-\alpha}{\sigma}\right)}$$

Similarly, we can use the same transformation to obtain that

$$E\left(Z_i | x_i = 1, \alpha, \sigma^2\right) = \alpha + \sigma \frac{\phi(n-\alpha)}{1 - \Phi\left(\frac{n-\alpha}{\sigma}\right)}$$

Now, let's look at the quadratic term. First, we may assume that $X_i = 0$. We see that

$$\begin{aligned} E\left(Y_i^2 | x_i = 0, \alpha, \sigma^2\right) &= \int_{-\infty}^{\sigma-\alpha} y^2 \frac{1}{2\pi} e^{-\frac{y^2}{2}} dy \\ &= -\int_{-\infty}^{\frac{u-\alpha}{\sigma}} y \cdot \frac{1}{2\pi} de^{-\frac{y^2}{2}} \\ &= -\frac{1}{2\pi} \cdot \frac{u-\alpha}{\sigma} \cdot e^{-\frac{1}{2}\left(\frac{u-a}{\sigma}\right)^2} + \Phi(\frac{u-\alpha}{\sigma}) \end{aligned}$$

Similarly, we should get that

$$\begin{aligned} & E\left(Y_i^2 | x_i = 1, \alpha, \sigma^2\right) \\ =& E\left(\left(-Y_i\right)^2 | x_i = 1, \alpha, \sigma^2\right) \\ =& -\frac{1}{2\pi} \cdot \frac{\alpha-\mu}{\sigma} e^{-\frac{1}{2}\left(\frac{\alpha-\mu}{\sigma}\right)^2} + \Phi\left(\frac{\alpha-\mu}{\sigma}\right) \end{aligned}$$

Therefore, to summarize, we get

$$\begin{aligned} E\left(z_i^2 | x_i, \alpha, \sigma^2\right) &= E\left(\sigma^2\left(Y_i^2 - 2Y_i\alpha + \alpha^2\right) | x_i, \alpha, \sigma^2\right) \\ &= \alpha^2 + \sigma^2 + \sigma(u-\alpha) H_i\left(\frac{n-\alpha}{\sigma}\right) \end{aligned}$$

Problem 2

Revisit the missing data problem (#4) from Homework 4.

- (a) **Give complete data log likelihood and derive the EM updates.**
- (b) **Implement your EM algorithm and use it to find the MLE for Σ.**
- (c) **Use what you learned in HW3 to demonstrate the potential sensitivity of the EM algorithm to initialization.**

Proof

- **(a)**

For the sake of simplicity of notation, we call

$$\Omega := \Sigma^{-1}, \quad \Omega = \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix}$$

$$\begin{aligned}
p\left(y_{obs}, y_{mis} | z\right) &\propto |\Omega|^{-\frac{n}{z}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(w_{11}y_{i1}^2 + w_{22}y_{i2}^2 + 2w_{12}y_{i1}y_{i2}\right)\right\} \\
l(\Omega) &= \frac{n}{2}\log|\Omega| - \frac{1}{2}\sum_{i=1}^{n}\left(w_{11}y_{i1}^2 + w_{22}y_{i2}^2 + 2w_{12}y_{i1}y_{i2}\right) + C \\
\mathcal{P} = P\left(\Sigma | y_{obs}, y_{mis}\right) &= \frac{n+3}{2}\log|\Omega| - \frac{1}{2}\sum_{i=1}^{n}\left(w_{11}y_{i1}^2 + w_{22}y_{i2}^2 + 2w_{12}y_{i1}y_{i2}\right)
\end{aligned}$$

With the calculated $\mathcal{P}$, we can proceed into the hear of EM:

$$\begin{aligned}
Q\left(\Omega^{(t+1)} | \Omega^{(t)}\right) &= E_{y_{\text{mis}}} | y_{\text{obs},\Omega^{(t)}} \mathcal{P} \\
&= \frac{n+3}{2}\log|\Omega| - \frac{1}{2}\sum_{i=1}^{4}\left(w_{11}y_{i1}^2 + w_{22}y_{i2}^2 + 2w_{12}y_{i1}y_{i2}\right) \\
&\quad - \frac{1}{2}\sum_{i=5}^{8}\left(w_{11}y_{i1}^2 + w_{22}E\left(y_{i2}^2 | \Omega^{(t)}, y_{obs}\right) + 2w_{12}y_{i1}E\left(y_{i2} | \Omega^{(t)}, y_{obs}\right)\right) \\
&\quad - \frac{1}{2}\sum_{i=9}^{12}\left(w_{11}E\left(y_{i1}^2\right)_{i1}\left(\Omega^{(t)}, y_{obs}\right) + 2w_{12}E\left(y_{i1} | \Omega^{(t)}, y_{obs}\right)y_{i2} + w_{22}y_{i2}^2\right) \\
&= \frac{n+3}{2}\log|\Omega| - \frac{1}{2}E_t\left(\operatorname{tr}(\Omega\sum_{i=1}^{n} y_i y_i^{\top}) | \Omega^{(t)}, y_{obs}\right)
\end{aligned}$$

Then we want to maximize the expectation. Follow the procedure by doing derivative and setting it to 0, we can get the result:

$$\Omega^{(t+1)} = (n+3)E\left(\sum_{i=1}^{n} y_i y_i^{\top} | \Omega^{(t)}, y_{0bs}\right)^{-1}$$

Due to the fact that each one of these are linear operations, we can move interior element out and calculate expectation explicitly and separately:

$$
\begin{aligned}
E\left(y_{i1}|\Omega^{(t)}, y_{obs}\right) &= -\frac{w_{12}^{(n)}}{w_{11}^{(t)}} y_{i_2} \\
E\left(y_{i1}^2|\Omega^{(t)}, y_{obs}\right) &= \frac{w_{12}^{(t)^2}}{w_{11}^{(t)2}} y_{i_2}^2 + |\Omega|^{-1}\left(w_{22}^{(t)} - \frac{w_{12}^{(t)^2}}{w_{11}^{(t)}}\right) \\
E\left(y_{i1}|\Omega^{(t)}, y_{obs}\right) &= -\frac{w_{12}^{(n)}}{w_{11}^{(t)}} y_{i_2} \\
E\left(y_{i2}^1|\Omega^{(t)}, y_{obs}\right) &= \frac{w_{21}^{(t)^2}}{w_{22}^{(t)2}} y_{i_1}^2 + |\Omega|^{-2}\left(w_{11}^{(t)} - \frac{w_{21}^{(t)^2}}{w_{22}^{(t)}}\right)
\end{aligned}
$$

- **(b)**

We'd use what we've found in hw4 that the mode has 0.8, -0.8, 0. The reuslts are shown below in part c. And as we're doing three different values, the 3 outputs can shed light on sensitivity of initial points directly. So no need to read part (b) where I just defined functions and did basic testing. You can directly skip to part 3 to check results.

```
# Create Data Matrix
y4 = t(matrix(c(1,1,-1,-1,1,-1,1,-1), ncol = 2, byrow = F))
S4 = y4 %*% t(y4)

# Start Sampling, use code from previous homework
n = 10000
Sig11 = rep(0, n)
Sig22 = rep(0, n)
RHO = rep(0, n)
for (i in 1:10000){
  rw = rWishart(1,4,solve(S4))
  riw = solve(rw[,,1])
  rho = riw[1,2]/sqrt(riw[1,1]* riw[2,2])
  RHO[i] = rho
  Sig11[i] = riw[1,1]
  Sig22[i] = riw[2,2]
}

Sig0 = matrix(c(  mean(Sig11),                          sqrt(mean(Sig11)*mean(Sig22))*mean(RHO),
                  sqrt(mean(Sig11)*mean(Sig22))*mean(RHO),mean(Sig22) ), ncol = 2, nrow = 2)
```

```
# Define functions requred for E step
# Conditional Expectation
conE <- function(idx, obs, Sig){
  return(Sig[idx, 3-idx]*obs / Sig[3-idx, 3-idx])
}

# Conditional Variance
conVar <- function(idx, obs, Sig){
  return(Sig[idx,idx] - (Sig[idx, 3-idx])^2/Sig[3-idx,3-idx])
}

# Conditional Sum of Square Matrix
conS <- function(idx, obs, Sig){
  S <- matrix(0, nrow=2, ncol=2)
```

```r
  S[3-idx, 3-idx] <- obs^2
  S[idx  , 3-idx] <- obs*conE(idx, obs, Sig)
  S[3-idx, idx  ] <- obs*conE(idx, obs, Sig)
  S[idx  , idx  ] <- (conE(idx, obs, Sig))^2 + conVar(idx, obs, Sig)
  return(S)
}
```

```r
# Define EM
EM <- function(itr, Sig0, W0, W){
  for (j in 1:itr){
  W0 <- W
  Sig <- solve(W)
  St <- S4
  yobs1 <- c(2,2,-2,-2)
  yobs2 <- yobs1

  for (r in 1:4){
    St <- St + conS(2, yobs1[r], Sig)
  }
  for (r in 1:4){
    St <- St + conS(1, yobs2[r], Sig)
  }

  W <- solve(St) * 15
  }

  thing <- solve(W)
  Rho <- thing[1,2] / sqrt(thing[1,1]*thing[2,2])

  print("Sigma0")
  print(Sig0)

  print("Sigma:")
  print(thing)

  print("Rho:")
  print(Rho)

  print("-------------------------------------------")
}



itr = 5000
W0 <- solve(Sig0)
W <- W0

EM(itr, Sig0, W0, W)
```

```
## [1] "Sigma0"
##          [,1]     [,2]
## [1,] 4.065978 0.018718
## [2,] 0.018718 3.596986
## [1] "Sigma:"
```

```
##           [,1]     [,2]
## [1,] 2.133333 1.453622
## [2,] 1.453622 2.133333
## [1] "Rho:"
## [1] 0.6813851
## [1] "-------------------------------------------"
```

- **(c)**

Let's make a list of initial input value to try MLEs 0.8, -0.8, 0

```
# We will try several different off diagonal terms: 0.8, -0.8, 0

try <- seq(-0.8, 0.8, by = 0.8)

for(i in try){
  Sig0 <- matrix(c(1, i, i, 1), ncol = 2)
  W0 <- solve(Sig0)
  W <- W0
  EM(itr, Sig0, W0, W)
}
```

```
## [1] "Sigma0"
##      [,1] [,2]
## [1,]  1.0 -0.8
## [2,] -0.8  1.0
## [1] "Sigma:"
##           [,1]      [,2]
## [1,]  2.133333 -1.453622
## [2,] -1.453622  2.133333
## [1] "Rho:"
## [1] -0.6813851
## [1] "-------------------------------------------"
## [1] "Sigma0"
##      [,1] [,2]
## [1,]    1    0
## [2,]    0    1
## [1] "Sigma:"
##          [,1]     [,2]
## [1,] 1.818182 0.000000
## [2,] 0.000000 1.818182
## [1] "Rho:"
## [1] 0
## [1] "-------------------------------------------"
## [1] "Sigma0"
##      [,1] [,2]
## [1,]  1.0  0.8
## [2,]  0.8  1.0
## [1] "Sigma:"
##          [,1]     [,2]
## [1,] 2.133333 1.453622
## [2,] 1.453622 2.133333
## [1] "Rho:"
## [1] 0.6813851
```

```
## [1] "-----------------------------------------"
```

It is not hard to see that the initializing point is really sensitive to the resulting estimation. Plus, this is just a normal model. If we run into something with much wilder distribution and likelihood, the initialization could mess things up.