STAT 740, Fall 2017: Homework 3

1. **(Calculus refresher):** Let $x \sim N(\mu, \sigma^2)$. Use integration by parts and the following

$$\int x e^{-\frac{1}{2}x^2} dx = k - e^{-\frac{1}{2}x^2},$$

for some constant k to show

$$E(x^2|x > c) = \mu^2 + \sigma^2 + \sigma(c + \mu)\frac{\phi(\frac{c-\mu}{\sigma})}{1 - \Phi(\frac{c-\mu}{\sigma})}.$$

2. Use the previous result along with the result for $E(x|x > c)$ in Tim's notes to derive the E-M algorithm for right-censored normal data. Use your algorithm on the V.A. data to find the MLE of μ and σ; take the log of the event times first. The following code uses `survreg` to obtain the MLEs so you can verify you get the same thing. For E-M starting values, simply use the censored data sample mean and standard deviation, e.g. $\mu^0 =$ `mean(ltime)` and $\sigma^0 =$ `sd(ltime)`.

```
library(MASS) # has VA data
library(survival) # has survfit & survreg

# function to give "histogram" from right-censored data
chist=function(x,d,J){
 a=min(x-(max(x)-min(x))/10000); b=max(x); J=10
 delta=(b-a)/J
 t=seq(a,b,length=(J+1))
 f=summary(survfit(Surv(x,d)~1),times=t)
 y=c(0,rep(f$surv[1:J]-f$surv[2:(J+1)],1,each=2),0)/delta
 x=rep(t,1,each=2)
 plot(x,y,type="l",col="blue",xlab="censored data",
 ylab="density",ylim=c(0,max(y*1.1)))
}

VAs=subset(VA,prior==0)
attach(VAs)
ltime=log(stime)
f=survreg(Surv(ltime,status)~1,dist="gaussian") # the easy way!
chist(stime,status,10) # histogram
x=seq(0,600,length=1000)
lines(x,dlnorm(x,f$coef,f$scale),col="red") # plot of log-normal fit
summary(f) # MLEs of mu and sigma; compare to values you get via E-M
```

Extra credit: obtain the AIC treating the data as *log-normal* and compare to that obtained from an exponential fit. Be careful; you need to obtained the maximized likelihood on *the original untransformed data scale*, e.g. $L(\hat{\mu}, \hat{\sigma}|\mathbf{x}, \boldsymbol{\delta})$, not $L(\hat{\mu}, \hat{\sigma}|\log(\mathbf{x}), \boldsymbol{\delta})$ to compare to exponential. Which is better? I don't care how you get the AICs...

3. Use either the bootstrap or direct numerical differentiation (e.g. `numDeriv` package) to find the standard errors for $\hat{\mu}$ and $\hat{\sigma}$ in Problem 2. Compare to those obtained from `survreg`. If you use the bootstrap, a bootstrap sample is $\{(x_{i_1}, \delta_{i_1}), \ldots, (x_{i_n}, \delta_{i_n})\}$ for $(i_1, \ldots, i_n) \subset \{1, \ldots, n\}^n$. That is, sample the pairs (x_i, δ_i) randomly from all n pairs with replacement.

4. In the notes on E-M for a finite mixture of normals, recall that by definition $w_{i\bullet} = w_{i1} + \cdots + w_{iJ} = 1$, thus $w_{\bullet\bullet} = n$. Derive the E-M maximization step for $\pi^{t+1} = w^t_{\bullet j}/n$. First show that taking first partial derivative w.r.t. to π_j gives

$$\frac{w^t_{\bullet j}}{\pi_j} - \frac{w^t_{\bullet J}}{\pi_J} = 0,$$

where $\pi_J = 1 - \sum_{j=1}^{J-1} \pi_j$. Multiply both sides by π_j then sum over j to show $\frac{w^t_{\bullet J}}{\pi_J} = n$. Finally, solve for π_j.

Hint for problem 2:

Augmented log-likelihood is

$$\log L(\mu, \sigma^2 | \mathbf{x}, \mathbf{z}) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i:\delta_i=1} (y_i - \mu)^2 - \frac{1}{2\sigma^2} \sum_{i:\delta_i=0} (t_i - \mu)^2.$$

Taking expectation w.r.t. $[\{t_i : \delta_i = 0\} | \{y_i : \delta_i = 1\}, \mu^t, \sigma^t]$ gives

$$-\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i:\delta_i=1} (y_i - \mu)^2 - \frac{1}{2\sigma^2} \sum_{i:\delta_i=0} E_{\mu^t,\sigma^t}\{(t_i - \mu)^2\}.$$

Note that

$$E_{\mu^t,\sigma^t}\{(t_i - \mu)^2\} = E_{\mu^t,\sigma^t}\{t_i^2 - 2t_i\mu + \mu^2\} = E_{\mu^t,\sigma^t}(t_i^2) - 2\mu E_{\mu^t,\sigma^t}(t_i) + \mu^2.$$

Use the result for $E(t_i | t_i > y_i, \mu^t, \sigma^t)$ and $E(t_i^2 | t_i > y_i, \mu^t, \sigma^t)$ in the notes and in problem 1. I redid my E-M notes for censored exponential data to be more clear, just refresh your browser.