Assignment 7

Tushar Ponkshe, tvp2110

November 30, 2018

(i)

```
moretti = read.csv('C:/Users/tusha/OneDrive/Documents/Columbia Fall 2018 Courses/moretti.csv')

poisLoglik = function(lambda, data) {
  likelihood = sum(log((lambda^data)*exp(-lambda)/factorial(data)))
}

ans = poisLoglik(lambda = 1, data = c(1, 0, 0, 1, 1))
ans
```

```
## [1] -5
```

(ii)

```
count_new_genres = function(year) {
  return(length(unique((moretti[moretti$Begin == year,])$Name)))
}
count_new_genres(1803)
```

```
## [1] 0
```

```
count_new_genres(1850)
```

```
## [1] 3
```

We get the correct values for the years 1803 and 1850

(iii)

```
i = c(1740:1900)
new_genres = rep(NA, length(i))
for (j in 1:length(new_genres)) {
  new_genres[j] = count_new_genres(i[j])
}
names(new_genres) <-c(1740:1900)
new_genres["1803"]
```

```
## 1803
##    0
```

```
new_genres["1850"]
```

```
## 1850
##    3
```

We get the correct values for the years 1803 and 1850

(iv)

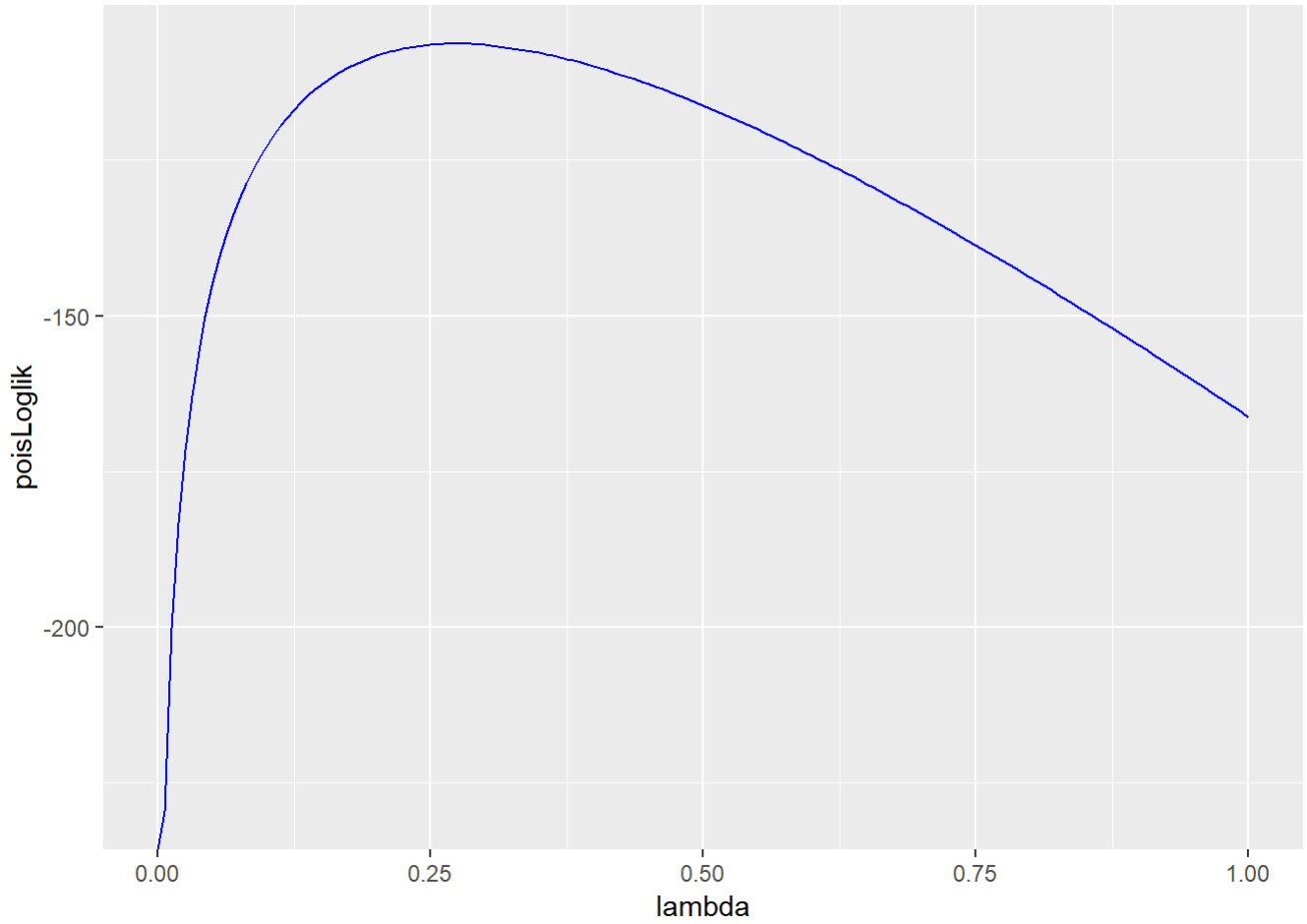
```
library(ggplot2)
lambda = seq(0, 1, by = 1/160)


poisLoglik = function(lambda, data) {
  likelihood = sum(log((lambda^data)*exp(-lambda)/factorial(data)))
}

y=rep(NA,length(new_genres))
for (i in 1:length(new_genres)) {
  y[i]=poisLoglik(lambda[i],new_genres)
}

new.data = data.frame(lambda, y)

ggplot(data=new.data, aes(x=lambda,y=y)) + xlab("lambda") + ylab("poisLoglik") + geom_line(color
="blue")
```



We can see that the poisLoglik function attains maximum value around $\lambda = 0.273$

v

```
poisLoglik = function(lambda, data) {
  likelihood = -sum(log((lambda^data)*exp(-lambda)/factorial(data)))
  return(likelihood)
}
#arg = c(lambda, new_genres)
nlm(poisLoglik, c(0.1), data=new_genres)$estimate
```

```
## [1] 0.2732914
```

We can see that $\lambda = 0.2732914$ maximizes the likelihood function

vi

```
intergenre_intervals = diff(moretti$Begin,lag = 1)

head(intergenre_intervals)
```

```
## [1]  8 11  7  2  2  3
```

```
mean(intergenre_intervals)
```

```
## [1] 3.44186
```

```
sd(intergenre_intervals)
```

```
## [1] 3.705224
```

```
coef.var = sd(intergenre_intervals)/mean(intergenre_intervals)
coef.var
```

```
## [1] 1.076518
```

vii (a)

```
intervals = function(vec.num){
  names(vec.num) = c(1740:1900)
  return_diff = c()
  for(i in 1:length(vec.num)){
    return_diff = c(return_diff,rep((names(vec.num[i])),vec.num[i]))
  }
  return(diff(as.numeric(return_diff), lag = 1))
}

intervals(new_genres)
```

```
##  [1]  8 11  7  2  2  3 16  1  1  9  4  4  6  8  3  1  2  2  0  2  6  1  7
## [24]  0  1  1  1  1  0  0  1  6 11  3  1  0  1  3  8  1  0  3  0
```

```
intergenre_intervals
```

```
##  [1]  8 11  7  2  2  3 16  1  1  9  4  4  6  8  3  1  2  2  0  2  6  1  7
## [24]  0  1  1  1  1  0  0  1  6 11  3  1  0  1  3  8  1  0  3  0
```

```
all.equal(intergenre_intervals, intervals(new_genres))
```

```
## [1] TRUE
```

We see that when we pass new_genres as argument we get intergenre_intervals

vii (b)

```
coef.variantion = function(gg) {
  return(sd(gg)/mean(gg))
}

pois.simu = function(years.num, genres.num) {
  store.pois = rpois(years.num, genres.num)
  store.intervals = intervals(store.pois)
  return(list(store.intervals, coef.variantion(store.intervals)))
}

pois.simu(161, 0.273)
```

```
## [[1]]
##  [1] 10  4  1  2  0  4  1  2  3  1  1  1  8 12  1  3 11  4  9  1  7  2  1
## [24]  3  4  8  1  2  1  1 10  4 12  1  3  8
##
## [[2]]
## [1] 0.9019524
```

```
mean(pois.simu(161, 0.273)[[1]])
```

```
## [1] 3.452381
```

We can see that the mean is generally between 3 and 4.

viii

```
coef.var.vec = c()

for (i in 1:10000) {
  coef.var.vec[i] = pois.simu(161, 0.273)[[2]]
}
mean(coef.var.vec > coef.var)
```

```
## [1] 0.2304
```

```
#(sum(coef.var.vec) - sum(coef.var))/10000
```

ix The fraction 0.22 does not give us enough evidence for saying that genres tend to appear together in bursts. By running the simulation 10,000 times to see if we observe a Poison process (randomness), we found clusters only about 22 percent of the time which is not conclusive to say that genres appear in clusters.