hw7_yw3204

wyh

11/25/2018

i.

```
poisLoglik <- function(l, data) {
  num <- (l ** data) * exp(-l)
  deno <- factorial(data)
  ratio <- num / deno
  res <- sum(log(ratio))
  return(res)
}

poisLoglik(1, c(1, 0, 0, 1, 1))
```

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

ii.

```
# read data
moretti <- read.csv("moretti.csv")

count_new_genres <- function(year) {
  return(sum(moretti$Begin == year))
}

count_new_genres(1803)
```

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

```
count_new_genres(1850)
```

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

The values are 0 and 3 seperately.

iii.

```
new_genres <- c()

for(i in c(1740:1900)) {
  new_genres <- c(new_genres, count_new_genres(i))
}

new_genres[64]
```

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

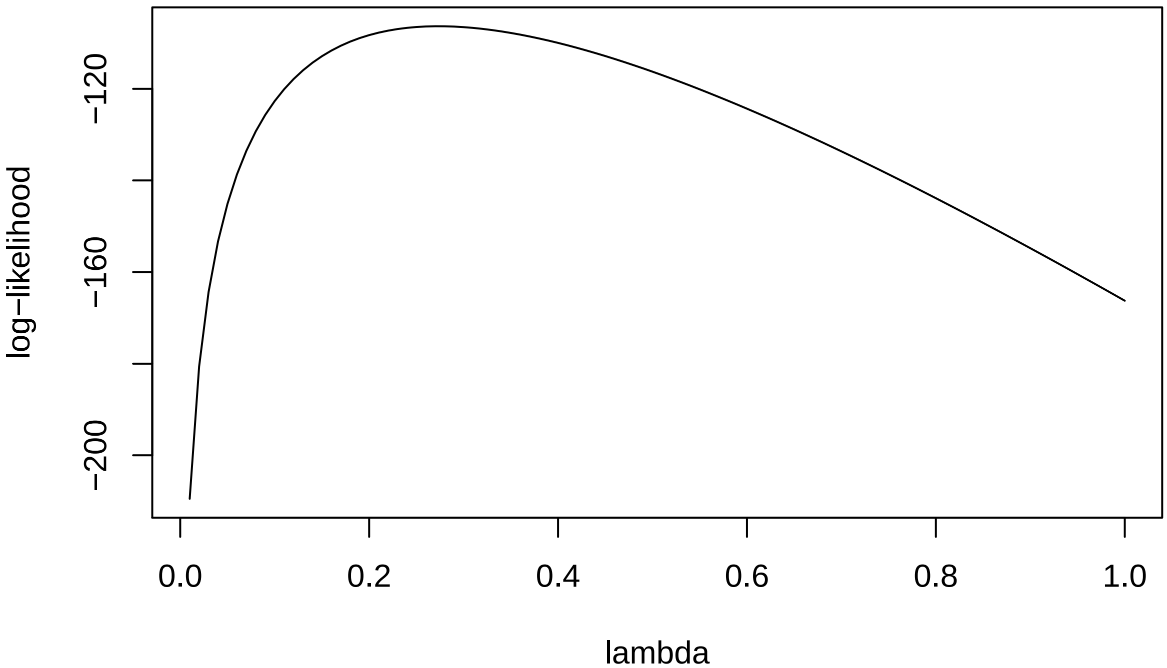
```r
new_genres[111]
```

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

The positions in the vector correspond to the years 1803 and 1850 are 64 and 111. The values should be 0 and 3 which are exactly what in the corresponding positions in vector new_genres.

iv.

```r
lambdas <- seq(0.01, 1, 0.01)
logs <- c()
for(l in lambdas) {
  logs <- c(logs, poisLoglik(l, new_genres))
}
plot(lambdas, logs, type = "l", xlab = "lambda", ylab = "log-likelihood")
```



The maximum is reached at around 0.2.

v.

```r
negPoisLoglik <- function(l, data = new_genres) {
  num <- (l ** data) * exp(-l)
  deno <- factorial(data)
  ratio <- num / deno
  res <- sum(log(ratio))
  return(-res)
}

nlm(negPoisLoglik, p = 1)
```

```
## Warning in log(ratio): NaNs produced
```

```
## Warning in nlm(negPoisLoglik, p = 1): NA/Inf replaced by maximum positive
## value

## Warning in log(ratio): NaNs produced

## Warning in nlm(negPoisLoglik, p = 1): NA/Inf replaced by maximum positive
## value

## Warning in log(ratio): NaNs produced

## Warning in nlm(negPoisLoglik, p = 1): NA/Inf replaced by maximum positive
## value

## Warning in log(ratio): NaNs produced

## Warning in nlm(negPoisLoglik, p = 1): NA/Inf replaced by maximum positive
## value

## Warning in log(ratio): NaNs produced

## Warning in nlm(negPoisLoglik, p = 1): NA/Inf replaced by maximum positive
## value

## Warning in log(ratio): NaNs produced

## Warning in nlm(negPoisLoglik, p = 1): NA/Inf replaced by maximum positive
## value

## $minimum
## [1] 106.3349
##
## $estimate
## [1] 0.2732914
##
## $gradient
## [1] 2.842171e-07
##
## $code
## [1] 1
##
## $iterations
## [1] 11
```

According to the result above, the minimum of the negative log-likelihhod function is reached at around 0.273. Thus, the maximum of the log-likelihhod function is reached at around 0.273.

vi.

```
# store the time interval
year_lap <- c()

# take first order difference
year_lap <- diff(moretti$Begin)

year_lap
```

```
##  [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
```

```r
mean(year_lap)
```

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

```r
sd(year_lap)
```

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

```r
coe_var <- sd(year_lap)/mean(year_lap)
```

vii.

a.

```r
cal_interval <- function(vec) {
  inds <- c()
  for(i in c(1:length(vec))) {
    if(vec[i] != 0) {
      inds <- c(inds, rep(i, vec[i]))
    }
  }

  intergenre_intervals <- diff(inds)
  return(intergenre_intervals)
}

cal_interval(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
```

b.

```r
cal_CoeVar <- function(num_ys, lambda) {
  n_genres <- rpois(num_ys, lambda)

  intervals <- cal_interval(n_genres)

  coe_var <- sd(intervals)/mean(intervals)

  return(list(intervals, coe_var))
}

cal_CoeVar(161, 0.273)
```

```
## [[1]]
##  [1] 10  3  4  1  1  2  8  1  0  1  8  1  4  0  2  6  6  3  3  4  2  0  3
## [24]  6  1  5  8  2  1  7  0  2  5  8  3  1  2  5  5  1  1  1  5  4  2  1
##
## [[2]]
## [1] 0.8029082
```

```r
# chek teh mean
# mean(cal_CoeVar(162, 0.273)[[1]])
```

viii.

```r
coes <- c()

for(i in c(1:10000)) {
  coes <- c(coes, cal_CoeVar(161, 0.273)[[2]])
}

# fraction of simulations having a higher coefficient of variation than Moretti's data
sum(coes > coe_var) / length(coes)
```

```
## [1] 0.2332
```

ix.

The proportion from the last question is relatively small which means the coefficient of variation for the moretti's data is higher than that of the supposed Poisson process. That's to say, the appearance of genres is not following a Poisson process. And thus, it doesn't have stationary and independent increment which means genres tends to appear together in burst.