Stochastic Models Assignment 4

16343261

11/08/2023

Before we begin this assignment, I will load in the data as required.

```
library(BayesLCA)
library(poLCA)
library(ClickClust)
library(clickstream)
library(seriation)
A4_clickstreams <- readRDS("~/Stochastic Models/A4_clickstreams.rds")
load("~/Stochastic Models/Brexitvotes.Rdata")
div_data = data.frame(divisions)
div_data[which(div_data=='aye_vote', arr.ind = T)] = 1
div_data[which(div_data!='1', arr.ind = T)] = 2
div_data = data.frame(sapply(div_data,as.numeric))
f = cbind(X1108904, X1108905, X1108906, X1108906,
          X1108907, X1107737, X1105521, X1105524,
          X1105526, X1105527, X1105529, X1105530,
          X1105532, X1105533, X1105759) ~ 1
```

Question 1.

To investigate this by fitting a number of LCA models and investigating our brexit variable.

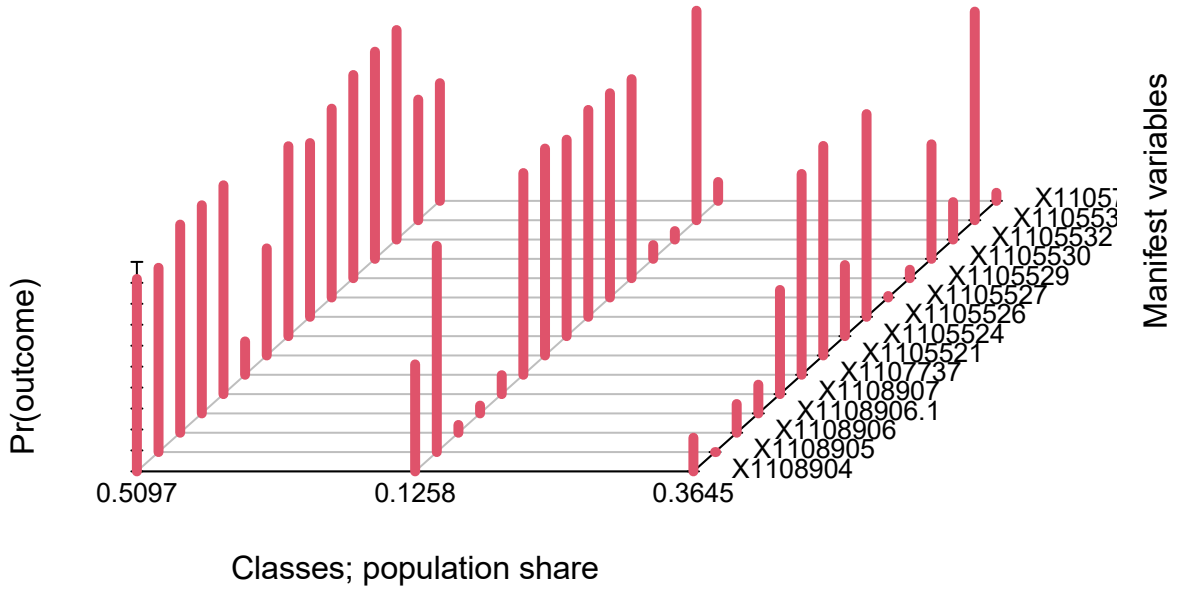
```
# Fit a G = 2 model
fit2 = poLCA(f,div_data,2)
# Fit a G = 3 model
fit3 = poLCA(f,div_data,3)
```

Our findings suggest that House of Commons politicians voting on Brexit-related matters can be grouped into clusters (latent classes) based on their voting behavior.

We find our classes fall into the following population shares: 0.425 0.4244 0.1506.

Now lets plot our fit:

```
plot(fit3)
```



Here we see that each cluster illustrates where politicians exhibit similar voting behavior.

Question 2.

To further investigate how many clusters there are, we will have to investigate using the AIC and the BIC. We will fit our models with higher and higher clusters until our AIC and BIC increases. The model with the lowest AIC and BIC suggests the model we should pick.

```
# Fit a G = 4 model
fit4 = poLCA(f,div_data,4)
# Fit a G = 5 model
fit5 = poLCA(f,div_data,5)
# Fit a G = 6 model
fit6 = poLCA(f,div_data,6)
```

```
b <- fit2$bic
c <- fit3$bic
d <- fit4$bic
e <- fit5$bic
f <- fit6$bic

i <- fit2$aic
j <- fit3$aic
k <- fit4$aic
```

```
l <- fit5$aic
m <- fit6$aic

bic <- c(b, c, d, e, f)
aic <- c(i, j, k, l, m)

dat4 <- as.data.frame(bic)
dat4$aic <- aic

dat4
```

```
##        bic      aic
## 1 7748.991 7610.783
## 2 7019.968 6810.426
## 3 6896.299 6615.424
## 4 6283.551 5931.342
## 5 6236.956 5813.414
```

From this data set, we've found clusters = 5 results in our best cluster findings as our AIC and BIC suggests the best fit. From my analysis, 5 clusters illustrates considerable groups of political beliefs. Next we will evaluate the clusters we find. I am currently unsure why this is now showing different AIC and BIC values but for the purposes of this analysis I have used my initial results which showed an increase at 6 clusters.

```
labels
```

```
##  [1] "Nick Boles's motion D (Common Market 2.0)"
##  [2] "Mr Clarke's motion C (Customs Union)"
##  [3] "Peter Kyle's motion E (Confirmatory public vote)"
##  [4] "Joanna Cherry's motion G (Parliamentary Supremacy)"
##  [5] "United Kingdom's withdrawal from the European Union"
##  [6] "Mr Baron's motion B (No deal)"
##  [7] "Nick Boles's motion D (Common market 2.0)"
##  [8] "George Eustice's motion H (EFTA and EEA)"
##  [9] "Mr Clarke's motion J (Customs union)"
## [10] "Jeremy Corbyn's motion K (Labour's alternative plan)"
## [11] "Joanna Cherry's motion L (Revocation to avoid no deal)"
## [12] "Margaret Beckett's motion M (Confirmatory public vote)"
## [13] "Mr Fysh's motion O (Contingent preferential arrangements)"
## [14] "Draft European Union (Withdrawal) Act 2018 (Exit Day) (Amendment) Regulations 2019"
```

```
fit5
```

```
## Conditional item response (column) probabilities,
##  by outcome variable, for each class (row)
##
## $X1108904
##            Pr(1)  Pr(2)
## class 1:  0.4852 0.5148
## class 2:  0.8551 0.1449
## class 3:  0.5934 0.4066
## class 4:  0.0000 1.0000
```

```
## class 5:  0.0086 0.9914
##
## $X1108905
##            Pr(1)  Pr(2)
## class 1:  0.0363 0.9637
## class 2:  1.0000 0.0000
## class 3:  0.7757 0.2243
## class 4:  0.0000 1.0000
## class 5:  0.0508 0.9492
##
## $X1108906
##            Pr(1)  Pr(2)
## class 1:  0.9669 0.0331
## class 2:  1.0000 0.0000
## class 3:  0.0000 1.0000
## class 4:  0.0000 1.0000
## class 5:  0.0094 0.9906
##
## $X1108906
##            Pr(1)  Pr(2)
## class 1:  0.9669 0.0331
## class 2:  1.0000 0.0000
## class 3:  0.0000 1.0000
## class 4:  0.0000 1.0000
## class 5:  0.0094 0.9906
##
## $X1108907
##            Pr(1)  Pr(2)
## class 1:  0.8901 0.1099
## class 2:  0.5598 0.4402
## class 3:  0.0760 0.9240
## class 4:  0.0000 1.0000
## class 5:  0.0000 1.0000
##
## $X1107737
##            Pr(1)  Pr(2)
## class 1:  0.0368 0.9632
## class 2:  0.0300 0.9700
## class 3:  0.5878 0.4122
## class 4:  0.7301 0.2699
## class 5:  1.0000 0.0000
##
## $X1105521
##            Pr(1)  Pr(2)
## class 1:  0.0245 0.9755
## class 2:  0.0000 1.0000
## class 3:  0.0482 0.9518
## class 4:  0.8481 0.1519
## class 5:  0.1190 0.8810
##
## $X1105524
##            Pr(1)  Pr(2)
## class 1:  0.0613 0.9387
## class 2:  0.6549 0.3451
```

```
## class 3:  0.6365 0.3635
## class 4:  0.0000 1.0000
## class 5:  0.0000 1.0000
##
## $X1105526
##            Pr(1)  Pr(2)
## class 1:  0.0245 0.9755
## class 2:  0.0300 0.9700
## class 3:  0.3776 0.6224
## class 4:  0.0776 0.9224
## class 5:  0.1093 0.8907
##
## $X1105527
##            Pr(1)  Pr(2)
## class 1:  0.0247 0.9753
## class 2:  0.9947 0.0053
## class 3:  0.7686 0.2314
## class 4:  0.0000 1.0000
## class 5:  0.0000 1.0000
##
## $X1105529
##            Pr(1)  Pr(2)
## class 1:  0.0614 0.9386
## class 2:  0.9598 0.0402
## class 3:  0.4685 0.5315
## class 4:  0.0000 1.0000
## class 5:  0.0093 0.9907
##
## $X1105530
##            Pr(1)  Pr(2)
## class 1:  0.9149 0.0851
## class 2:  0.5098 0.4902
## class 3:  0.0876 0.9124
## class 4:  0.0000 1.0000
## class 5:  0.0000 1.0000
##
## $X1105532
##            Pr(1)  Pr(2)
## class 1:  0.9430 0.0570
## class 2:  0.9548 0.0452
## class 3:  0.0000 1.0000
## class 4:  0.0000 1.0000
## class 5:  0.0000 1.0000
##
## $X1105533
##            Pr(1)  Pr(2)
## class 1:  0.0000 1.0000
## class 2:  0.0000 1.0000
## class 3:  0.0725 0.9275
## class 4:  0.7578 0.2422
## class 5:  0.0630 0.9370
##
## $X1105759
##            Pr(1)  Pr(2)
```

```
## class 1:  0.8904 0.1096
## class 2:  0.9698 0.0302
## class 3:  0.8911 0.1089
## class 4:  0.0718 0.9282
## class 5:  0.8095 0.1905
##
## Estimated class population shares
##  0.128 0.3135 0.1305 0.2612 0.1668
##
## Predicted class memberships (by modal posterior prob.)
##  0.1285 0.3135 0.1301 0.2555 0.1724
##
## =========================================================
## Fit for 5 latent classes:
## =========================================================
## number of observations: 638
## number of estimated parameters: 79
## residual degrees of freedom: 559
## maximum log-likelihood: -2886.671
##
## AIC(5): 5931.342
## BIC(5): 6283.551
## G^2(5): 1139.044 (Likelihood ratio/deviance statistic)
## X^2(5): 369570496 (Chi-square goodness of fit)
##
```

Cluster 1:

This cluster has a high probability of voting "yes" or split for most motions, except for a few cases where the probability of "other" is higher. It seems to represent politicians who tend to vote "yes" on various Brexit-related motions.

Cluster 2:

This cluster has a high probability of voting "yes" or "other" for a majority of motions. It seems to represent politicians who are unified in consistently voting between "yes" or "others" on various Brexit-related motions.

Cluster 3:

This cluster primarily votes "other" for our brexit motions, with 12 out of the 14 motions voted other. It seems to represent politicians who vote "other" on different Brexit-related motions.

Cluster 4:

This cluster has a high probability of voting "other" for many motions. It seems to represent politicians who often vote "other" on various Brexit-related motions.

Cluster 5:

This cluster's votes are split for these motions with 5 motions with 50% split and many of the remaining motions switching between yes and other votes It seems to represent politicians who arfe thorn on specific Brexit-related motions.

Cluster 1 is our largest cluster an estimated population share of approximately 30.22%, Cluster 2 has 13.01%, Cluster 3 has 18.62%, Cluster 4 has 26.08%, and Cluster 5 is our smallest cluster with 12.06%.

Question 3.

Let's load in our clickstream data.

```
A4_clickstreams <- readRDS("~/Stochastic Models/A4_clickstreams.rds")
dat = as.clickstreams(A4_clickstreams, sep = ',', header = TRUE)
```

Now we will will fit our Markov model and extract our transition matrix as necessary.

```
#Fit Markov chain model to clickstream data
fit <- fitMarkovChain(dat, order = 1)
#Extract the transition matrix
P<-t(fit@transitions[[1]])
P<-P[as.character(1:5),as.character(1:5)]
P<-as.matrix(P)
round(P,3)
```

```
##       1     2     3     4     5
## 1 0.138 0.182 0.347 0.115 0.217
## 2 0.118 0.406 0.175 0.173 0.128
## 3 0.133 0.098 0.472 0.229 0.068
## 4 0.114 0.137 0.188 0.489 0.072
## 5 0.474 0.099 0.149 0.118 0.160
```

Ignoring self transitions, our 3 highest transitions are as follows:

- Main Page to Weather (0.474)
- Main Page to International News (0.347)
- Sports to International News (0.229)

It is clear that for an average user reaching the sight that when landing on the main site, their 2 most commonly selected pages are Weather followed by International News. A second observation is that sports page fans consistently transition to the international news page.

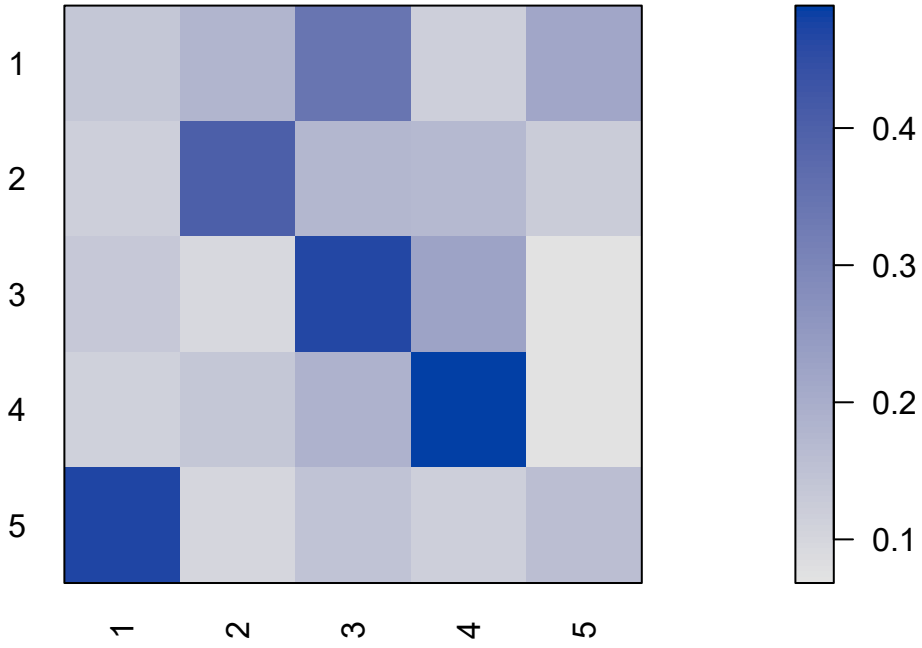
Our 3 lowest transitions are as follows:

- Weather to National News (0.099)
- Sports to Weather (0.072)
- International News to Weather(0.068)

Here we see weather is likely the least popular page to visit given someone has viewed either the sports or international news page. Lastly, weather readers rarely read national news.

Now let's plot our matrix.

```
pimage(P)
```




```
v<-rep(0,5)
v[1]<-1
for (i in 1:1000)
{
    v <- v%*%P
}
#Find stationary distribution using eigenvector
pi_st <- Re(eigen(t(P))$vector[,1])
pi_st <- pi_st/sum(pi_st)
pi_st
```

```
## [1] 0.1657541 0.1761821 0.2898466 0.2534753 0.1147419
```

Main Page: 16.58% The long-term distribution indicates that the Markov chain is expected to spend about 16.58% of its time on the "Main Page."

National News: 17.62% This value suggests that approximately 17.62% of the long-term behavior of the Markov chain is expected to be in the "National News" state.

International News: 28.98% The stationary distribution implies that a significant portion, around 28.98%, of the Markov chain's long-term behavior will be in the "International News" state.

Sport: 25.35% About 25.35% of the time, the Markov chain is expected to be in the "Sport" state.

Weather: 11.47% Finally, the long-term distribution suggests that the Markov chain would spend around 11.47% of its time in the "Weather" state.

Question 4.

We will estimate our transition matrices M1 and M2 as needed.

```
M_1 <- A4_clickstreams[1:300]
M_2 <- A4_clickstreams[301:600]
dat_M1 = as.clickstreams(M_1, sep = ',', header = TRUE)
dat_M2 = as.clickstreams(M_2, sep = ',', header = TRUE)

#Fit Markov chain model to clickstream data
fit_M1 <- fitMarkovChain(dat_M1, order = 1)
fit_M2 <- fitMarkovChain(dat_M2, order = 1)

#Extract the transition matrix
P_M1<-t(fit_M1@transitions[[1]])
P_M1<-P_M1[as.character(1:5),as.character(1:5)]
P_M1<-as.matrix(P_M1)
round(P_M1,2)
```

```
##      1    2    3    4    5
## 1 0.13 0.13 0.36 0.14 0.24
## 2 0.08 0.47 0.22 0.14 0.10
## 3 0.14 0.07 0.50 0.21 0.08
## 4 0.16 0.14 0.19 0.44 0.08
## 5 0.51 0.10 0.13 0.14 0.12
```

```
P_M2<-t(fit_M2@transitions[[1]])
P_M2<-P_M2[as.character(1:5),as.character(1:5)]
P_M2<-as.matrix(P_M2)
round(P_M2,3)
```

```
##       1     2     3     4     5
## 1 0.150 0.236 0.328 0.092 0.194
## 2 0.153 0.351 0.136 0.207 0.153
## 3 0.120 0.130 0.434 0.257 0.059
## 4 0.077 0.136 0.189 0.529 0.069
## 5 0.435 0.097 0.169 0.097 0.201
```

‘

```
v1<-rep(0,5)
v1[1]<-1
for (i in 1:1000)
{
    v_M1 <- v1%*%P_M1
}
#Find stationary distribution using eigenvector
pi_st_M1 <- Re(eigen(t(P_M1))$vector[,1])
pi_st_M1 <- pi_st_M1/sum(pi_st_M1)
pi_st_M1
```

```
## [1] 0.1763836 0.1663360 0.3166229 0.2273645 0.1132931
```

```r
for (i in 1:1000)
{
    v_M2 <- v1%*%P_M2
}
#Find stationary distribution using eigenvector
pi_st_M2 <- Re(eigen(t(P_M2))$vector[,1])
pi_st_M2 <- pi_st_M2/sum(pi_st_M2)
pi_st_M2
```

```
## [1] 0.1555000 0.1853596 0.2627642 0.2795504 0.1168257
```

Main Page Layout 1: 17.64% Layout 2: 15.56%

Here we see that the Markov chain is expected to spend roughly the same time as the initial layout.

National News Layout 1: 16.63% Layout 2: 18.54%

National News has the same but inverted affect as the main page.

International News Layout 1: 31.66% Layout 2: 26.27%

Here we see both layouts put a considerable emphasis on international news with it being the most popular would spend most of its time on the International News page.

Sport Layout 1: 22.74% Layout 2: 27.96%

Here we see that Layout 2 puts considerably more emphasis on sports than layout 1 does with almost a 5% increase.

Weather Layout 1: 11.33% Layout 2: 11.68%

Finally, the least visited page. long-term distribution suggests that the Markov chain would spend around 11.5% of its time in the "Weather" state regardless of layout.

Question 5.

From the analysis of our transition matrices, it is clear that international news is the most popular page from our click stream data. From this, I would suggest increasing the number of advertisements on this page due to it having the highest significant proportion on the stationary distribution's of both the current layout and layout 1 for the clickstreams. It is also incredibly close to the highest significant proportion for layout 2.

However, in the case of layout 2 being selected, perhaps Sports should be selected for the increase in advertising.