

# exploring 6 21

Renata Diaz

6/21/2018

## Paper LDA model and LDATS changepoint

Load Christensen 2018 data and source paper functions:

```
library(topicmodels)
source('previous-work/AIC_model_selection.R')

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union

source('previous-work/LDA-distance.R')

dat <- read.csv('paper_dat.csv', stringsAsFactors = F)

dates <- dat[,1]
dat <- dat[,2:22]
```

Run paper LDA model:

```
seeds = 2*seq(200)

# repeat LDA model fit and AIC calculation with a bunch of different seeds to test robustness of the an
best_ntopic = repeat_VEM(dat,
                        seeds,
                        topic_min=2,
                        topic_max=6)
```

Note that `repeat_VEM` returns, for each seed, the number of topics that produces the lowest AIC.

Histogram of best # of topics:

```
# histogram of how many seeds chose how many topics
hist(best_ntopic$k,breaks=c(0.5,1.5,2.5,3.5,4.5,5.5,6.5,7.5,8.5,9.5),xlab='best # of topics', main='')
```

Four topics is overwhelmingly the best (it is the best for the most seeds).

```
ntopic = 4
# =====
# 2b. how different is species composition of 4 community-types when LDA is run with different seeds?
# =====
# get the best 100 seeds where 4 topics was the best LDA model
seeds_4topics = best_ntopic %>%
```

```

filter(k == 4) %>%
arrange(aic) %>%
head(100) %>%
pull(SEED)

# best seed for 4 is 206
# choose seed with highest log likelihood for all following analyses
# (also produces plot of community composition for 'best' run compared to 'worst')
best_seed = calculate_LDA_distance(dat,seeds_4topics, k =4)
mean_dist = unlist(best_seed)[2]
max_dist = unlist(best_seed)[3]

best_seed
mean_dist
max_dist

```

Run the LDA model with the selected seed and number of topics.

```
ldamodel= LDA(dat,ntopics, control = list(seed = SEED),method='VEM')
```

Run the LDATS changepoint model with this LDA model.

Don't actually run this because it is time-consuming. Save and reload the changepoint model to look at it.

```

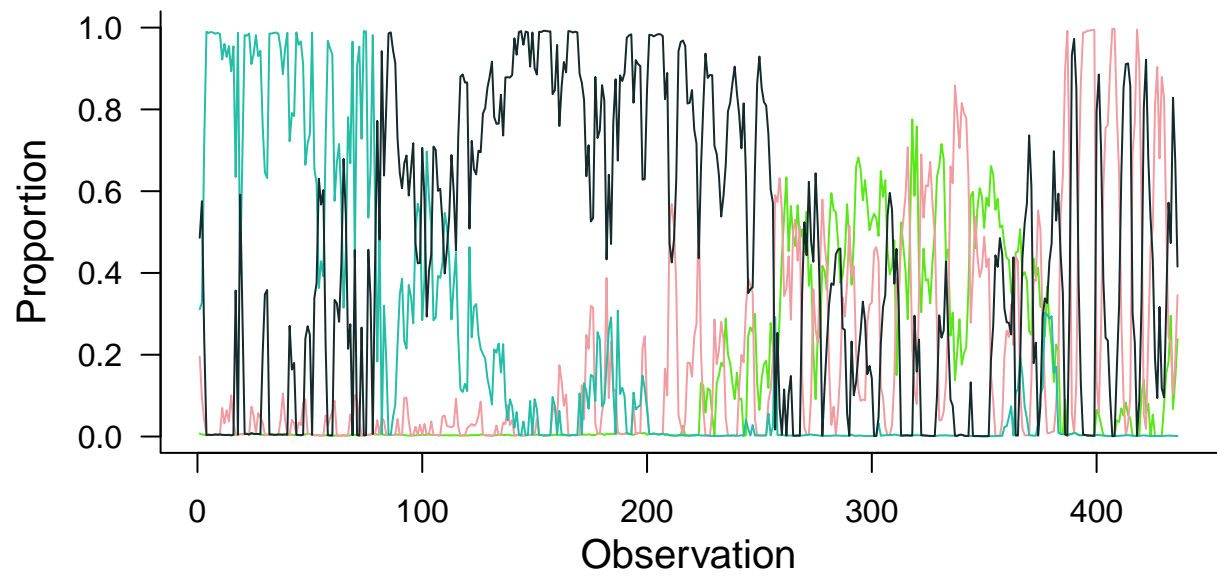
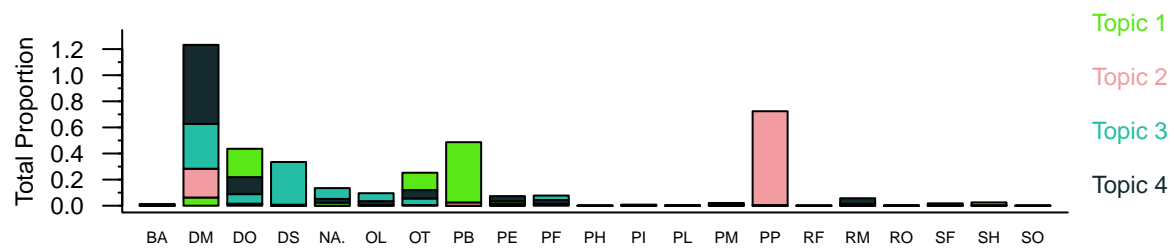
library(LDATS)
source('functions/run_changepoint_model.R')
dat$date <- as.Date(dates)

changepoint = run_rodent_cpt(rodent_data = dat, selected = ldamodel,
                           changepoints_vector = c(2, 3, 4, 5, 6))

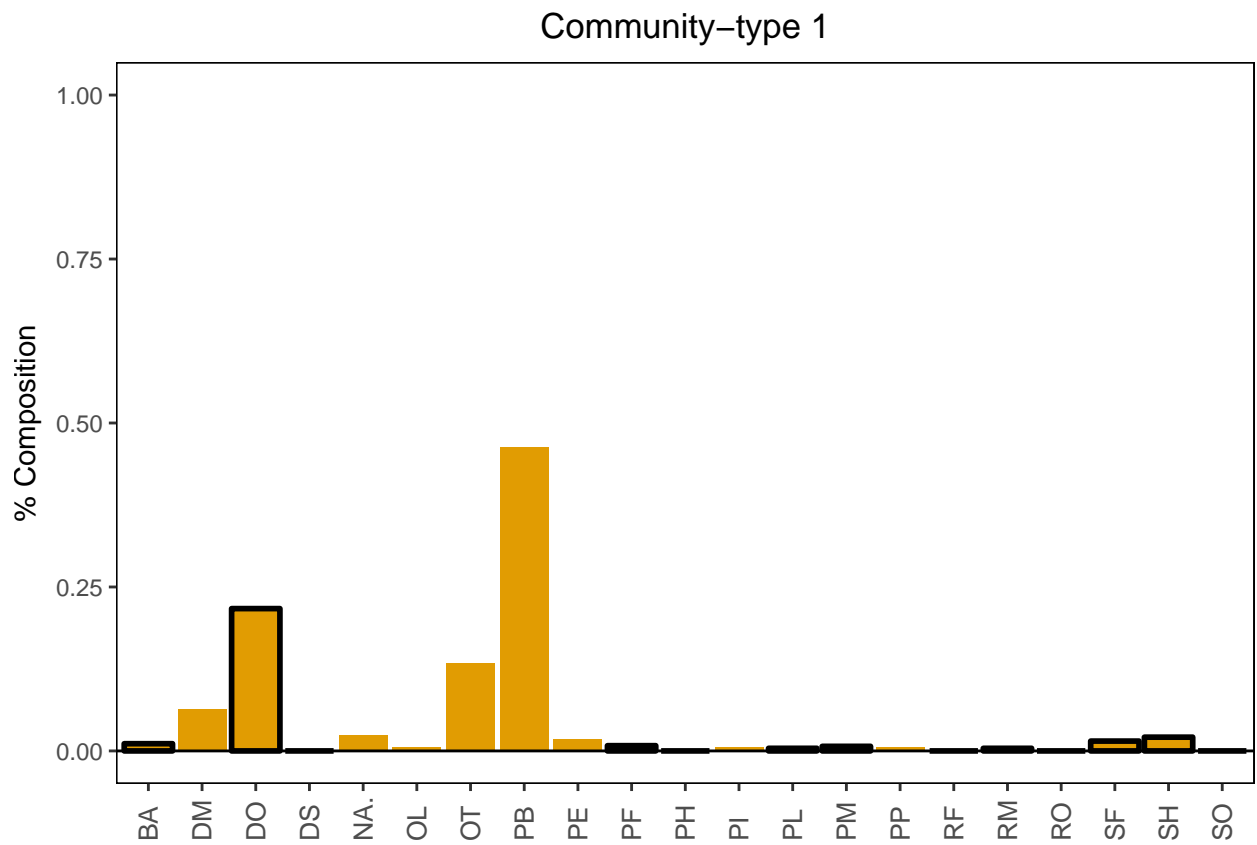
save(dat, ldamodel, changepoint, file = 'models/paperLDA_LDATScpt.Rdata')

```

Look at results.

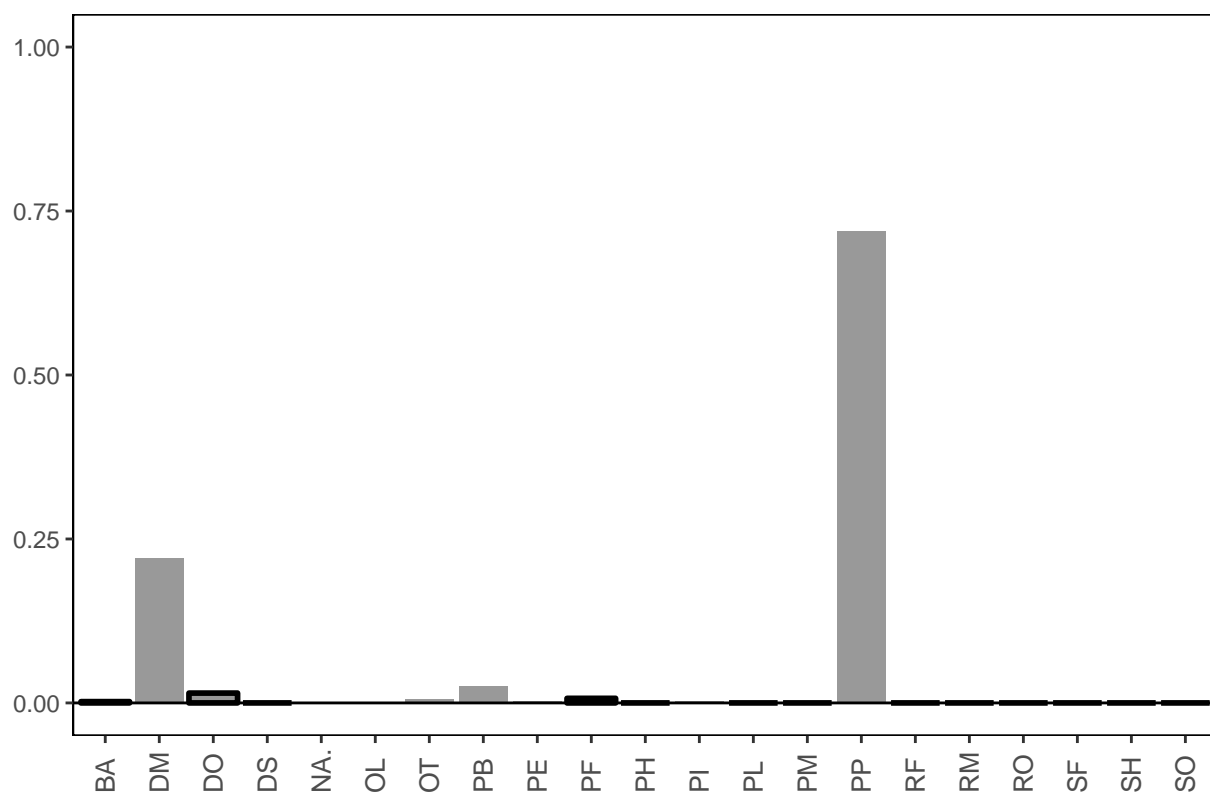


## [[1]]



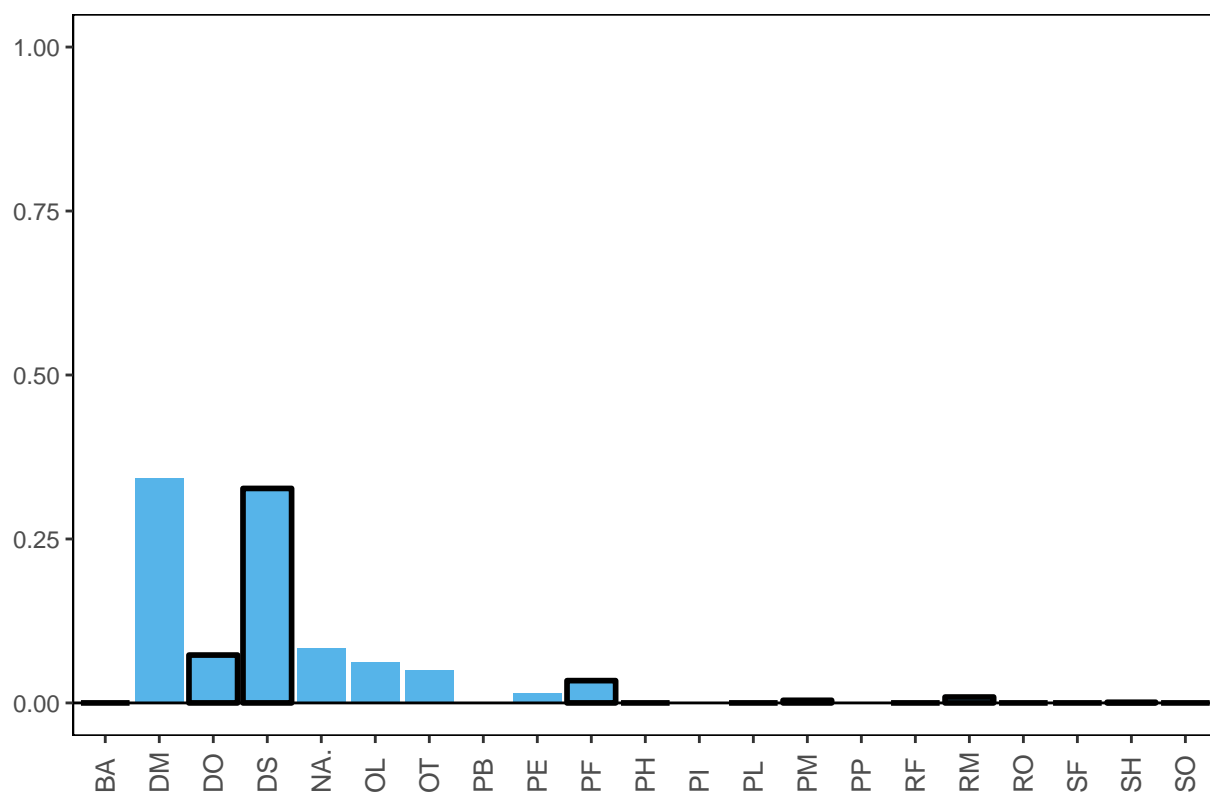
##  
## [[2]]

Community-type 2



##  
## [[3]]

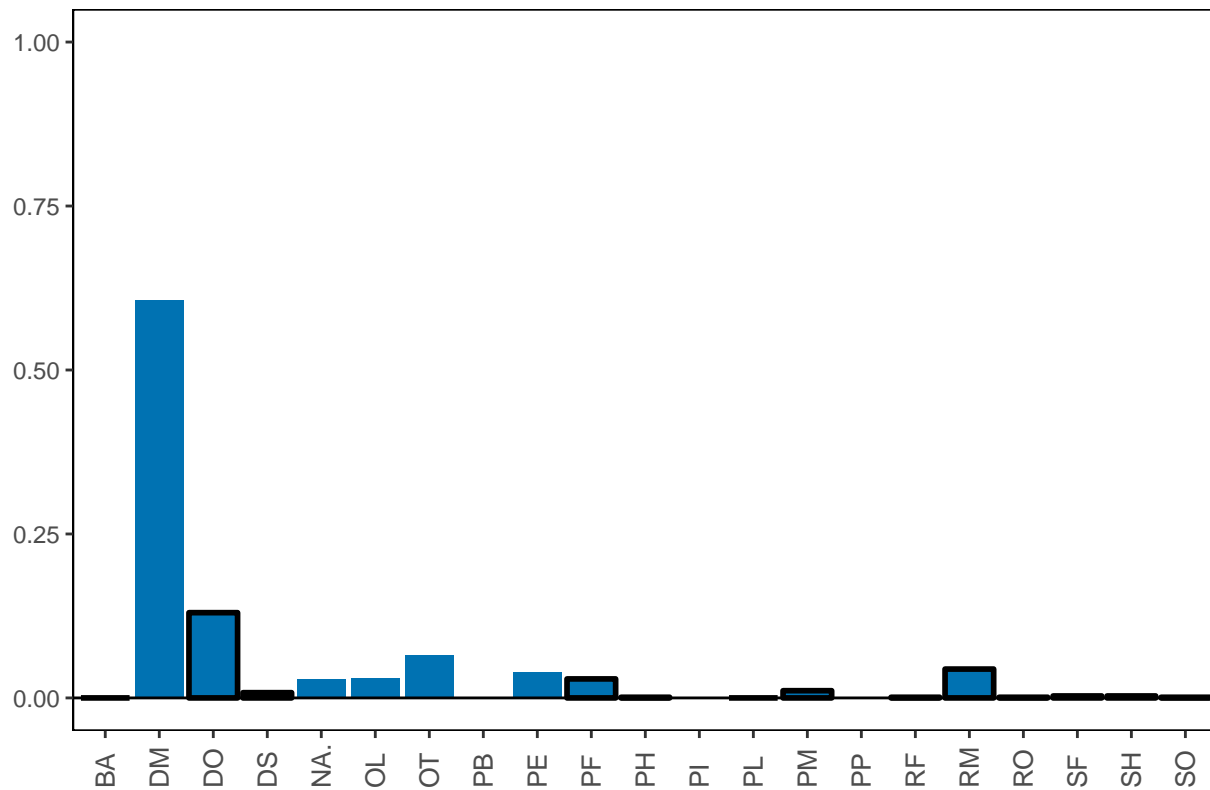
Community-type 3



##

## [[4]]

## Community-type 4



```
##           Mean Median Lower Upper      SD MCMCerr   AC10      ESS
## Changepoint_1 111.36   113    60   156 29.34  0.2949 0.2666 611.1389
## Changepoint_2 251.46   252   219   265 19.11  0.1921 0.0709 1398.7034

## $acceptance_rates
## [1] 0.6055556 0.8178788 0.9042424 0.9305051 0.9397980 0.9419192
##
## $swapping_rates
## [1] 0.3133333 0.5123232 0.7335354 0.8718182 0.9144444
##
## $strip_counts
## [1] 124 107 108 116 131 148
##
## $strip_rates
## [1] 0.01252525 0.01080808 0.01090909 0.01171717 0.01323232 0.01494949
```

## Looking at various chains in changepoint models

This will take more memory, because I need to fiddle around with the `select_changepoint_model` procedure.