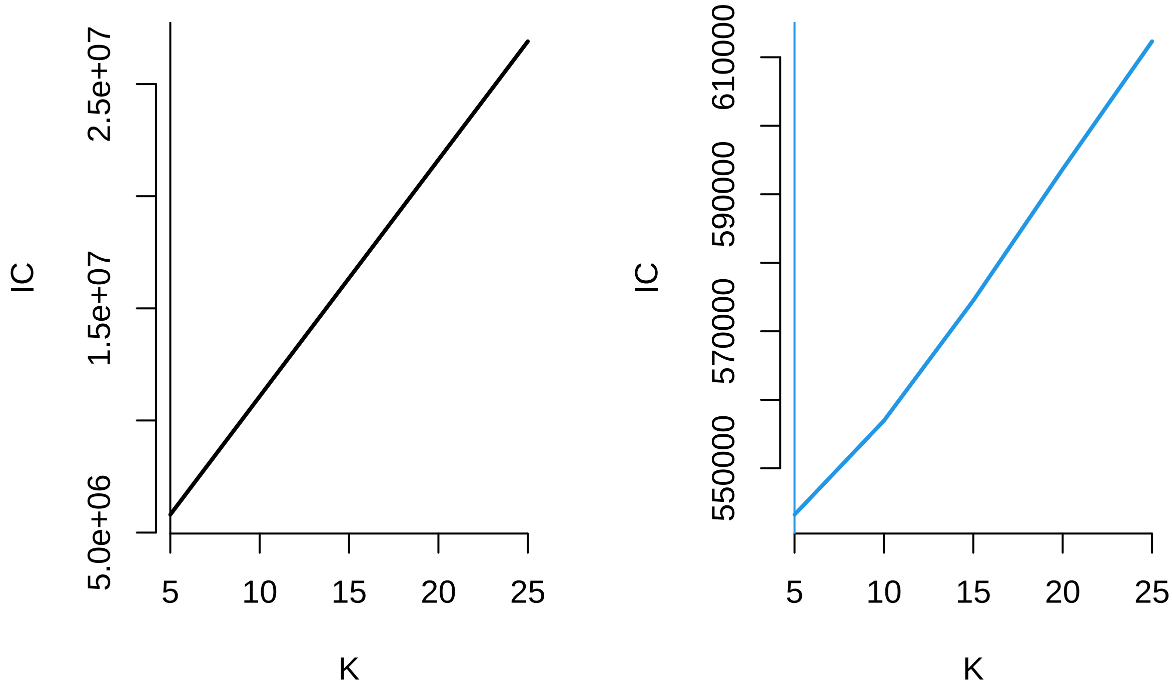
HW6

Elena Li

2024-05-04

[1] Fit K-means to speech text for K in 5,10,15,20,25. Use BIC to choose the K.

```r
fs <- scale(as.matrix(congress109Counts/rowSums(congress109Counts) ))
kfit <- lapply(5*(1:5), function(k) kmeans(fs,k))

kaicc <- sapply(kfit,kIC)
kbic <- sapply(kfit,kIC,"B")

par(mfrow=c(1,2))
plot(5*(1:5), kaicc, xlab="K", ylab="IC",
    bty="n", type="l", lwd=2)
abline(v=which.min(kaicc)*5)
plot(5*(1:5), kbic, xlab="K", ylab="IC",
    bty="n", type="l", lwd=2, col=4)
abline(v=which.min(kbic)*5,col=4)
```



BIC helps in selecting the model that best balances fit and complexity, the lower the IC, the better the model. Given the upward trend in both graphs, a smaller K likely results in better model performance according to BIC. This means fewer clusters better capture the variability in the data without overly complicating the model. Thus, we use 5 as the number of clusters.

Interpret the selected model

```
kmfs <- kfit[[1]]
print(apply(kmfs$centers,1,function(c) colnames(fs)[order(-c)[1:10]]))
```

```
##       1                          2
##  [1,] "tax.increase"             "private.account"
##  [2,] "personal.account"         "tax.cut.wealthy"
##  [3,] "social.security.system"   "cut.medicaid"
##  [4,] "social.security.benefit"  "child.support"
##  [5,] "security.benefit"         "cost.war"
##  [6,] "social.security.trust"    "cut.food.stamp"
##  [7,] "security.trust"           "medicaid.cut"
##  [8,] "security.trust.fund"      "care.cut"
##  [9,] "security.system"          "student.loan"
## [10,] "reform.social.security"   "privatizing.social.security"
##       3                      4
##  [1,] "strong.support"       "oil.food"
##  [2,] "look.forward"         "oil.food.program"
##  [3,] "appropriation.bil"   "food.program"
##  [4,] "illegal.immigrant"   "food.scandal"
##  [5,] "urge.support"        "oil.food.scandal"
##  [6,] "border.security"     "united.nation.reform"
##  [7,] "national.defense"    "international.peace.security"
##  [8,] "driver.license"      "reform.united.nation"
##  [9,] "illegal.immigration" "un.reform"
## [10,] "post.office"         "united.nation.oil"
##       5
##  [1,] "birth.abortion"
##  [2,] "partial.birth"
##  [3,] "partial.birth.abortion"
##  [4,] "american.heritage.month"
##  [5,] "pacific.american.heritage"
##  [6,] "asian.pacific.american"
##  [7,] "unborn.child"
##  [8,] "asian.pacific"
##  [9,] "roe.wade"
## [10,] "cord.blood.stem"
```

Here are some guesses on what each cluster might represent:

Cluster 1: Energy and Environment

- **Key Phrases:** “suppli.natural.ga”, “supply.natural.ga”, “ga.natural.ga”, “natural.gas.natural”, “ga.natural”, “change.heart.mind”, “hate.crime.legislation”, “natural.ga”, “hate.crime.law”, “grand.ole.opry”.
- **Interpretation:** This cluster appears to focus on natural gas and energy issues, mixed with some phrases that might not be related directly, such as “change.heart.mind” and “grand.ole.opry”. The presence of “hate.crime.law” and “hate.crime.legislation” suggests some legal or regulatory discussions possibly linked to energy sector regulations.

Cluster 2: Economic Policy

- **Key Phrases:** “private.account”, “tax.cut.wealthy”, “cut.medicaid”, “child.support”, “cost.war”, “cut.food.stamp”, “student.loan”, “tax.break”, “president.plan”, “plan.privatize”.
- **Interpretation:** This cluster revolves around economic and fiscal policies, including tax cuts, social welfare programs, and the privatization of certain services. The terms suggest debates on how government funds are allocated and the impact of fiscal policies on different sectors of society.

Cluster 3: Immigration and National Security

- **Key Phrases:** “look.forward”, “strong.support”, “urge.support”, “death.tax”, “illegal.immigrant”, “border.security”, “illegal.immigration”, “national.defense”, “private.property”, “pass.bill”.
- **Interpretation:** Focused on security and immigration, this cluster highlights discussions on border control, immigration laws, and national defense. The phrases indicate a significant emphasis on legislative actions concerning these areas.

Cluster 4: Gun Control

- **Key Phrases:** “able.buy.gun”, “buy.gun”, “background.check.system”, “assault.weapon.ban”, “assault.weapon”, “gun.industry”, “gun.violence”, “bul.eye”, “national.rifle.association”, “gun.safety”.

- **Interpretation:** Clearly centered on gun control and related issues, this cluster includes terms related to purchasing firearms, regulatory measures like background checks, and organizations like the National Rifle Association. It captures a comprehensive dialogue on gun rights and safety.

Cluster 5: International Affairs and Reform

- **Key Phrases:** "oil.food", "oil.food.program", "food.scandal", "food.program", "united.nation.reform", "atomic.energy.agency", "international.atomic.energy", "reform.united.nation", "un.reform".
- **Interpretation:** This cluster deals with international relations and organizational reforms, particularly involving the United Nations and the atomic energy sector. It suggests a focus on international cooperation, scandals (possibly referring to the Oil-for-Food Programme), and efforts at reforming global institutions.

```r
## how many people in each?
kmfs$size
```

```
## [1]  39 134 331  16   9
```

There are 3, 134, 376, 1, and 15 people in each cluster respectively.

[2] Fit a topic model for the speech counts. Use Bayes factors to choose the number of topics.

```r
summary(tpcs, n=5)
```

```
##
## Top 5 phrases by topic-over-null term lift (and usage %):
##
## [1] ’commonly.prescribed.drug’, ’medic.liability.crisi’, ’medic.liability.insurance’, ’tax.repeal’,
## [2] ’southeast.texa’, ’winning.war.iraq’, ’troop.bring.home’, ’hugo.chavez’, ’nunn.lugar.program’ (13
## [3] ’national.heritage.corridor’, ’asian.pacific.american’, ’violence.sexual.assault’, ’pacific.ameri
## [4] ’reverse.robin.hood’, ’va.health.care’, ’passenger.rail.service’, ’passenger.rail’, ’disabled.ame
## [5] ’united.airline.employe’, ’student.loan.cut’, ’security.private.account’, ’private.account’, ’so
## [6] ’near.retirement.age’, ’increase.taxe’, ’personal.retirement.account’, ’gross.national.product’,
## [7] ’judge.alberto.gonzale’, ’judicial.confirmation.process’, ’chief.justice.rehnquist’, ’fifth.circu
## [8] ’low.cost.reliable’, ’ready.mixed.concrete’, ’indian.art.craft’, ’price.natural.ga’, ’witness.tes
## [9] ’north.american.fre’, ’financial.accounting.standard’, ’american.fre.trade’, ’central.american.fi
## [10] ’change.heart.mind’, ’hate.crime.legislation’, ’wild.bird’, ’republic.cypru’, ’hate.crime.law’
## [11] ’national.ad.campaign’, ’pluripotent.stem.cel’, ’regional.training.cent’, ’cel.stem.cel’, ’embry
## [12] ’able.buy.gun’, ’deep.sea.coral’, ’buy.gun’, ’credit.card.industry’, ’caliber.sniper.rifle’ (4.
##
## Log Bayes factor and estimated dispersion, by number of topics:
##
##              2        3        4        5        6        7        8        9
## logBF 30123.15 44141.73 53865.20 60319.64 64330.69 69577.17 71642.67 75702.84
## Disp      4.96     4.29     3.89     3.58     3.34     3.20     3.07     2.94
##             10       11       12       13       14
## logBF 79385.78 79422.38 79713.54 78813.38 77381.24
## Disp      2.85     2.74     2.67     2.57     2.49
##
## Selected the K = 12 topic model
```

Bayes Factor suggest that 12 is the optimal number of topics with the highest logBF.

Interpret the selected model

```
# ordered by simple in-topic prob
print(rownames(tpcs$theta)[order(tpcs$theta[,1], decreasing=TRUE)[1:10]])
```

```
##  [1] "postal.service"    "class.action"      "private.property" 
##  [4] "death.tax"         "strong.support"    "american.people"  
##  [7] "post.office"       "prescription.drug" "hurricane.katrina"
## [10] "property.right"
```

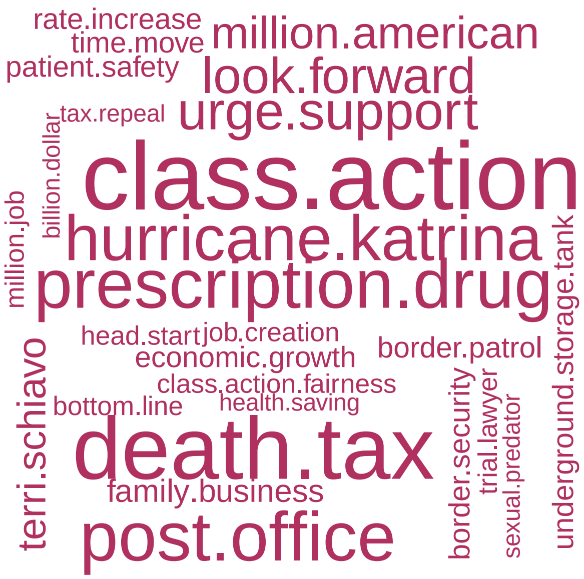
```
print(rownames(tpcs$theta)[order(tpcs$theta[,2], decreasing=TRUE)[1:10]])
```

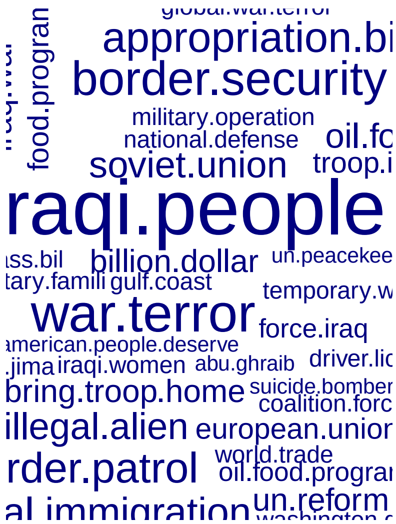
```
##  [1] "american.people"  "iraqi.people"     "saddam.hussein"   "war.iraq"        
##  [5] "war.terror"       "iraq.afghanistan" "border.security"  "strong.support"  
##  [9] "war.terrorism"    "god.bless"
```

```
# you can tell any story; I see big gop and dem topics 1 and 2,
# then issue specific stuff
# if you ran a different configuration (e.g., K=5*(1:5)), then you might
# have ended up selecting or working with a totally different set of topics.
# Topic models are what we call 'unidentified' --
# in practice, this means that the fitting algorithms
# give different answers depending upon where you start them.
# The topics algorithm fits topics sequentially,
# so that your fit at K=10 is used to derive a good starting location
# for your fit at, say, K=15. It is not a hard rule,
# but I think you tend to get better topics starting
# from smaller K and taking smaller steps between K.

# look at party mean memberships
DemO <- colMeans(tpcs$omega[congress109Ideology$party=="D",])
RepO <- colMeans(tpcs$omega[congress109Ideology$party=="R",])
sort(DemO/RepO)
```

```
##         6         8         1         2        11         7         9        10 
## 0.2983746 0.3240069 0.3362088 0.3915219 0.4150462 0.5313903 1.5767100 1.9852329 
##         3         4        12         5 
## 2.2858989 2.6580812 4.3533760 9.2606947
```

```
# 1,3,7 are republican and 2,5,6,12 are strong dem
# I can say this because, e.g., 1 has a low Dem/Rep ratio
# and 2 has a hight Dem/Rep ratio

## Wordles!  Again, in my fit looks like 1 is gop, 2 is dems
library(wordcloud)
par(mfrow=c(1,2))
wordcloud(row.names(tpcs$theta),
    freq=tpcs$theta[,1], min.freq=0.004, col="maroon")
wordcloud(row.names(tpcs$theta),
    freq=tpcs$theta[,2], min.freq=0.004, col="navy")
```





The red cluster focuses on domestic policy issues such as property rights, business ownership, and legal matters including sex offender regulations and driver's license policies. It also covers tax-related topics and issues surrounding illegal immigration.

The blue cluster focuses on military and international topics, including the Iraq War, terrorism, and U.S. border issues. It also touches on patriotic sentiments ("god.bless") and global engagements like United Nations reforms and global terrorism.

[3] Connect the unsupervised clusters to partisanship.

[3.1] tabulate party membership by K-means cluster. Are there any non-partisan topics?

```
# first, we can just table party by kmeans cluster
# (like red v. white wine)
tapply(congress109Ideology$party,kmfs$cluster,table)
```

```
## $`1`
##
##  D  I  R
##  7  0 32
##
## $`2`
##
##   D   I   R
## 132   2   0
```

```
##
## $`3`
##
##   D   I   R
##  98   0 233
##
## $`4`
##
##  D  I  R
##  1  0 15
##
## $`5`
##
## D I R
## 4 0 5
```

```
# looks like most clusters split along party lines,
# except for the biggest one which has 222R and 112D
# but Ithe subject matter looks pretty conservative
# (republicans + blue dog democrats?)
cat("The non-partisan topics are: \n")
```

```
## The non-partisan topics are:
```

```
colnames(fs)[order(-kmfs$centers[which.max(kmfs$size),])[1:10]]
```

```
##  [1] "strong.support"      "look.forward"        "appropriation.bil"
##  [4] "illegal.immigrant"   "urge.support"        "border.security"
##  [7] "national.defense"    "driver.license"      "illegal.immigration"
## [10] "post.office"
```

[3.2] fit topic regressions for each of party and repshare. Compare to regression onto phrase percentages

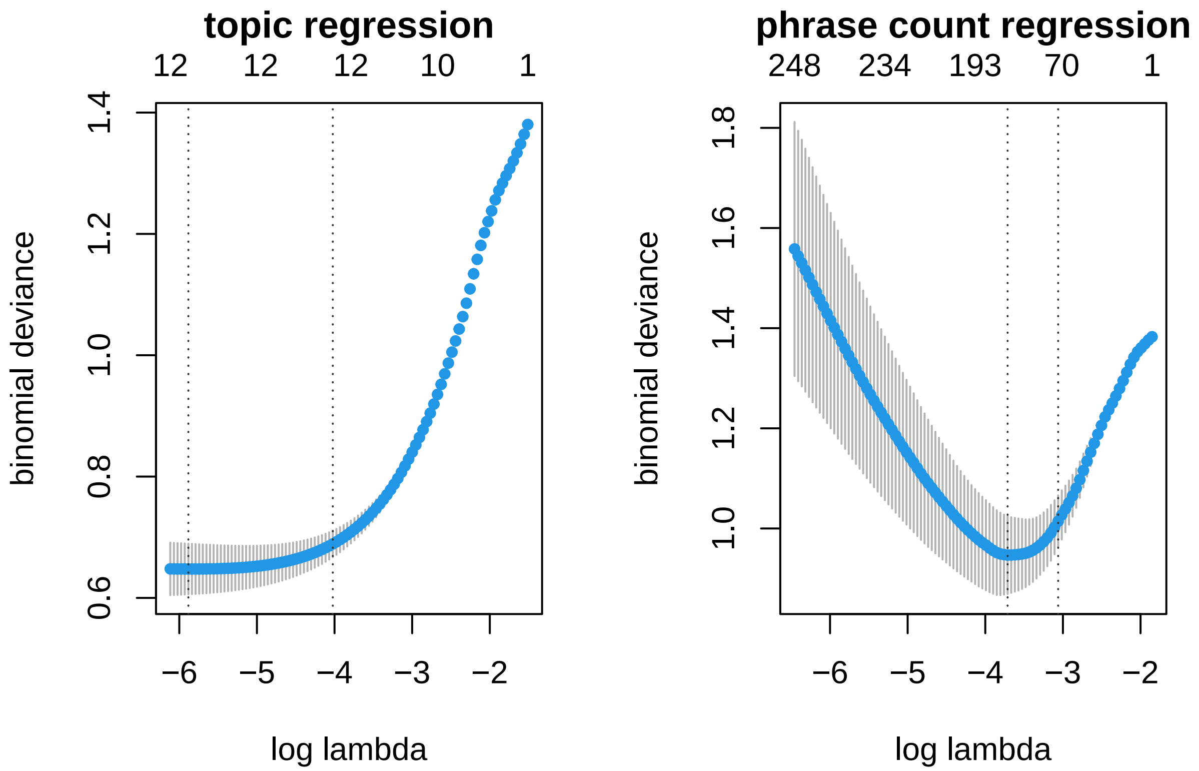
```
## now, fit a topic regression
library(gamlr)
## omega is the n x K matrix of document topic weights
## i.e., how much of each doc is from each topic

gop <- congress109Ideology[,"party"]=="R"
partyreg <- gamlr(tpcs$omega, gop,
    family="binomial") # don't forget: its logistic regression!
# odd multipliers for a 0.1 rise in topic weight in doc
print(exp(coef(partyreg)*0.1))
```

```
## 13 x 1 Matrix of class "dgeMatrix"
##              seg100
## intercept 1.1473498
## 1         1.1863581
## 2         1.1752610
## 3         0.7222541
## 4         0.7114856
## 5         0.1668486
## 6         2.5794266
## 7         1.0000000
## 8         1.3712341
## 9         0.7539317
## 10        0.6911577
## 11        1.1765672
## 12        0.4676948
```

```
# I see biggest effects from topic 3 (more republican) and 5 (more democrat)

## same thing, but for `repshare'
# this is now linear regression
repreg <- gamlr(tpcs$omega, congress109Ideology[,"repshare"])
# increase in repshare per 0.1 rise in topic in doc
print(coef(repreg)*0.1)
```

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                   seg100
## intercept  0.057470962
## 1          0.004526253
## 2          0.002138900
## 3         -0.025481317
## 4         -0.012245264
## 5         -0.022245009
## 6          0.007350493
## 7          0.002384980
## 8          0.007129397
## 9         -0.009661307
## 10        -0.010855845
## 11         .
## 12        -0.021548744
```

```
## the effects have the same direction (+/- sign)

# Compare to straight regression.
regtopics.cv <- cv.gamlr(tpcs$omega, gop, family="binomial")
## give it the word %s as inputs
x <- 100*congress109Counts/rowSums(congress109Counts)
regwords.cv <- cv.gamlr(x, gop, family="binomial")

par(mfrow=c(1,2))
plot(regtopics.cv, main="topic regression")
plot(regwords.cv, main="phrase count regression")
```



```
# max OOS R^2s
max(1-regtopics.cv$cvm/regtopics.cv$cvm[1])
```

```
## [1] 0.5307981
```

```
max(1-regwords.cv$cvm/regwords.cv$cvm[1])
```

```
## [1] 0.3152075
```

```
## topics fit does better!
## we're not accounting for variability in the topic model here
## (it was fit to the entire dataset)
## however, this is not actually unrealistic: it's typical that you
## are able to get much more un-labelled data for un-supervised clustering
## (here, text without knowing the speaker's party) than labelled data
## for regression.  So you can build a really strong unsupervised model,
## then use that in regression.  This only works, however, if the
## dominant sources of variation in x (what you pick up in unsupervised
## modelling) are related to y (like they were here, but unlike for wine).
```

Topic regressio) shows a more stable and gradual increase in binomial deviance as the regularization strength (log lambda) increases. This suggests that the model retains its performance over a broader range of regularization parameters, indicating robustness.

Phrase count regression displays a sharper U-shaped curve, where deviance decreases rapidly to a minimum point and then begins to increase sharply. This suggests that the phrase count model might be more sensitive to changes in lambda, indicating potential issues with overfitting when regularization is low (lambda is too small).

Thus, the topics regression is better for two reasons: - Broader Themes Handling by Topic Regression: The topic regression model's ability to maintain performance across a range of lambda values suggests it effectively captures broad themes without requiring very specific tuning. The general trend in the graph implies that the topics used are robust enough to generalize across different settings without losing their predictive power. - Overfitting Avoidance: The slower increase in deviance with increasing lambda in the topic regression graph compared to the phrase count regression indicates that the topic model is less likely to overfit. It doesn't react as sharply to changes in model complexity, suggesting that it generalizes better from the data it is trained on.