Measuring Ideological Position and Polarization from Texts

Machine Learning Application

Andrew Peterson

10.04.2018

Université de Genève

Overview

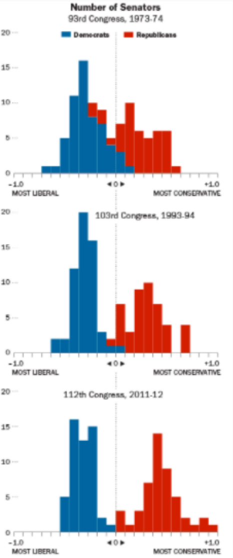
Table of contents

Overview

Code is available on Github at:
https://github.com/aristotle-tek

Classification Accuracy & Polarization

The Challenge



How to measure ideological polarization in the UK Parliament when votes are along party lines?

Preview of Results (3rd Party Evaluation: N. Goet)

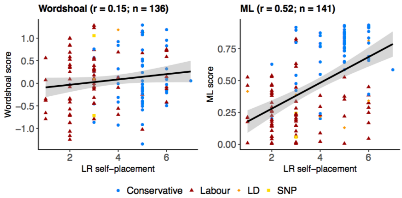


Figure 6: Correlation with British Candidate Survey Data

Table 3: Validation Scores

Validity	ML	Wordshoal
1. Face validity		
1.1 General test	✓	✓/✗
1.2 Detailed test	✓	✗
2. Convergent validity		
2.1 Session-level estimates	✓	✓
2.2 Individual estimates	✓	✗
3. Construct validity		
3.1 Between-session consistency	✓	✗
3.2 Individual-level distribution	✓	✗
3.3 Explanatory power of party	✓	✗

Challenge: Measuring Parliamentary Polarization

Roll calls: useful in US (e.g. NOMINATE) but uninformative in (very) cohesive legislatures with strategic voting (Spirling & McLean, 2007)

Early Day Motions: work well for those that sign in modern period (Kellerman, 2012) but selection problems and data unavailable for most of 20th C.

Speeches: can place parties (e.g. Slapin & Proksch, 2008), but what about locating individual MPs on something other than 'gov-vs-oppn' scale (e.g. Lauderdale & Herzog, 2016)

Alternative Approaches to Measuring Speech Polarization

Ideological slant - difference in word occurrence frequencies (Gentzkow and Shapiro, 2010)

3.1. *Selecting Phrases for Analysis*

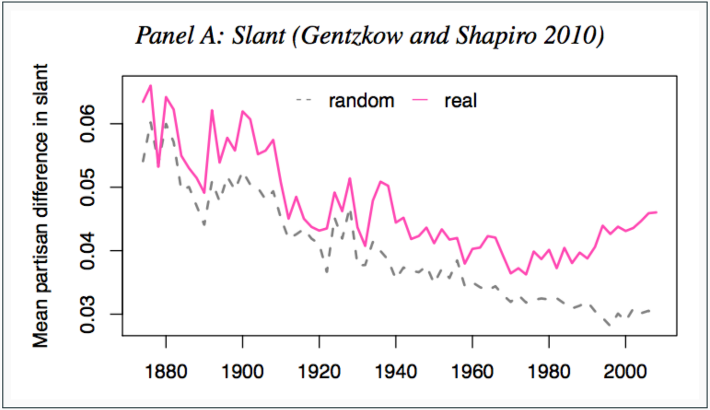
Let f_{pld} and f_{plr} denote the total number of times phrase p of length l (two or three words) is used by Democrats and Republicans, respectively. Let $f_{\sim pld}$ and $f_{\sim plr}$ denote the total occurrences of length-l phrases that are *not* phrase p spoken by Democrats and Republicans, respectively. Let χ^2_{pl} denote Pearson's χ^2 statistic for each phrase:

$$(1) \qquad \chi^2_{pl} = \frac{(f_{plr} f_{\sim pld} - f_{pld} f_{\sim plr})^2}{(f_{plr} + f_{pld})(f_{plr} + f_{\sim plr})(f_{pld} + f_{\sim pld})(f_{\sim plr} + f_{\sim pld})}.$$

Perfectly simple and reasonable. Any objections?

Alternative Approaches to Measuring Speech Polarization

Gentzkow, Shapiro & Taddy* Frequency approaches can be biased by changes in the size of the vocabulary. Null test: estimate on randomized labels.



* "Measuring Polarization in High-dimensional Data: Method and Application to Congressional Speech."

Alternative Approaches

- "highly parametric" approach: Gentzkow, et al. Penalized maximum likelihood estimation via Poisson approximation
- scaling, CA, Wordfish, wordshoals, etc. (unsupervised)
- ML

Alternative Approaches

- Niels Goet – evaluating alternative approaches

Table 3: Validation Scores

Validity	ML	Wordshoal
1. Face validity		
1.1 General test	✓	✓/✗
1.2 Detailed test	✓	✗
2. Convergent validity		
2.1 Session-level estimates	✓	✓
2.2 Individual estimates	✓	✗
3. Construct validity		
3.1 Between-session consistency	✓	✗
3.2 Individual-level distribution	✓	✗
3.3 Explanatory power of party	✓	✗

Alternative Approaches

- Niels Goet – evaluating alternative approaches

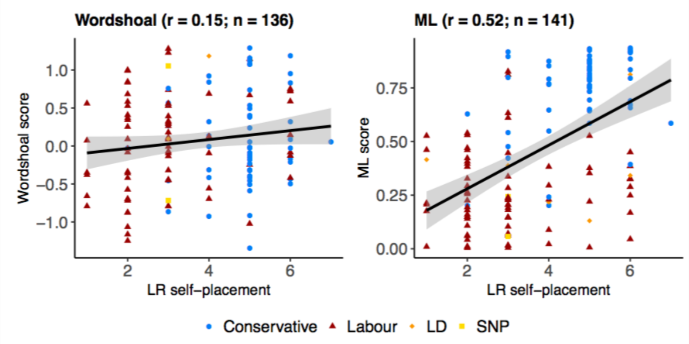


Figure 6: Correlation with British Candidate Survey Data

Approach: Polarization as Distinctness of Partisan Speech

Data: 3.5 million Commons speeches, 1935–2013 (Beelen, et al 2016). Restrict to Labour and Conservative members (approx 85% of all members)

Assumption: in more polarized periods, speeches from parties will be 'more different', more distinct.

Attempt to predict party of MP delivering speech, using only the speech document term vector.

Idea: when we can predict party well (say, 90% correct) we have highly polarized chamber; when prediction accuracy is poor (say, 50% correct) we have period of low polarization.

Intuition 1: Attlee vs Eden, 1948—low polarization

Deputy Leader of Oppn: May I ask the Prime Minister if he has been made aware that there is considerable concern that the Minister of Labour should have left the country ...and whether it would not have been possible to retain the Minister of Labour here until these difficult negotiations were completed?

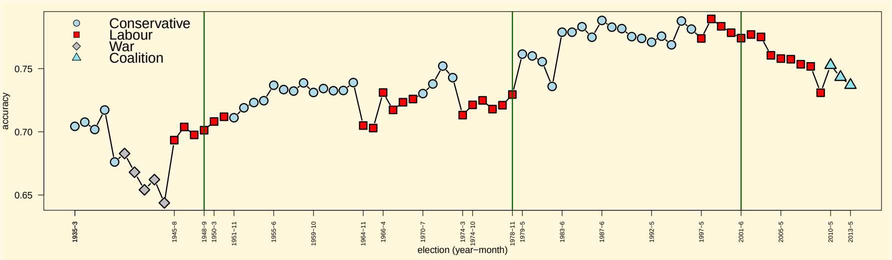
Prime Minister: At the time my right hon. Friend the Minister of Labour left for the important Conference of the I.L.O. it was thought that the matter had been settled...I can assure the right hon. Gentleman that everything will be done...

Intuition 2: Kinnock vs Thatcher—high polarization

Prime Minister: I am happy that my successor will carry on the excellent policies that have finished with the decline of socialism, brought great prosperity to this country, raised Britain's standing in the world and brought about a truly capital-owning democracy.

Leader of Oppn: If the Prime Minister thinks that nothing should be changed, can she tell us why on earth all those now competing for her job are desperately wriggling around trying to find a way out of the poll tax trap?

Aggregate Polarization of Commons



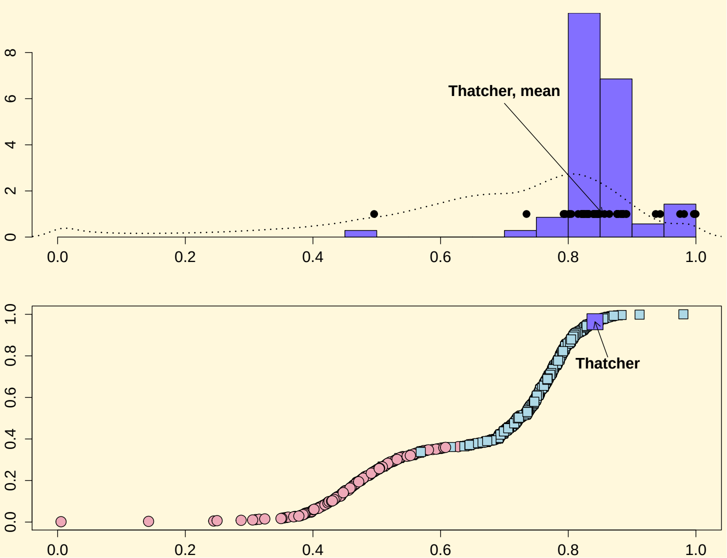
- Change points after WWII; Thatcher I (1979); Blair II (2001)
- Modern polarization where it was $\sim$ 1965

Member Estimates

Can predict Pr(speech is Conservative) by predicting the probability that each of their speeches is Conservative.

$\rightarrow$ Conceive of MP position as simply being mean of their speeches (in L–C, 0–1 space) in given session.

Example: Thatcher, 1984



Implementation Details

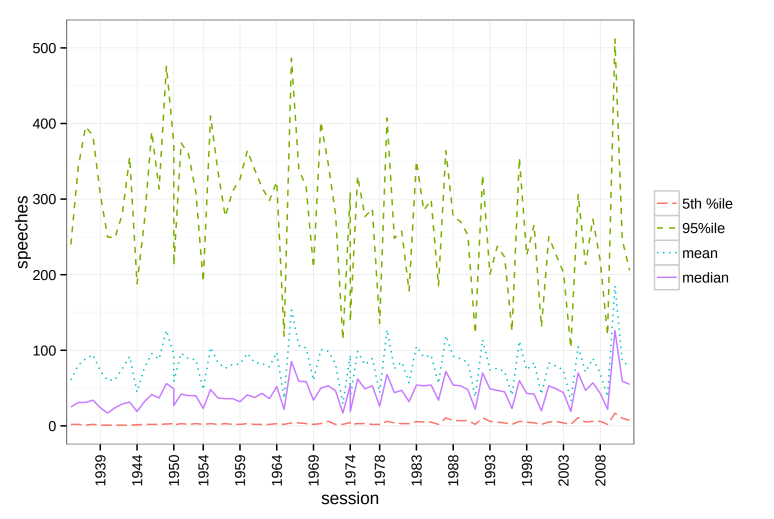
XML to dataframe

- data obtained in `xml` format:

```xml
<speech pm:speaker="The PRIME MINISTER" pm:party="Conservative" pm:role="gov" pm:
member-ref="uk.m.20736" pm:id="uk.proc.d.1935-12-03.3.10.12" pm:function="The PRIME
MINISTER" pm:party-ref="uk.p.Con">
  <p pm:id="uk.proc.d.1935-12-03.3.10.12.1">I do not think that is quite true, but I
  will come to that point by and by. In this country there is to-day a much greater area
  of confidence, there is more work and the purchasing power of the country has
  improved. Unemployment generally has continued to improve, even at a time of the year
  when a large seasonal decline is generally in evidence. So far as they go those signs
  are good; they are based on confidence and will only continue as long as that
  confidence can be maintained. Apart from any unforeseen disaster that may occur
  abroad, I am confident that those conditions will remain as long as we can all pull
  together as a nation. It is true that the rate of improvement has slowed down a little
  this year, but that is not unnatural. If you are recovering from the kind of economic
  disaster that lasted from 1929 to 1931, or into 1932, recovery must be slow, and if it
```

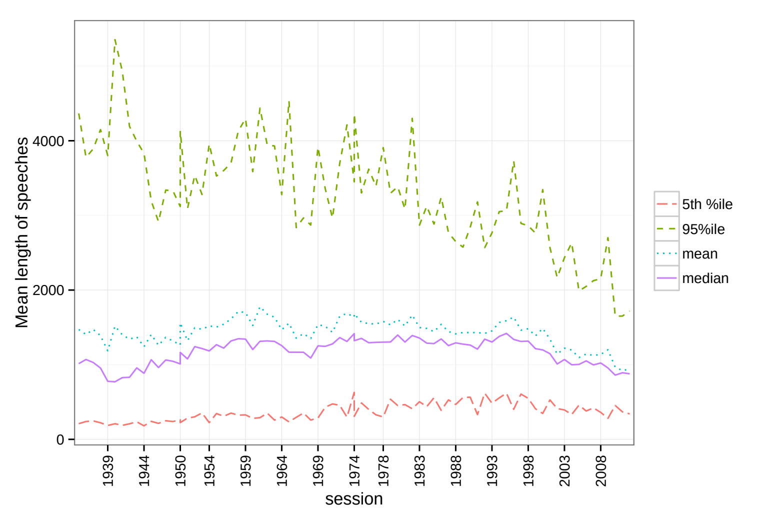
- parsed relevant parts (speaker, party, date, text...) using Python, lxml
- (NB: errors in govt role...)

Temporal Stability of the Data

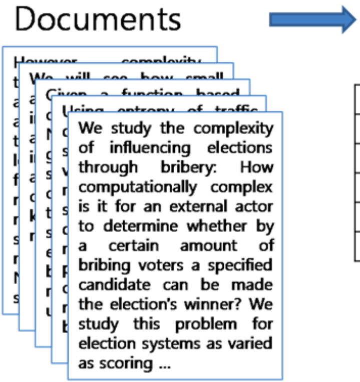


Number of Speeches By Member Per Session

Mean Length of Speeches By Member Per Session



Bag-of-Words (Document term matrix)



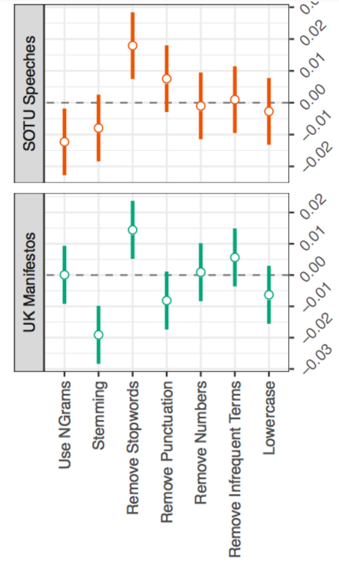
Vector-space representation

	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Term-document matrix

Pre-processing

- Denny & Spirling (2017)



Pre-processing

- (Python `Scikit-Learn`)
- Impose min speech length
- Counts (versus tf-idf)
- no stemming, stopword removal; require terms appear in 200 speeches
- max absolute scale (preserves sparsity)
- augment with topic dummies
- save as sparse matrices (`scipy.sparse.csr_matrix`)

Pre-processing

```
vectorizer = CountVectorizer(decode_error='ignore',
    vocabulary=presavedvocab, min_df =200)
X = vectorizer.transform(text)
X, topics = augment_with_topics(X, df)
X = sklearn.preprocessing.maxabs_scale(X)
scipy.io.mmwrite(mat_dir + 'foo.mtx' , X)


# now v0.19: sklearn.preprocessing.MaxAbsScaler
# if not sparse: X = preprocessing.scale(X)
```

10-fold cross-validation

Dealing with Imbalanced data

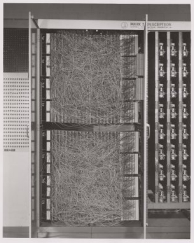
- reweight loss function for party strength:
- weights inversely proportional to the class (party) frequencies: $\frac{n}{2 \cdot n_p}$, where n is the total number of speeches and n_p is the number of speeches by members of that party.
- (weight up the speeches of the less commonly observed party in a given session)

Four Classifiers

- logistic regression, SGD
- logistic regression, SAG
- Perceptron classifier
- the 'passive aggressive' classifier

Perceptron

$$f(x) = \begin{cases} 1, & \text{if } w \cdot x + b > 0 \\ 0, & \text{otherwise} \end{cases}$$



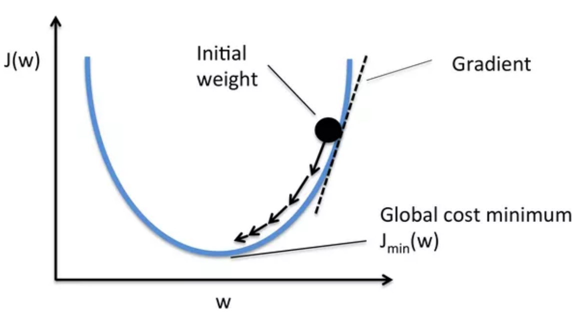
(1957) Cornell University Library

Stochastic Gradient Descent

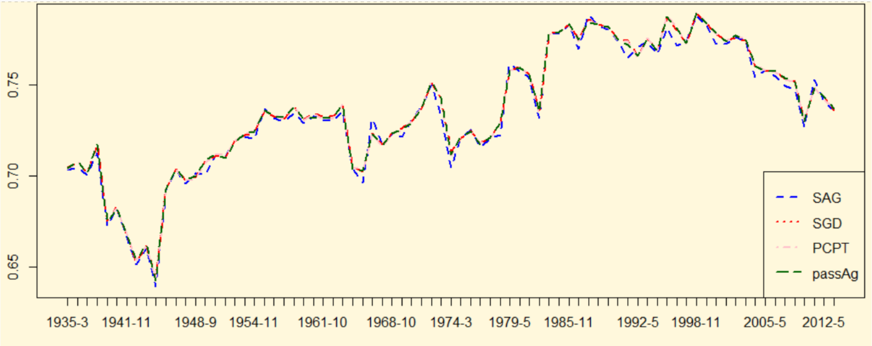
- Logistic regression

$$y = \frac{1}{1 + exp(-\beta_0 + \beta x)}$$

- Update parameters on batches of randomly selected subsets of the data (Bottou 2004)



Rough Equivalence



Validation & Robustness

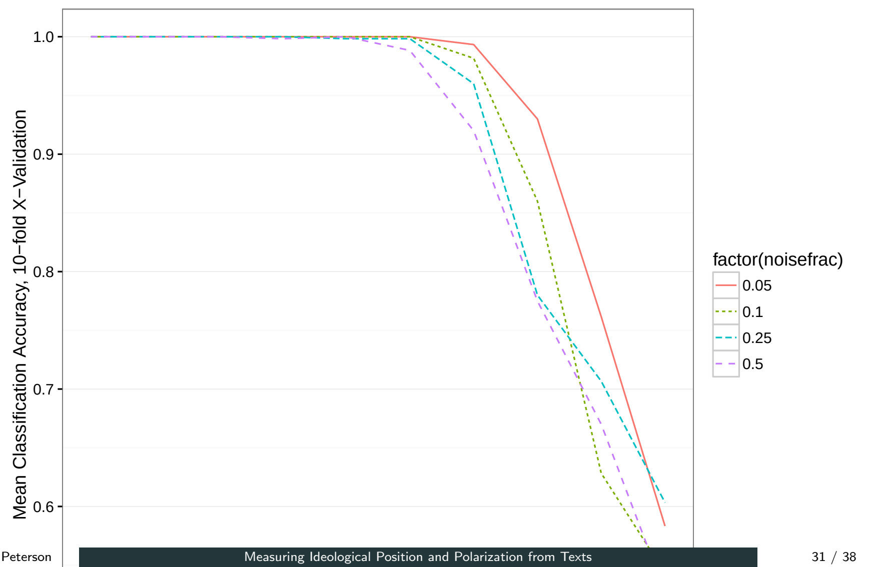
Possible Objection: Bias and Noise

noise and uncertainty confused for centrality?

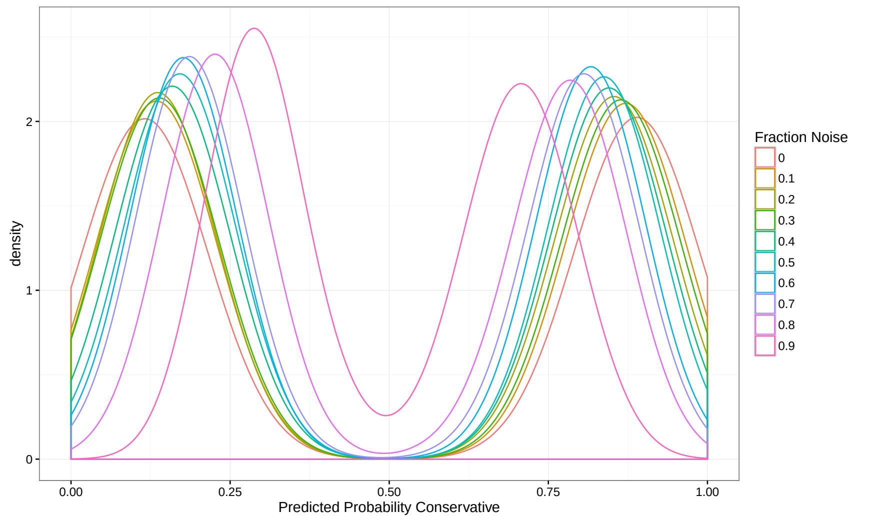
simulation of Parliaments with noise

Robustness to Noise

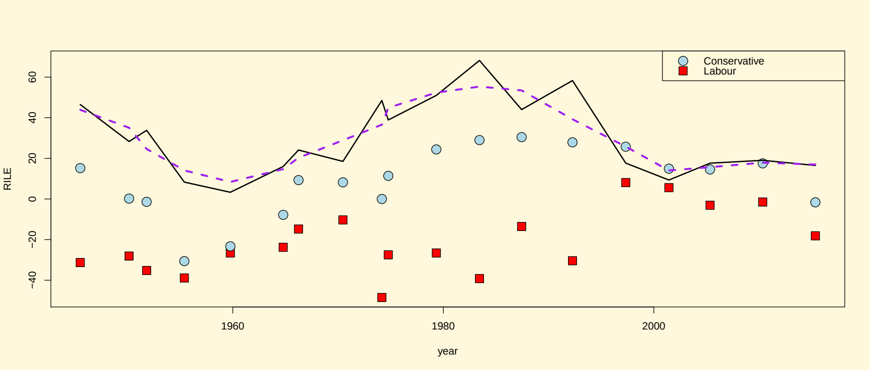
Classification Accuracy (y-axis) for Different Levels of Separation (x-axis) at different levels of noise.



Density of Predicted Probability Conservative by Noise



vs Left/right (RILE) scores from the Manifesto Project



Higher scores correspond to more right wing policies. Lines are difference between the parties (solid) and lowess (broken) of the same.

Reviewer objection

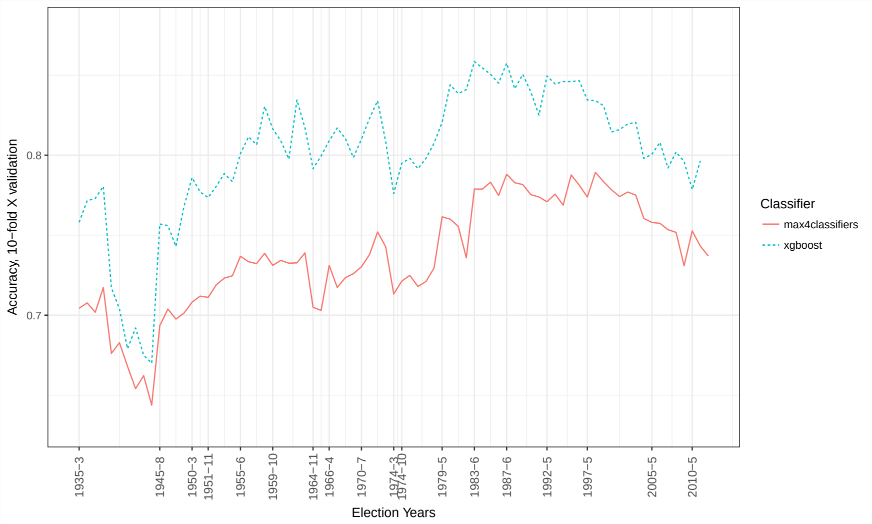
"If you had a better classifier, you'd get different results"

– Oouu! I do have a better classifier...

XGBoost

- boosted tree model: add new regression trees to optimize objective function given the residuals from the current model, with regularization
- exact greedy algorithm: first sorts the features according to their importance and then identifies the optimal point at which to make a split for each of these features.
- max depth 14, 400 estimators, logistic regression for binary classification as the objective, and learning rate of 0.1.

Accuracy vs XGBoost ($\rho = .89$)



With percentile bootstrap confidence interval

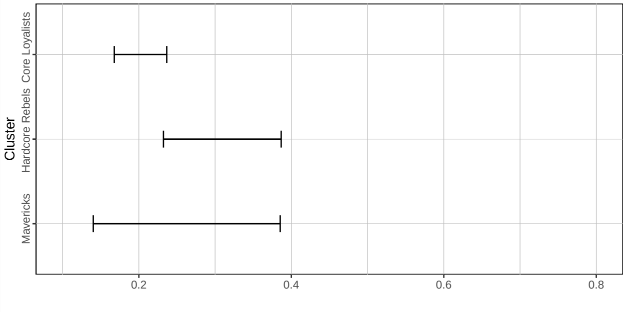
(resampling texts within each session)

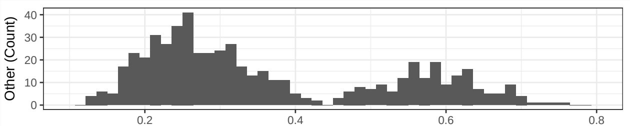
Questions?

Additional slides

Comparing our individual level scores to various clusters of MPs

(‘Core Loyalists’, ‘Hardcore Rebels’, ‘Mavericks’) from Spirling & Quinn 2010. Histogram is of all MPs in 1998.





Mean Variance by Session