Text as Data

Justin Grimmer

Associate Professor
Department of Political Science
Stanford University

October 7th, 2014
Estimating Word Discrimination

1) Task
   a) Classification \(\rightarrow\) learn word weights for dictionaries
   b) Fictitious prediction problem \(\rightarrow\) Identify features that discriminate between groups to learn features that are indicative of some group

2) Objective function

\[ f(\theta, X) = f(\theta, X, Y) \]

where:
\( Y = \) Document Labels
\( X = \) Document Features
\( \theta = \) Parameters that measure words discrimination between categories

3) Optimization \(\rightarrow\) method specific

4) Validation \(\rightarrow\) depends on task
   i) Classification \(\rightarrow\) Accuracy, Precision, Recall
   ii) Fictitious prediction \(\rightarrow\) Face, convergent, discriminatory, and confound
Stylometry → Who Wrote Disputed Federalist Papers?

Federalist papers → Mosteller and Wallace (1963)
- Persuade citizens of New York State to adopt constitution
- Canonical texts in study of American politics
- 77 essays
  - Published from 1787-1788 in Newspapers
  - And under the name Publius, anonymously

Who Wrote the Federalist papers?
- Jay wrote essays 2, 3, 4, 5, and 64
- Hamilton: wrote 43 papers
- Madison: wrote 12 papers

Disputed: Hamilton or Madison?
- Essays: 49-58, 62, and 63
- Joint Essays: 18-20

Task: identify authors of the disputed papers.
Task: Classify papers as Hamilton or Madison using dictionary methods
Setting up the Analysis

**Training** ⇝ papers Hamilton, Madison are known to have authored

**Test** ⇝ unlabeled papers

**Preprocessing:**
- Hamilton/Madison both discuss similar issues
- Differ in extent they use stop words
- Focus analysis on the stop words
Setting up the Analysis

- $Y = (Y_1, Y_2, \ldots, Y_N) = (\text{Hamilton, Hamilton, Madison, \ldots, Hamilton})$
  
  $N \times 1$ matrix with author labels

- Define the number of words in federalist paper $i$ as $\text{num}_i$
  
  $X = \begin{pmatrix}
  \frac{1}{\text{num}_1} & \frac{2}{\text{num}_1} & \frac{0}{\text{num}_1} & \cdots & \frac{3}{\text{num}_1} \\
  \frac{0}{\text{num}_2} & \frac{1}{\text{num}_2} & \frac{0}{\text{num}_2} & \cdots & \frac{0}{\text{num}_2} \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  \frac{0}{\text{num}_N} & \frac{0}{\text{num}_N} & \frac{1}{\text{num}_N} & \cdots & \frac{0}{\text{num}_N}
  \end{pmatrix}
  
  $N \times J$ counting stop word usage rate

- $\theta = (\theta_1, \theta_2, \ldots, \theta_J)$
  
  Word weights.
Objective Function

Heuristically: find $\theta^* = (\theta_1^*, \theta_2^*, \ldots, \theta_J^*)$ used to create score

$$p_i = \sum_{j=1}^{J} \theta_j^* X_{ij}$$

that maximally discriminates between categories
Define:

\[ \mu_{\text{Madison}} = \frac{1}{N_{\text{Madison}}} \sum_{i=1}^{N} I(Y_i = \text{Madison}) X_i \]

\[ \mu_{\text{Hamilton}} = \frac{1}{N_{\text{Hamilton}}} \sum_{i=1}^{N} I(Y_i = \text{Hamilton}) X_i \]
Objective Function

We can then define functions that describe the “projected” mean and variance for each author

\[
g(\theta, X, Y, \text{Madison}) = \frac{1}{N_{\text{Madison}}} \sum_{i=1}^{N} I(Y_i = \text{Madison}) \theta' X_i = \theta' \mu_{\text{Madison}}
\]

\[
g(\theta, X, Y, \text{Hamilton}) = \frac{1}{N_{\text{Hamilton}}} \sum_{i=1}^{N} I(Y_i = \text{Hamilton}) \theta' X_i = \theta' \mu_{\text{Hamilton}}
\]

\[
s(\theta, X, Y, \text{Madison}) = \sum_{i=1}^{N} I(Y_i = \text{Madison})(\theta' X_i - \theta' \mu_{\text{Madison}})^2
\]

\[
s(\theta, X, Y, \text{Hamilton}) = \sum_{i=1}^{N} I(Y_i = \text{Hamilton})(\theta' X_i - \theta' \mu_{\text{Hamilton}})^2
\]
Objective Function \implies Optimization

\[ f(\theta, X, Y) = \frac{(g(\theta, X, Y, \text{Hamilton}) - g(\theta, X, Y, \text{Madison}))^2}{s(\theta, X, Y, \text{Hamilton}) + s(\theta, X, Y, \text{Madison})} \]

\[ = \frac{\left(\theta' (\mu_{\text{Hamilton}} - \mu_{\text{Madison}})\right)^2}{\text{Scatter}_{\text{Hamilton}} + \text{Scatter}_{\text{Madison}}} \]

Optimization \implies \text{find } \theta^* \text{ to maximize } f(\theta, X, Y), \text{ assuming independence across dimensions.}

(Fisher’s) Linear Discriminant Analysis
Optimization \rightarrow \text{Word Weights}

For each word $j$, construct weight $\theta^*_j$,

$$
\begin{align*}
\mu_{j, \text{Hamilton}} &= \frac{\sum_{i=1}^N I(Y_i = \text{Hamilton})X_{ij}}{\sum_{j=1}^J \sum_{i=1}^N I(Y_i = \text{Hamilton})X_{ij}} \\
\mu_{j, \text{Madison}} &= \frac{\sum_{i=1}^N I(Y_i = \text{Madison})X_{ij}}{\sum_{j=1}^J \sum_{i=1}^N I(Y_i = \text{Madison})X_{ij}} \\
\sigma^2_{j, \text{Hamilton}} &= \text{Var}(X_{i,j} | \text{Hamilton}) \\
\sigma^2_{j, \text{Madison}} &= \text{Var}(X_{i,j} | \text{Madison})
\end{align*}
$$

We can then generate weight $\theta^*_j$ as

$$
\theta^*_j = \frac{\mu_{j, \text{Hamilton}} - \mu_{j, \text{Madison}}}{\sigma^2_{j, \text{Hamilton}} + \sigma^2_{j, \text{Madison}}}
$$
Trimming the Dictionary

- Trimming weights: Focus on discriminating words (very simple regularization)
- Cut off: For all $|\theta_j^*| < 0.025$ set $\theta_j^* = 0$. 
For each disputed document $i$, compute discrimination statistic

$$p_i = \sum_{j=1}^{J} \theta_j^* X_{ij}$$

$p_i \rightarrow$ classification (linear discriminator)
- Above midpoint in training set $\rightarrow$ Hamilton text
- Below midpoint in training set $\rightarrow$ Madison text

Findings: Madison is the author of the disputed Federalist papers.
Inferring Separating Words
Classification → Custom Dictionaries

Fictitious Prediction Problem
Infer words that are indicative of some class/group
- Difference in Republican, Democratic language
  → Partisan words
- Difference in Liberal, Conservative language
  → Ideological Language
- Difference in Secret/Not Secret Language
  → Secretive Language (Gill and Spirling 2014)
- Difference in Toy advertising
- Difference in Language across groups
  → Labeling output from Clustering/Topic Models

Vague and Difficult to derive beforehand
Inferring Separating Words

Classification $\Rightarrow$ Custom Dictionaries
- Stylometry $\Rightarrow$ Classify Authors

Fictitious Prediction Problem
- Infer words that are indicative of some class/group
- Difference in Republican, Democratic language $\Rightarrow$ Partisan words
- Difference in Liberal, Conservative language $\Rightarrow$ Ideological Language
- Difference in Secret/Not Secret Language $\Rightarrow$ Secretive Language (Gill and Spirling 2014)
- Difference in Toy advertising
- Difference in Language across groups

Labeling output from Clustering/Topic Models

Vague and Difficult to derive beforehand
Inferring Separating Words

Classification ⇝ Custom Dictionaries
- Stylometry ⇝ Classify Authors
- Dictionary based classification ⇝ Gentzkow and Shapiro (2010) and measures of media slant

- Fictitious Prediction Problem
  - Infer words that are indicative of some class/group
    - Difference in Republican, Democratic language ⇝ Partisan words
    - Difference in Liberal, Conservative language ⇝ Ideological Language
    - Difference in Secret/Not Secret Language ⇝ Secretive Language (Gill and Spirling 2014)
    - Difference in Toy advertising
- Difference in Language across groups ⇝ Labeling output from Clustering/Topic Models
  - Vague and Difficult to derive beforehand
Inferring Separating Words
Classification $\mapsto$ Custom Dictionaries
- Stylometry $\mapsto$ Classify Authors
- Dictionary based classification $\mapsto$ Gentzkow and Shapiro (2010) and measures of media slant
- Dictionary based classification $\mapsto$ Customized to particular setting
Inferring Separating Words
Classification \( \rightsquigarrow \) Custom Dictionaries
- Stylometry \( \rightsquigarrow \) Classify Authors
- Dictionary based classification \( \rightsquigarrow \) Gentzkow and Shapiro (2010) and measures of media slant
- Dictionary based classification \( \rightsquigarrow \) Customized to particular setting

Fictitious Prediction Problem \( \rightsquigarrow \) Infer words that are indicative of some class/group
Inferring Separating Words
Classification ⇝ Custom Dictionaries
- Stylometry ⇝ Classify Authors
- Dictionary based classification ⇝ Gentzkow and Shapiro (2010) and measures of media slant
- Dictionary based classification ⇝ Customized to particular setting

Fictitious Prediction Problem ⇝ Infer words that are indicative of some class/group
- Difference in Republican, Democratic language ⇝ Partisan words
Inferring Separating Words

Classification⇝ Custom Dictionaries
- Stylometry⇝ Classify Authors
- Dictionary based classification ⇝ Gentzkow and Shapiro (2010) and measures of media slant
- Dictionary based classification⇝ Customized to particular setting

Fictitious Prediction Problem ⇝ Infer words that are indicative of some class/group
- Difference in Republican, Democratic language ⇝ Partisan words
- Difference in Liberal, Conservative language ⇝ Ideological Language
Inferring Separating Words

Classification $\rightarrow$ Custom Dictionaries
- Stylometry $\rightarrow$ Classify Authors
- Dictionary based classification $\rightarrow$ Gentzkow and Shapiro (2010) and measures of media slant
- Dictionary based classification $\rightarrow$ Customized to particular setting

Fictitious Prediction Problem $\rightarrow$ Infer words that are indicative of some class/group
- Difference in Republican, Democratic language $\rightarrow$ Partisan words
- Difference in Liberal, Conservative language $\rightarrow$ Ideological Language
- Difference in Secret/Not Secret Language $\rightarrow$ Secretive Language (Gill and Spirling 2014)
Inferring Separating Words

Classification $\Rightarrow$ Custom Dictionaries
- Stylometry $\Rightarrow$ Classify Authors
- Dictionary based classification $\Rightarrow$ Gentzkow and Shapiro (2010) and measures of media slant
- Dictionary based classification $\Rightarrow$ Customized to particular setting

Fictitious Prediction Problem $\Rightarrow$ Infer words that are indicative of some class/group
- Difference in Republican, Democratic language $\Rightarrow$ Partisan words
- Difference in Liberal, Conservative language $\Rightarrow$ Ideological Language
- Difference in Secret/Not Secret Language $\Rightarrow$ Secretive Language (Gill and Spirling 2014)
- Difference in Toy advertising
Inferring Separating Words
Inferring Separating Words
Classification $\leadsto$ Custom Dictionaries
  - Stylometry $\leadsto$ Classify Authors
  - Dictionary based classification $\leadsto$ Gentzkow and Shapiro (2010) and measures of media slant
  - Dictionary based classification $\leadsto$ Customized to particular setting
Fictitious Prediction Problem $\leadsto$ Infer words that are indicative of some class/group
  - Difference in Republican, Democratic language $\leadsto$ Partisan words
  - Difference in Liberal, Conservative language $\leadsto$ Ideological Language
  - Difference in Secret/Not Secret Language $\leadsto$ Secretive Language (Gill and Spirling 2014)
  - Difference in Toy advertising
  - Difference in Language across groups $\leadsto$ Labeling output from Clustering/Topic Models
Inferring Separating Words

Classification $\Rightarrow$ Custom Dictionaries
- Stylometry $\Rightarrow$ Classify Authors
- Dictionary based classification $\Rightarrow$ Gentzkow and Shapiro (2010) and measures of media slant
- Dictionary based classification $\Rightarrow$ Customized to particular setting

Fictitious Prediction Problem $\Rightarrow$ Infer words that are indicative of some class/group
- Difference in Republican, Democratic language $\Rightarrow$ Partisan words
- Difference in Liberal, Conservative language $\Rightarrow$ Ideological Language
- Difference in Secret/Not Secret Language $\Rightarrow$ Secretive Language (Gill and Spirling 2014)
- Difference in Toy advertising
- Difference in Language across groups $\Rightarrow$ Labeling output from Clustering/Topic Models

Vague and Difficult to derive before hand
Congressional Language Across Sources

- Collected 64,033 press releases
- Problem: are they distinct from floor statements (approx. 52,000 during same time)?
  - Yes: press releases have different purposes, targets, and need not relate to official business
  - No: press releases are just reactive to floor activity, will follow floor statements
- Deeper question: what does it mean for two text collections to be different?
  - One Answer: texts used for different purposes
  - Partial answer: identify words that distinguish press releases and floor speeches
Congressional Language Across Sources

Congressional Press Releases and Floor Speeches
Congressional Language Across Sources

Congressional Press Releases and Floor Speeches
- Collected 64,033 press releases
Congressional Language Across Sources

Congressional Press Releases and Floor Speeches

- Collected 64,033 press releases
- Problem: are they distinct from floor statements (approx. 52,000 during same time)?
Congressional Language Across Sources

Congressional Press Releases and Floor Speeches

- Collected 64,033 press releases
- Problem: are they distinct from floor statements (approx. 52,000 during same time)?
  - Yes: press releases have different purposes, targets, and need not relate to official business
Congressional Language Across Sources

Congressional Press Releases and Floor Speeches

- Collected 64,033 press releases
- Problem: are they distinct from floor statements (approx. 52,000 during same time)?
  - Yes: press releases have different purposes, targets, and need not relate to official business
  - No: press releases are just reactive to floor activity, will follow floor statements
Congressional Language Across Sources

Congressional Press Releases and Floor Speeches

- Collected 64,033 press releases
- Problem: are they distinct from floor statements (approx. 52,000 during same time)?
  - Yes: press releases have different purposes, targets, and need not relate to official business
  - No: press releases are just reactive to floor activity, will follow floor statements
- Deeper question: what does it mean for two text collections to be different?
Congressional Language Across Sources

Congressional Press Releases and Floor Speeches

- Collected 64,033 press releases
- Problem: are they distinct from floor statements (approx. 52,000 during same time)?
  - Yes: press releases have different purposes, targets, and need not relate to official business
  - No: press releases are just reactive to floor activity, will follow floor statements
- Deeper question: what does it mean for two text collections to be different?
- One Answer: texts used for different purposes
Congressional Language Across Sources

Congressional Press Releases and Floor Speeches
- Collected 64,033 press releases
- Problem: are they distinct from floor statements (approx. 52,000 during same time)?
  - Yes: press releases have different purposes, targets, and need not relate to official business
  - No: press releases are just reactive to floor activity, will follow floor statements
- Deeper question: what does it mean for two text collections to be different?
- One Answer: texts used for different purposes
- Partial answer: identify words that distinguish press releases and floor speeches
A Method for Identifying Distinguishing Words

Mutual Information

- **Unconditional uncertainty (entropy):**
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
  - Maximum: No. press releases = No. floor statements
  - Minimum: All documents in one category

- **Conditional uncertainty ($X_j$):** (conditional entropy)
  - Condition on presence of word $X_j$
  - Randomly sample a press release
  - Guess press release/floor statement
  - Word presence reduces uncertainty
  - Unrelated: Conditional uncertainty = uncertainty
  - Perfect predictor: Conditional uncertainty = 0

- **Mutual information ($X_j$):** uncertainty - conditional uncertainty ($X_j$)
  - Maximum: Uncertainty $\rightarrow X_j$ is perfect predictor
  - Minimum: 0 $\rightarrow X_j$ fails to separate speeches and floor statements
A Method for Identifying Distinguishing Words

Mutual Information
A Method for Identifying Distinguishing Words

Mutual Information
- Unconditional uncertainty (entropy):
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
  - Maximum: No. press releases = No. floor statements
  - Minimum: All documents in one category

- Conditional uncertainty ($X_j$) (conditional entropy)
  - Condition on presence of word $X_j$
  - Randomly sample a press release
  - Guess press release/floor statement
  - Word presence reduces uncertainty
  - Unrelated: Conditional uncertainty = uncertainty
  - Perfect predictor: Conditional uncertainty = 0

- Mutual information($X_j$):
  - Uncertainty - conditional uncertainty ($X_j$)
  - Maximum: Uncertainty → $X_j$ is perfect predictor
  - Minimum: 0 → $X_j$ fails to separate speeches and floor statements
A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
  - Randomly sample a press release
A Method for Identifying Distinguishing Words

Mutual Information
- Unconditional uncertainty (entropy):
  - Randomly sample a press release
  - Guess press release/floor statement
- Conditional uncertainty ($X_j$) (conditional entropy)
  - Condition on presence of word $X_j$
  - Randomly sample a press release
  - Guess press release/floor statement
  - Word presence reduces uncertainty
- Mutual information ($X_j$): uncertainty - conditional uncertainty ($X_j$)
  - Maximum: Uncertainty $\rightarrow$ $X_j$ is perfect predictor
  - Minimum: 0 $\rightarrow$ $X_j$ fails to separate speeches and floor statements
A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
    - Maximum: No. press releases = No. floor statements

- Conditional uncertainty \((X_j)\) (conditional entropy)
  - Condition on presence of word \(X_j\)
  - Randomly sample a press release
  - Guess press release/floor statement
  - Word presence reduces uncertainty
  - Unrelated: Conditional uncertainty = uncertainty
  - Perfect predictor: Conditional uncertainty = 0

- Mutual information \((X_j)\):
  - Uncertainty - conditional uncertainty \((X_j)\)
  - Maximum: Uncertainty \(\rightarrow X_j\) is perfect predictor
  - Minimum: 0 \(\rightarrow X_j\) fails to separate speeches and floor statements
A Method for Identifying Distinguishing Words

Mutual Information
- Unconditional uncertainty (entropy):
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
    - Maximum: No. press releases = No. floor statements
    - Minimum: All documents in one category

Conditional uncertainty ($X_j$) (conditional entropy)
- Condition on presence of word $X_j$
- Randomly sample a press release
- Guess press release/floor statement
- Word presence reduces uncertainty
- Unrelated: Conditional uncertainty = uncertainty
- Perfect predictor: Conditional uncertainty = 0

Mutual information ($X_j$):
- Uncertainty - conditional uncertainty ($X_j$)
- Maximum: Uncertainty → $X_j$ is perfect predictor
- Minimum: 0 → $X_j$ fails to separate speeches and floor statements
A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
    - Maximum: No. press releases = No. floor statements
    - Minimum: All documents in one category

- Conditional uncertainty ($X_j$) (conditional entropy)
A Method for Identifying Distinguishing Words

Mutual Information

- **Unconditional uncertainty (entropy):**
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
    - Maximum: No. press releases = No. floor statements
    - Minimum: All documents in one category

- **Conditional uncertainty (\(X_j\)) (conditional entropy)**
  - Condition on presence of word \(X_j\)
A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
    - Maximum: No. press releases = No. floor statements
    - Minimum: All documents in one category

- Conditional uncertainty ($X_j$) (conditional entropy)
  - Condition on presence of word $X_j$
  - Randomly sample a press release
A Method for Identifying Distinguishing Words

Mutual Information

- **Unconditional uncertainty (entropy):**
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
    - Maximum: No. press releases = No. floor statements
    - Minimum: All documents in one category

- **Conditional uncertainty ($X_j$) (conditional entropy)**
  - Condition on presence of word $X_j$
  - Randomly sample a press release
  - Guess press release/floor statement
A Method for Identifying Distinguishing Words

Mutual Information

- **Unconditional uncertainty (entropy):**
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
    - Maximum: No. press releases = No. floor statements
    - Minimum: All documents in one category

- **Conditional uncertainty ($X_j$) (conditional entropy)**
  - Condition on presence of word $X_j$
  - Randomly sample a press release
  - Guess press release/floor statement
  - Word presence reduces uncertainty
A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
    - Maximum: No. press releases = No. floor statements
    - Minimum: All documents in one category

- Conditional uncertainty ($X_j$) (conditional entropy)
  - Condition on presence of word $X_j$
  - Randomly sample a press release
  - Guess press release/floor statement
  - Word presence reduces uncertainty
    - Unrelated: Conditional uncertainty = uncertainty
A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
    - Maximum: No. press releases = No. floor statements
    - Minimum: All documents in one category

- Conditional uncertainty ($X_j$) (conditional entropy)
  - Condition on presence of word $X_j$
  - Randomly sample a press release
  - Guess press release/floor statement
  - Word presence reduces uncertainty
    - Unrelated: Conditional uncertainty = uncertainty
    - Perfect predictor: Conditional uncertainty = 0
A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
    - Maximum: No. press releases = No. floor statements
    - Minimum: All documents in one category

- Conditional uncertainty ($X_j$) (conditional entropy)
  - Condition on presence of word $X_j$
  - Randomly sample a press release
  - Guess press release/floor statement
  - Word presence reduces uncertainty
    - Unrelated: Conditional uncertainty = uncertainty
    - Perfect predictor: Conditional uncertainty = 0

- Mutual information($X_j$): uncertainty - conditional uncertainty ($X_j$)
A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
    - Maximum: No. press releases = No. floor statements
    - Minimum: All documents in one category

- Conditional uncertainty ($X_j$) (conditional entropy)
  - Condition on presence of word $X_j$
  - Randomly sample a press release
  - Guess press release/floor statement
  - Word presence reduces uncertainty
    - Unrelated: Conditional uncertainty = uncertainty
    - Perfect predictor: Conditional uncertainty = 0

- Mutual information($X_j$): uncertainty - conditional uncertainty ($X_j$)
  - Maximum: Uncertainty $\rightarrow X_j$ is perfect predictor
A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
    - Maximum: No. press releases = No. floor statements
    - Minimum: All documents in one category

- Conditional uncertainty ($X_j$) (conditional entropy)
  - Condition on presence of word $X_j$
  - Randomly sample a press release
  - Guess press release/floor statement
  - Word presence reduces uncertainty
    - Unrelated: Conditional uncertainty = uncertainty
    - Perfect predictor: Conditional uncertainty = 0

- Mutual information($X_j$): uncertainty - conditional uncertainty ($X_j$)
  - Maximum: Uncertainty $\rightarrow X_j$ is perfect predictor
  - Minimum: 0 $\rightarrow X_j$ fails to separate speeches and floor statements
A Method for Identifying Distinguishing Words

Pr(Press) ≡ Probability selected document press release
Pr(Speech) ≡ Probability selected document speech

Define entropy

\[ H(Doc) = -\sum_{t \in \{Pre, Spe\}} Pr(t) \log_2 Pr(t) \]

- Encodes bits
- Maximum: Pr(Press) = Pr(Speech) = 0.5
- Minimum: Pr(Press) \rightarrow 0 (or Pr(Press) \rightarrow 1)
A Method for Identifying Distinguishing Words

- $\text{Pr}(\text{Press}) \equiv \text{Probability selected document press release}$
A Method for Identifying Distinguishing Words

- \( \text{Pr(Press)} \equiv \text{Probability selected document press release} \)
- \( \text{Pr(Speech)} \equiv \text{Probability selected document speech} \)
A Method for Identifying Distinguishing Words

- \( \text{Pr}(\text{Press}) \equiv \text{Probability selected document press release} \)
- \( \text{Pr}(\text{Speech}) \equiv \text{Probability selected document speech} \)
- Define \textbf{entropy} \( H(\text{Doc}) \)
A Method for Identifying Distinguishing Words

- $\Pr(\text{Press}) \equiv$ Probability selected document press release
- $\Pr(\text{Speech}) \equiv$ Probability selected document speech
- Define entropy $H(\text{Doc})$

\[
H(\text{Doc}) = - \sum_{t \in \{\text{Pre,Spe}\}} \Pr(t) \log_2 \Pr(t)
\]
A Method for Identifying Distinguishing Words

- $\Pr(\text{Press}) \equiv$ Probability selected document press release
- $\Pr(\text{Speech}) \equiv$ Probability selected document speech
- Define entropy $H(\text{Doc})$

$$H(\text{Doc}) = - \sum_{t \in \{\text{Pre, Spe}\}} \Pr(t) \log_2 \Pr(t)$$

- $\log_2$? Encodes bits
A Method for Identifying Distinguishing Words

- $\Pr(\text{Press}) \equiv \text{Probability selected document press release}$
- $\Pr(\text{Speech}) \equiv \text{Probability selected document speech}$
- Define entropy $H(\text{Doc})$

$$H(\text{Doc}) = - \sum_{t \in \{\text{Pre, Spe}\}} \Pr(t) \log_2 \Pr(t)$$

- $\log_2$? Encodes bits
- Maximum: $\Pr(\text{Press}) = \Pr(\text{Speech}) = 0.5$
A Method for Identifying Distinguishing Words

- \( \Pr(\text{Press}) \equiv \text{Probability selected document press release} \)
- \( \Pr(\text{Speech}) \equiv \text{Probability selected document speech} \)
- Define entropy \( H(\text{Doc}) \)

\[
H(\text{Doc}) = - \sum_{t \in \{\text{Pre}, \text{Spe}\}} \Pr(t) \log_2 \Pr(t)
\]

- \( \log_2 \)? Encodes bits
- Maximum: \( \Pr(\text{Press}) = \Pr(\text{Speech}) = 0.5 \)
- Minimum: \( \Pr(\text{Press}) \rightarrow 0 \) (or \( \Pr(\text{Press}) \rightarrow 1 \))
A Method for Identifying Distinguishing Words

- Consider presence/absence of word $X_j$
- Define conditional entropy $H(Doc | X_j)$

$$H(Doc | X_j) = -\sum_{s=0}^{\sum_t \in \{Pre, Spe\}} Pr(t, X_j = s) \log_2 Pr(t | X_j = s)$$

- Maximum: $X_j$ unrelated to Press Releases/Floor Speeches
- Minimum: $X_j$ is a perfect predictor of press release/floor speech
A Method for Identifying Distinguishing Words

- Consider presence/absence of word $X_j$

$$H(Doc | X_j) = -\sum_{s=0}^{\sum_{t \in \{Pre, Spe\}} Pr(t, X_j = s) \log_2 Pr(t | X_j = s)$$

- Maximum: $X_j$ unrelated to Press Releases/Floor Speeches
- Minimum: $X_j$ is a perfect predictor of press release/floor speech
A Method for Identifying Distinguishing Words

- Consider presence/absence of word $X_j$
- Define conditional entropy $H(\text{Doc}|X_j)$
A Method for Identifying Distinguishing Words

- Consider presence/absence of word $X_j$
- Define **conditional entropy** $H(Doc|X_j)$

$$H(Doc|X_j) = -\sum_{s=0}^{1} \sum_{t \in \{Pre, Spe\}} \Pr(t, X_j = s) \log_2 \Pr(t|X_j = s)$$

- Maximum: $X_j$ unrelated to Press Releases/Floor Speeches
- Minimum: $X_j$ is a perfect predictor of press release/floor speech
A Method for Identifying Distinguishing Words

- Consider presence/absence of word $X_j$
- Define conditional entropy $H(Doc|X_j)$

$$H(Doc|X_j) = - \sum_{s=0}^{1} \sum_{t \in \{Pre, Spe\}} Pr(t, X_j = s) \log_2 Pr(t|X_j = s)$$

- Maximum: $X_j$ unrelated to Press Releases/Floor Speeches
- Minimum: $X_j$ is a perfect predictor of press release/floor speech
A Method for Identifying Distinguishing Words

- Consider presence/absence of word $X_j$
- Define conditional entropy $H(Doc|X_j)$

$$H(Doc|X_j) = - \sum_{s=0}^{1} \sum_{t \in \{Pre,Spe\}} \Pr(t, X_j = s) \log_2 \Pr(t|X_j = s)$$

- Maximum: $X_j$ unrelated to Press Releases/Floor Speeches
- Minimum: $X_j$ is a perfect predictor of press release/floor speech
A Method for Identifying Distinguishing Words

- Define Mutual Information ($\text{Mutual Information}(X_j)$) as
  \[ \text{Mutual Information}(X_j) = H(\text{Doc}) - H(\text{Doc} | X_j) \]
  - Maximum: entropy $\Rightarrow H(\text{Doc} | X_j) = 0$
  - Minimum: 0 $\Rightarrow H(\text{Doc} | X_j) = H(\text{Doc})$.

Bigger mutual information $\Rightarrow$ better discrimination

Objective function and optimization $\Rightarrow$ estimate probabilities that we then place in mutual information
A Method for Identifying Distinguishing Words

- Define Mutual Information($X_j$) as

$$\text{Mutual Information}(X_j) = H(\text{Doc}) - H(\text{Doc}|X_j)$$

- Maximum: entropy $\Rightarrow H(\text{Doc}|X_j) = 0$
- Minimum: 0 $\Rightarrow H(\text{Doc}|X_j) = H(\text{Doc})$
A Method for Identifying Distinguishing Words

- Define **Mutual Information** ($X_j$) as

\[
\text{Mutual Information}(X_j) = H(\text{Doc}) - H(\text{Doc}|X_j)
\]

Bigger mutual information ⇒ better discrimination

**Objective function and optimization** → estimate probabilities that we then place in mutual information
A Method for Identifying Distinguishing Words

- Define **Mutual Information**($X_j$) as

\[
\text{Mutual Information}(X_j) = H(Doc) - H(Doc|X_j)
\]

- Maximum: entropy $\Rightarrow H(Doc|X_j) = 0$

Bigger mutual information $\Rightarrow$ better discrimination

Objective function and optimization $\Rightarrow$ estimate probabilities that we then place in mutual information
A Method for Identifying Distinguishing Words

- Define Mutual Information ($X_j$) as

$$\text{Mutual Information}(X_j) = H(\text{Doc}) - H(\text{Doc} | X_j)$$

- Maximum: entropy $\Rightarrow H(\text{Doc} | X_j) = 0$
- Minimum: $0 \Rightarrow H(\text{Doc} | X_j) = H(\text{Doc})$. 
A Method for Identifying Distinguishing Words

- Define **Mutual Information** \( (X_j) \) as

\[
\text{Mutual Information}(X_j) = H(\text{Doc}) - H(\text{Doc} | X_j)
\]

- Maximum: entropy \( \Rightarrow H(\text{Doc} | X_j) = 0 \)
- Minimum: 0 \( \Rightarrow H(\text{Doc} | X_j) = H(\text{Doc}) \).

Bigger mutual information \( \Rightarrow \) better discrimination
A Method for Identifying Distinguishing Words

- Define **Mutual Information**$(X_j)$ as

\[
\text{Mutual Information}(X_j) = H(\text{Doc}) - H(\text{Doc}|X_j)
\]

- Maximum: entropy $\Rightarrow H(\text{Doc}|X_j) = 0$
- Minimum: $0 \Rightarrow H(\text{Doc}|X_j) = H(\text{Doc})$.

Bigger mutual information $\Rightarrow$ better discrimination

Objective function and optimization $\Rightarrow$ estimate probabilities that we then place in mutual information
A Method for Identifying Distinguishing Words

Formula for mutual information
(based on ML estimates of probabilities)

\[ n_p = \text{Number Press Releases} \]
\[ n_s = \text{Number of Speeches} \]
\[ D = n_p + n_s \]
\[ n_j = \sum_{i=1}^{D} X_{i,j} \quad \text{(No. docs } X_j \text{ appears)} \]
\[ n_{-j} = \text{No. docs } X_j \text{ does not appear} \]
\[ n_{j,p} = \text{No. press and } X_j \]
\[ n_{j,s} = \text{No. speech and } X_j \]
\[ n_{-j,p} = \text{No. press and not } X_j \]
\[ n_{-j,s} = \text{No. speech and not } X_j \]
A Method for Identifying Distinguishing Words

Formula for Mutual Information

\[
\text{MI}(X_j) = \frac{n_{j,p}}{D} \log_2 \left( \frac{n_{j,p}D}{n_j n_p} \right) + \frac{n_{j,s}}{D} \log_2 \left( \frac{n_{j,s}D}{n_j n_s} \right) \\
+ \frac{n_{-j,p}}{D} \log_2 \left( \frac{n_{-j,p}D}{n_{-j} n_p} \right) + \frac{n_{-j,s}}{D} \log_2 \left( \frac{n_{-j,s}D}{n_{-j} n_s} \right).
\]
What’s Different About Press Releases

What’s Different?

Justin Grimmer (Stanford University)
What’s Different About Press Releases

What’s Different?
What’s Different About Press Releases

What’s Different?
What’s Different About Press Releases

What’s Different?
What’s Different About Press Releases

What’s Different?
- Press Releases: Credit Claiming
What’s Different About Press Releases

- Press Releases: Credit Claiming
- Floor Speeches: Procedural Words
What’s Different About Press Releases

What’s Different?

- Press Releases: Credit Claiming
- Floor Speeches: Procedural Words
- Validate: Manual Classification
What’s Different About Press Releases

What’s Different?

- Press Releases: Credit Claiming
- Floor Speeches: Procedural Words
- Validate: Manual Classification
- Sample 500 Press Releases, 500 Floor Speeches
What’s Different About Press Releases

What’s Different?

- Press Releases: Credit Claiming
- Floor Speeches: Procedural Words
- Validate: Manual Classification
- Sample 500 Press Releases, 500 Floor Speeches
- Credit Claiming: 36% Press Releases, 4% Floor Speeches
What’s Different About Press Releases

What’s Different?

- Press Releases: Credit Claiming
- Floor Speeches: Procedural Words
- Validate: Manual Classification
- Sample 500 Press Releases, 500 Floor Speeches
- Credit Claiming: 36% Press Releases, 4% Floor Speeches
- Procedural: 0% Press Releases, 44% Floor Speeches
What’s Different About Press Releases

- Press Releases: Credit Claiming
- Floor Speeches: Procedural Words
- Validate: Manual Classification
- Sample 500 Press Releases, 500 Floor Speeches
- Credit Claiming: 36% Press Releases, 4% Floor Speeches
- Procedural: 0% Press Releases, 44% Floor Speeches
- Validate: Topic Classification

Justin Grimmer (Stanford University)
Fightin’ Words

An Introduction to Regularization

Monroe, Colaresi, and Quinn (2009) what makes a word partisan?

Argue for using Log Odds Ratio, weighted by variance

Recall: For some event \( E \) and \( F \)

\[
\text{Odds}(E) = \frac{P(E)}{1 - P(E)}
\]

\[
\text{Odds Ratio}(E, F) = \frac{P(E)}{1 - P(E)} \div \frac{P(F)}{1 - P(F)}
\]

\[
\text{Log Odds Ratio}(E, F) = \log \left(\frac{P(E)}{1 - P(E)}\right) - \log \left(\frac{P(F)}{1 - P(F)}\right)
\]
Fightin’ Words $\Rightarrow$ An Introduction to Regularization

Monroe, Colaresi, and Quinn (2009) $\Rightarrow$ what makes a word partisan? Argue for using Log Odds Ratio, weighted by variance

Recall: For some event $E$ and $F$

$$
\begin{align*}
\text{Odds}(E) &= \frac{P(E)}{1 - P(E)} \\
\text{Odds Ratio}(E, F) &= \frac{P(E)}{1 - P(E)} \div \frac{P(F)}{1 - P(F)} \\
\log\text{Odds Ratio}(E, F) &= \log\left(\frac{P(E)}{1 - P(E)}\right) - \log\left(\frac{P(F)}{1 - P(F)}\right)
\end{align*}
$$

Strategy $\Rightarrow$ Construct objective function on proportions (and then calculate log-odds)
Fightin’ Words — An Introduction to Regularization

Monroe, Colaresi, and Quinn (2009) — what makes a word partisan? Argue for using Log Odds Ratio, weighted by variance

Recall: For some event $E$ and $F$

$$P(E) = 1 - P(E^c)$$

Odds($E$) = $P(E) / (1 - P(E))$

Odds Ratio($E$, $F$) = $P(E) / (1 - P(E)) / P(F) / (1 - P(F))$

Log Odds Ratio($E$, $F$) = $\log \left( P(E) / (1 - P(E)) \right) - \log \left( P(F) / (1 - P(F)) \right)$

Strategy — Construct objective function on proportions (and then calculate log-odds)
Fightin’ Words ⇔ An Introduction to Regularization

Monroe, Colaresi, and Quinn (2009) ⇔ what makes a word partisan?
Argue for using Log Odds Ratio, weighted by variance
Recall: For some event $E$ and $F$

$$P(E) = 1 - P(E^c)$$
Monroe, Colaresi, and Quinn (2009)\textarrow{} what makes a word partisan? Argue for using \textbf{Log Odds Ratio}, weighted by variance

Recall: For some event $E$ and $F$

\[
P(E) = 1 - P(E^c)\]

\[
\text{Odds}(E) = \frac{P(E)}{1 - P(E)}\]

\[
\text{Odds Ratio}(E, F) = \frac{P(E)}{1 - P(E)} \frac{1 - P(F)}{1 - P(F)}\]

\[
\text{Log Odds Ratio}(E, F) = \log \left( \frac{P(E)}{1 - P(E)} \right) - \log \left( \frac{P(F)}{1 - P(F)} \right)\]
Fightin’ Words\textarrow An Introduction to Regularization

Monroe, Colaresi, and Quinn (2009)\textarrow what makes a word partisan?
Argue for using Log Odds Ratio, weighted by variance
Recall: For some event $E$ and $F$

\[
P(E) = 1 - P(E^c)
\]
\[
\text{Odds}(E) = \frac{P(E)}{1 - P(E)}
\]
\[
\text{Odds Ratio}(E, F) = \frac{P(E)}{1 - P(E)} \cdot \frac{P(F)}{1 - P(F)}
\]
Fightin’ Words ⇾ An Introduction to Regularization

Monroe, Colaresi, and Quinn (2009) ⇾ what makes a word partisan?
Argue for using Log Odds Ratio, weighted by variance
Recall: For some event $E$ and $F$

$$P(E) = 1 - P(E^c)$$

$$\text{Odds}(E) = \frac{P(E)}{1 - P(E)}$$

$$\text{Odds Ratio}(E, F) = \frac{\frac{P(E)}{1 - P(E)}}{\frac{P(F)}{1 - P(F)}}$$

$$\text{Log Odds Ratio}(E, F) = \log \left( \frac{P(E)}{1 - P(E)} \right) - \log \left( \frac{P(F)}{1 - P(F)} \right)$$
Fightin’ Words⇝ An Introduction to Regularization

Monroe, Colaresi, and Quinn (2009)⇝ what makes a word partisan?
Argue for using Log Odds Ratio, weighted by variance
Recall: For some event $E$ and $F$

\[
P(E) = 1 - P(E^c)
\]
\[
\text{Odds}(E) = \frac{P(E)}{1 - P(E)}
\]
\[
\text{Odds Ratio}(E, F) = \frac{P(E)}{(1 - P(E))} \frac{P(F)}{1 - P(F)}
\]
\[
\text{Log Odds Ratio}(E, F) = \log \left( \frac{P(E)}{1 - P(E)} \right) - \log \left( \frac{P(F)}{1 - P(F)} \right)
\]

Strategy⇝ Construct objective function on proportions (and then calculate log-odds)
Fightin’ Words—An Introduction to Regularization

Monroe, Colaresi, and Quinn (2009)—what makes a word partisan? Argue for using Log Odds Ratio, weighted by variance

Recall: For some event $E$ and $F$

\[
P(E) = 1 - P(E^c)
\]

\[
\text{Odds}(E) = \frac{P(E)}{1 - P(E)}
\]

\[
\text{Odds Ratio}(E, F) = \frac{P(E)}{1-P(E)} \frac{(1-P(E))}{P(F)}
\]

\[
\text{Log Odds Ratio}(E, F) = \log \left( \frac{P(E)}{1 - P(E)} \right) - \log \left( \frac{P(F)}{1 - P(F)} \right)
\]

Strategy—Construct objective function on *proportions* (and then calculate log-odds)
Objective Function

Suppose we’re interested in how a word separates partisan speech.
\( Y = (\text{Republican, Republican, Democrat, \ldots, Republican}) \)
\( X = \text{Unnormalized matrix of word counts } N \times J \)
Define

\[
\mathbf{x}_{\text{Republican}} = (\sum_{i=1}^{N} I(Y_i = \text{Republican})X_{i1}, \sum_{i=1}^{N} I(Y_i = \text{Republican})X_{i2}, \ldots, \sum_{i=1}^{N} I(Y_i = \text{Republican})X_{iJ})
\]

with \( N_{\text{Republican}} = \text{Total number of Republican words} \)
Objective Function

\[ \pi \sim \text{Dirichlet}(\alpha) \]

\[ x_{\text{Republican}} \mid \pi_{\text{Republican}} \sim \text{Multinomial}(N_{\text{Republican}}, \pi_{\text{Republican}}) \]

This implies an objective function on \( \pi \),

\[ p(\pi \mid \alpha, X, Y) \propto p(\pi \mid \alpha) p(x_{\text{Republican}} \mid \pi, \alpha, Y) \]

\[ \propto \Gamma \left( \sum_{j=1}^{J} \alpha_j \right) \prod_{j=1}^{J} \Gamma(\alpha_j) \prod_{j=1}^{J} \pi_{\alpha_j}^{x_{\text{Republican}}_j - 1} \]

\[ p(\pi \mid \alpha, X, Y) \]

is a Dirichlet distribution:

\[ \pi_{\text{Republican}, j} = x_{\text{Republican}, j} + \alpha_j \]

\[ N_{\text{Republican}} + \sum_{j=1}^{J} \alpha_j \]
Objective Function

\[ \pi \text{Republican} \sim \text{Dirichlet}(\alpha) \]
Objective Function

\[ \pi_{\text{Republican}} \sim \text{Dirichlet}(\alpha) \]

\[ x_{\text{Republican}} | \pi_{\text{Republican}} \sim \text{Multinomial}(N_{\text{Republican}}, \pi_{\text{Republican}}) \]
Objective Function

\[ \pi_{\text{Republican}} \sim \text{Dirichlet}(\alpha) \]
\[ x_{\text{Republican}}|\pi_{\text{Republican}} \sim \text{Multinomial}(N_{\text{Republican}}, \pi_{\text{Republican}}) \]

This implies an objective function on \( \pi \),
Objective Function

\[
\begin{align*}
\pi_{\text{Republican}} & \sim \text{Dirichlet}(\alpha) \\
x_{\text{Republican}} | \pi_{\text{Republican}} & \sim \text{Multinomial}(N_{\text{Republican}}, \pi_{\text{Republican}})
\end{align*}
\]

This implies an objective function on \(\pi\),

\[
p(\pi | \alpha, X, Y) \propto p(\pi | \alpha) p(x_{\text{Republican}} | \pi \alpha, Y)
\]
Objective Function

\[ \pi_{\text{Republican}} \sim \text{Dirichlet}(\alpha) \]
\[ x_{\text{Republican}} | \pi_{\text{Republican}} \sim \text{Multinomial}(N_{\text{Republican}}, \pi_{\text{Republican}}) \]

This implies an objective function on \( \pi \),

\[
p(\pi | \alpha, X, Y) \propto p(\pi | \alpha)p(x_{\text{Republican}} | \pi \alpha, Y)
\]
\[
\propto \frac{\Gamma(\sum_{j=1}^{J} \alpha_j)}{\prod_{j} \Gamma(\alpha_j)} \prod_{j=1}^{J} \pi_{j}^{\alpha_j - 1} \pi_{\text{Republican}, j}^{x_{\text{Republican}, j}}
\]
Objective Function

\[ \pi_{\text{Republican}} \sim \text{Dirichlet}(\alpha) \]
\[ x_{\text{Republican}} | \pi_{\text{Republican}} \sim \text{Multinomial}(N_{\text{Republican}}, \pi_{\text{Republican}}) \]

This implies an objective function on \( \pi \),

\[
p(\pi | \alpha, X, Y) \propto p(\pi | \alpha)p(x_{\text{Republican}} | \pi \alpha, Y)
\]
\[
\propto \frac{\Gamma(\sum_{j=1}^{J} \alpha_j)}{\prod_{j} \Gamma(\alpha_j)} \prod_{j=1}^{J} \pi_j^{\alpha_j - 1} x_{\text{Republican},j} \pi_j
\]

\( p(\pi | \alpha, X, Y) \) is a Dirichlet distribution:
Objective Function

\[ \pi_{\text{Republican}} \sim \text{Dirichlet}(\alpha) \]
\[ x_{\text{Republican}} | \pi_{\text{Republican}} \sim \text{Multinomial}(N_{\text{Republican}}, \pi_{\text{Republican}}) \]

This implies an objective function on \( \pi \),

\[
p(\pi | \alpha, X, Y) \propto p(\pi | \alpha)p(x_{\text{Republican}} | \pi, Y) \propto \frac{\Gamma(\sum_{j=1}^{J} \alpha_j)}{\prod_{j} \Gamma(\alpha_j)} \prod_{j=1}^{J} \pi_{j}^{\alpha_j-1} x_{\text{Republican},j} \]

\( p(\pi | \alpha, X, Y) \) is a Dirichlet distribution:

\[ \pi_{\text{Republican},j}^{*} = \frac{x_{\text{Republican},j} + \alpha_j}{N_{\text{Republican}} + \sum_{j=1}^{J} \alpha_j} \]
Objective Function

\[ \pi_{\text{Republican}} \sim \text{Dirichlet}(\alpha) \]
\[ x_{\text{Republican}} | \pi_{\text{Republican}} \sim \text{Multinomial}(N_{\text{Republican}}, \pi_{\text{Republican}}) \]

This implies an objective function on \( \pi \),

\[
p(\pi | \alpha, X, Y) \propto p(\pi | \alpha) p(x_{\text{Republican}} | \pi \alpha, Y)
\]
\[
\propto \frac{\Gamma(\sum_{j=1}^{J} \alpha_j)}{\prod_{j} \Gamma(\alpha_j)} \prod_{j=1}^{J} \pi_{j}^{\alpha_j-1} x_{\text{Republican},j} \pi_{j}
\]

\( p(\pi | \alpha, X, Y) \) is a Dirichlet distribution:

\[
\pi^*_{\text{Republican},j} = \frac{x_{\text{Republican},j} + \alpha_j}{N_{\text{Republican}} + \sum_{j=1}^{J} \alpha_j}
\]
Calculating Log Odds Ratio

Define log Odds Ratio\(_j\) as

\[
\text{log Odds Ratio}_j = \log \left( \frac{\pi_{\text{Republican}_j}}{1 - \pi_{\text{Republican}_j}} \right) - \log \left( \frac{\pi_{\text{Democratic}_j}}{1 - \pi_{\text{Democratic}_j}} \right)
\]

Var(log Odds Ratio\(_j\)) \approx \frac{1}{\chi_{jD}^2 + \alpha_j} + \frac{1}{\chi_{jR}^2 + \alpha_j}

Std. Log Odds\(_j\) = \frac{\text{log Odds Ratio}_j}{\sqrt{\text{Var(log Odds Ratio}_j)}}
Applying the Model

How do Republicans and Democrats differ in debate?
Condition on topic and examine word usage

- Press Releases (64,033)
- Topic Coded (Structural Topic Model)
- Given press release is about topic, what are the features that distinguish Republican and Democratic language?
Multinomial Inverse Regression

- In classification we’re generally interested in:

\[ E[Y|X] = g(X_1, X_2, \ldots, X_J) \]

- Problem: \( J \) might be very, very big.
- Potential solution \( \Rightarrow \) invert regression

\[ E[X|Y] = g(Y) \]

- Inversion is particularly useful for feature selection
Multinomial Inverse Regression: Objective Function (Taddy 2014)

As before, $x_{\text{Republican}}$ to be the Republican count vector.
Multinomial Inverse Regression: Objective Function (Taddy 2014)

As before, $\mathbf{x}_{\text{Republican}}$ to be the Republican count vector.

\[ \mathbf{x}_{\text{Republican}} \sim \text{Multinomial}(N_{\text{Republican}}, \pi_{\text{Republican}}) \]
Multinomial Inverse Regression: Objective Function (Taddy 2014)

As before, $x_{\text{Republican}}$ to be the Republican count vector.

\[
x_{\text{Republican}} \sim \text{Multinomial}(N_{\text{Republican}}, \pi_{\text{Republican}})
\]

\[
\pi_{\text{Republican},j} = \frac{\exp[\alpha_j + I(\text{Republican})\phi_j]}{\sum_{l=1}^{J} \exp[\alpha_l + I(\text{Republican})\phi_l]}
\]
Multinomial Inverse Regression: Objective Function
(Taddy 2014)

As before, \( x_{\text{Republican}} \) to be the Republican count vector.

\[
x_{\text{Republican}} \sim \text{Multinomial}(N_{\text{Republican}}, \pi_{\text{Republican}})
\]

\[
\pi_{\text{Republican}, j} = \frac{\exp[\alpha_j + I(\text{Republican})\phi_j]}{\sum_{l=1}^{J} \exp[\alpha_l + I(\text{Republican})\phi_l]}
\]

\[
\phi_j \sim \text{Laplace}(\lambda_j)
\]
Multinomial Inverse Regression: Objective Function (Taddy 2014)

As before, $x_{\text{Republican}}$ to be the Republican count vector.

\[
x_{\text{Republican}} \sim \text{Multinomial}(N_{\text{Republican}}, \pi_{\text{Republican}})
\]

\[
\pi_{\text{Republican}, j} = \frac{\exp[\alpha_j + I(\text{Republican})\phi_j]}{\sum_{l=1}^{J} \exp[\alpha_l + I(\text{Republican})\phi_l]}
\]

\[
\phi_j \sim \text{Laplace}(\lambda_j)
\]

\[
\lambda_j \sim \text{Gamma}(s, r)
\]
Multinomial Inverse Regression: Objective Function
(Taddy 2014)

As before, $x_{\text{Republican}}$ to be the Republican count vector.

$$x_{\text{Republican}} \sim \text{Multinomial}(N_{\text{Republican}}, \pi_{\text{Republican}})$$

$$\pi_{\text{Republican}, j} = \frac{\exp[\alpha_j + I(\text{Republican})\phi_j]}{\sum_{l=1}^{J} \exp[\alpha_l + I(\text{Republican})\phi_l]}$$

$$\phi_j \sim \text{Laplace}(\lambda_j)$$

$$\lambda_j \sim \text{Gamma}(s, r)$$

Laplace priors $\leadsto$ regularize or shrink estimates toward zero
Multinomial Inverse Regression: Objective Function
(Taddy 2014)

As before, $x_{\text{Republican}}$ to be the Republican count vector.

$$
\begin{align*}
    x_{\text{Republican}} & \sim \text{Multinomial}(N_{\text{Republican}}, \pi_{\text{Republican}}) \\
    \pi_{\text{Republican}, j} &= \frac{\exp[\alpha_j + I(\text{Republican})\phi_j]}{\sum_{l=1}^{J} \exp[\alpha_l + I(\text{Republican})\phi_l]} \\
    \phi_j & \sim \text{Laplace}(\lambda_j) \\
    \lambda_j & \sim \text{Gamma}(s, r)
\end{align*}
$$

Laplace priors $\sim$ regularize or shrink estimates toward zero
Laplace priors $\sim$ Equivalent to $L1$ or lasso penalization
Multinomial Inverse Regression: Objective Function (Taddy 2014)

As before, \( x_{\text{Republican}} \) to be the Republican count vector.

\[
\begin{align*}
    x_{\text{Republican}} & \sim \text{Multinomial}(N_{\text{Republican}}, \pi_{\text{Republican}}) \\
    \pi_{\text{Republican},j} &= \frac{\exp[\alpha_j + I(\text{Republican})\phi_j]}{\sum_{l=1}^{J} \exp[\alpha_l + I(\text{Republican})\phi_l]} \\
    \phi_j & \sim \text{Laplace}(\lambda_j) \\
    \lambda_j & \sim \text{Gamma}(s, r)
\end{align*}
\]

Laplace priors \( \rightsquigarrow \) regularize or shrink estimates toward zero
Laplace priors \( \rightsquigarrow \) Equivalent to \( L1 \) or lasso penalization
Gamma-Lasso prior
Multinomial Inverse Regression: Objective Function (Taddy 2014)

As before, $x_{\text{Republican}}$ to be the Republican count vector.

\[
x_{\text{Republican}} \sim \text{Multinomial}(N_{\text{Republican}}, \pi_{\text{Republican}})
\]

\[
\pi_{\text{Republican}, j} = \frac{\exp[\alpha_j + I(\text{Republican})\phi_j]}{\sum_{l=1}^{J} \exp[\alpha_l + I(\text{Republican})\phi_l]}
\]

\[
\phi_j \sim \text{Laplace}(\lambda_j)
\]

\[
\lambda_j \sim \text{Gamma}(s, r)
\]

Laplace priors $\Rightarrow$ regularize or shrink estimates toward zero
Laplace priors $\Rightarrow$ Equivalent to $L1$ or lasso penalization
Gamma-Lasso prior
Optimization $\Rightarrow$ Coordinate descent (paper is great!) $\Rightarrow$ textir package
Applying Multinomial Inverse Regression: Objective Function

Taddy (2014) considers speeches made on Congressional floor in 2005 “Most” Republican words
Applying Multinomial Inverse Regression: Objective Function

Taddy (2014) considers speeches made on Congressional floor in 2005 “Most” Republican words
un.official, term. care. insurance, weapons. grade. plutonium
million. illegal. immigrant, grand ole opry, ..., personal. injury. lawyer
Applying Multinomial Inverse Regression: Objective Function

Taddy (2014) considers speeches made on Congressional floor in 2005
“Most” Republican words
un.official, term.care.insurance, weapons.grade.plutonium
million.illegal.immigrant, grand ole opry, ..., personal.injury.lawyer
“Most” Democratic Words
Taddy (2014) considers speeches made on Congressional floor in 2005
“Most” Republican words
un.official, term.care.insurance, weapons.grade.plutonium
million.illegal.immigrant, grand ole opry, ..., personal.injury.lawyer
“Most” Democratic Words
wild.bird, death.penalty.system, record.budget.deficit
security.private.account, able.buy.gun