Day 4: Machine Learning and Scaling for Texts

Kenneth Benoit

TCD 2016: Quantitative Text Analysis

March 18, 2016

Classification as a goal

- Machine learning focuses on identifying classes (classification), while social science is typically interested in locating things on latent traits (scaling)
- But the two methods overlap and can be adapted – will demonstrate later using the Naive Bayes classifier
- Applying lessons from machine to learning to supervised scaling, we can
  - Apply classification methods to scaling
  - improve it using lessons from machine learning

Supervised v. unsupervised methods compared

- The goal (in text analysis) is to differentiate *documents* from one another, treating them as "bags of words"
- Different approaches:
  - *Supervised methods* require a training set that exmplify constrasting classes, identified by the researcher
  - *Unsupervised methods* scale documents based on patterns of similarity from the term-document matrix, without requiring a training step

Supervised v. unsupervised methods compared

- The goal (in text analysis) is to differentiate *documents* from one another, treating them as "bags of words"
- Different approaches:
  - *Supervised methods* require a training set that exmplify constrasting classes, identified by the researcher
  - *Unsupervised methods* scale documents based on patterns of similarity from the term-document matrix, without requiring a training step
- Relative advantage of supervised methods:
  You already know the dimension being scaled, because you set it in the training stage

Supervised v. unsupervised methods compared

- The goal (in text analysis) is to differentiate *documents* from one another, treating them as "bags of words"
- Different approaches:
  - *Supervised methods* require a training set that exmplify constrasting classes, identified by the researcher
  - *Unsupervised methods* scale documents based on patterns of similarity from the term-document matrix, without requiring a training step
- Relative advantage of supervised methods:
  You already know the dimension being scaled, because you set it in the training stage
- Relative disadvantage of supervised methods:
  You *must* already know the dimension being scaled, because you have to feed it good sample documents in the training stage

Supervised v. unsupervised methods: Examples

- General examples:
  - Supervised: Naive Bayes, k-Nearest Neighbor, Support Vector Machines (SVM)
  - Unsupervised: correspondence analysis, IRT models, factor analytic approaches

Supervised v. unsupervised methods: Examples

- General examples:
  - Supervised: Naive Bayes, k-Nearest Neighbor, Support Vector Machines (SVM)
  - Unsupervised: correspondence analysis, IRT models, factor analytic approaches
- Political science applications
  - Supervised: Wordscores (LBG 2003); SVMs (Yu, Kaufman and Diermeier 2008); Naive Bayes (Evans et al 2007)
  - Unsupervised "Wordfish" (Slapin and Proksch 2008); Correspondence analysis (Schonhardt-Bailey 2008); two-dimensional IRT (Monroe and Maeda 2004)

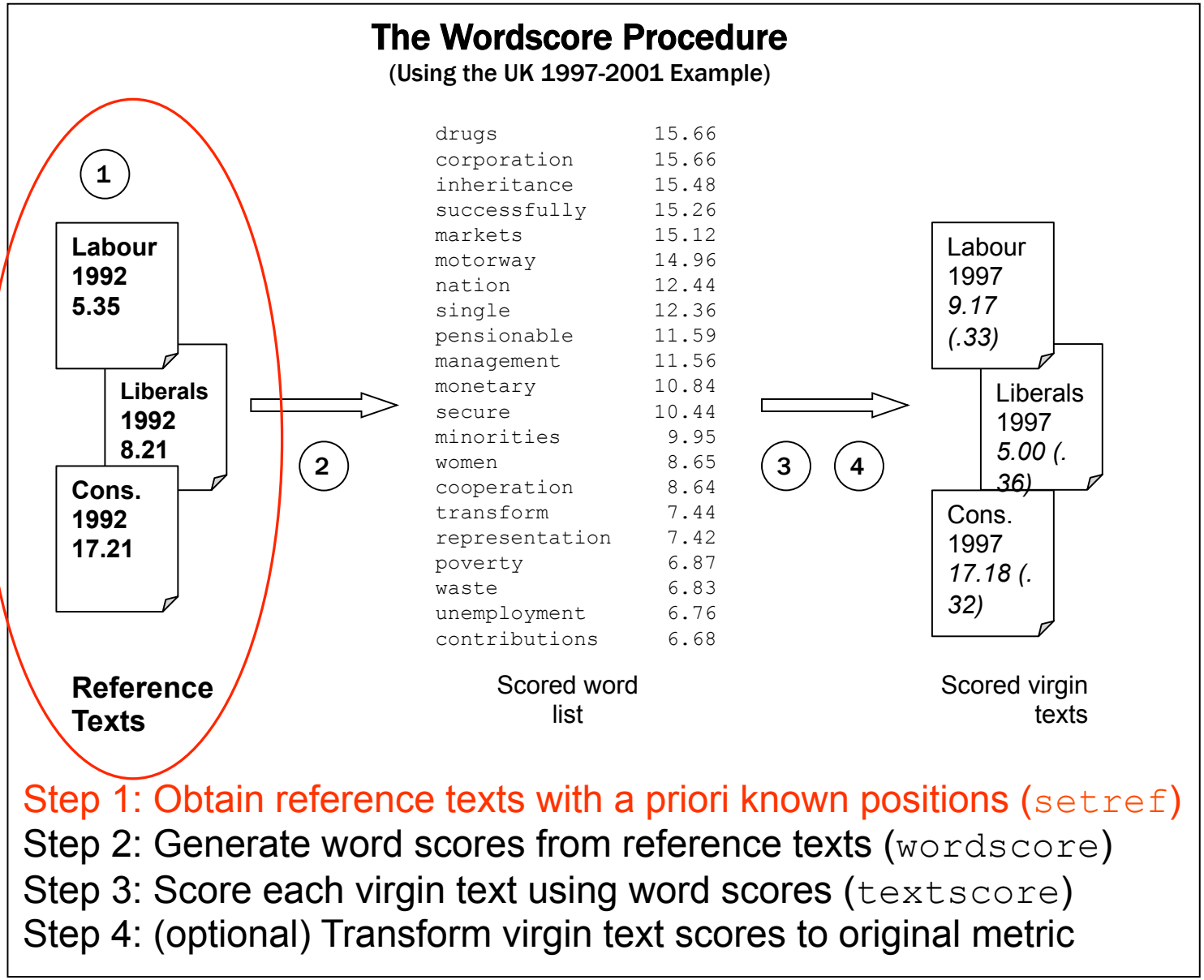
Classification to Scaling

- The class predictions for a collection of words from NB can be adapted to scaling
- The intermediate steps from NB turn out to be excellent for scaling purposes, and identical to Laver, Benoit and Garry's "Wordscores"
- There are certain things from machine learning that ought to be adopted when classification methods are used for scaling
  - Feature selection
  - Stemming/pre-processing

Wordscores conceptually

- Two sets of texts
  - Reference texts: texts about which we know something (a scalar dimensional score)
  - Virgin texts: texts about which we know nothing (but whose dimensional score wed like to know)
- These are analogous to a "training set" and a "test set" in classification
- Basic procedure:
  1. Analyze reference texts to obtain word scores
  2. Use word scores to score virgin texts

Wordscores Procedure



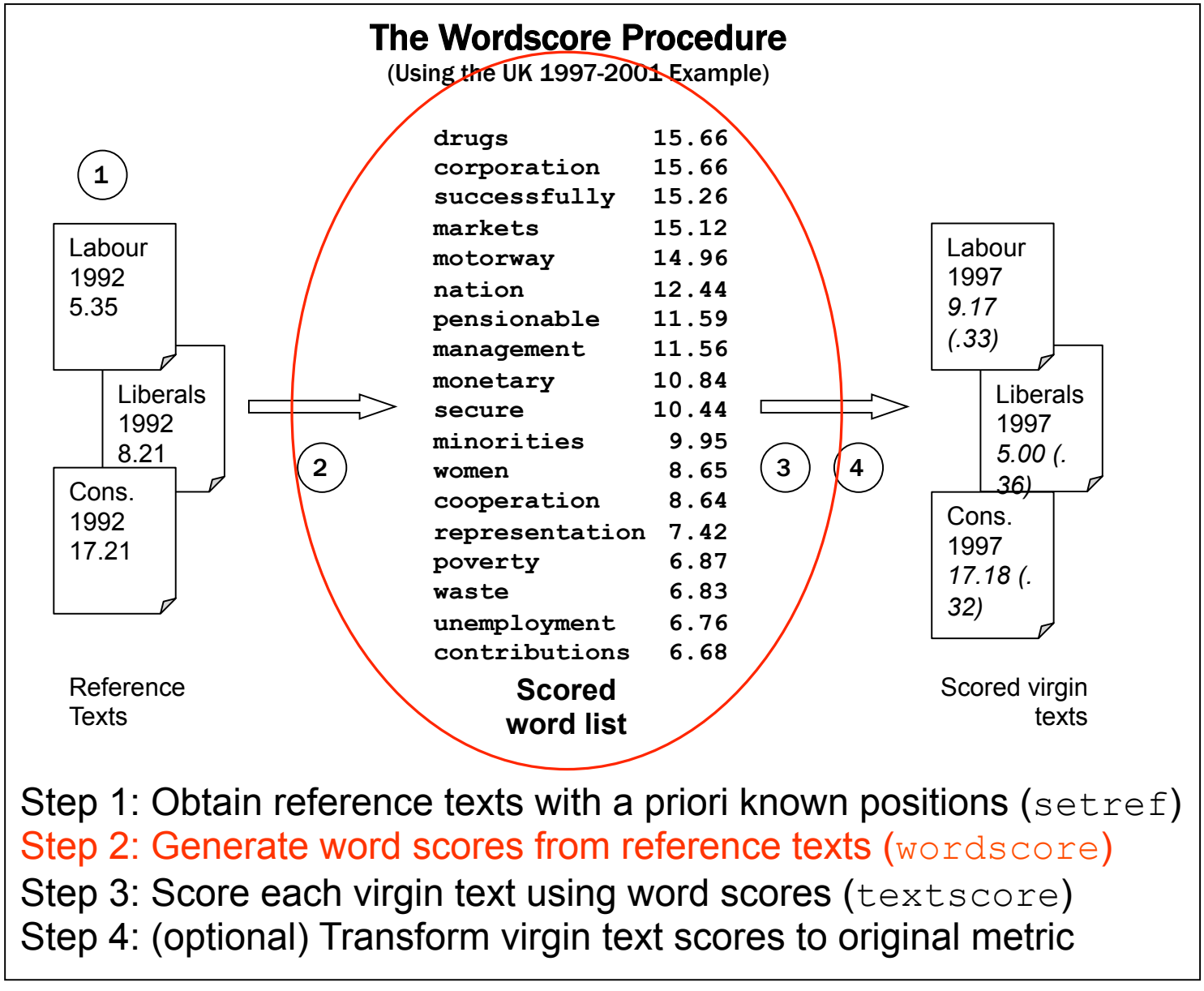
Step 1: Obtain reference texts with a priori known positions (`setref`)

Step 2: Generate word scores from reference texts (`wordscore`)

Step 3: Score each virgin text using word scores (`textscore`)

Step 4: (optional) Transform virgin text scores to original metric

Wordscores Procedure



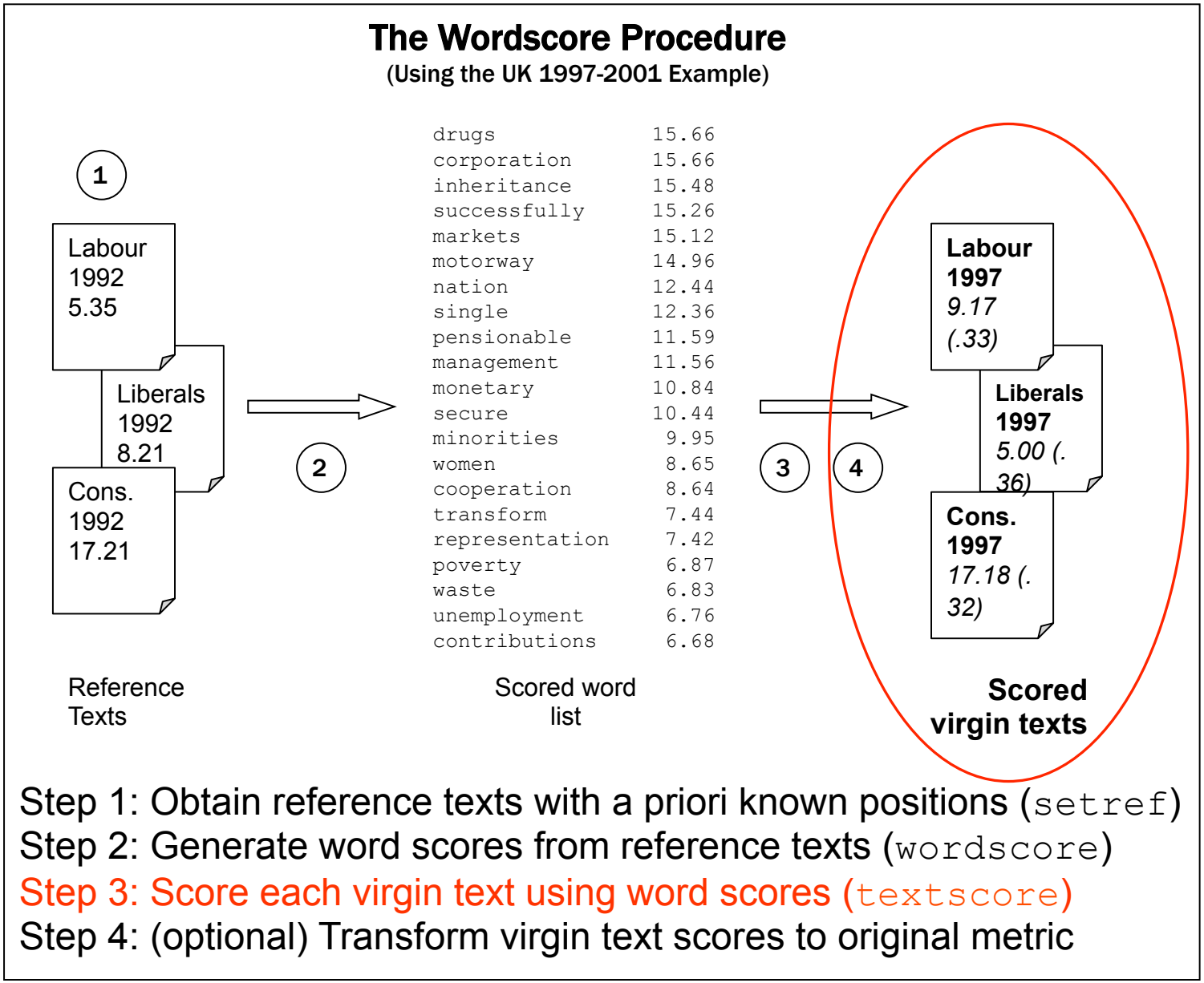
Step 1: Obtain reference texts with a priori known positions (`setref`)

Step 2: Generate word scores from reference texts (`wordscore`)

Step 3: Score each virgin text using word scores (`textscore`)

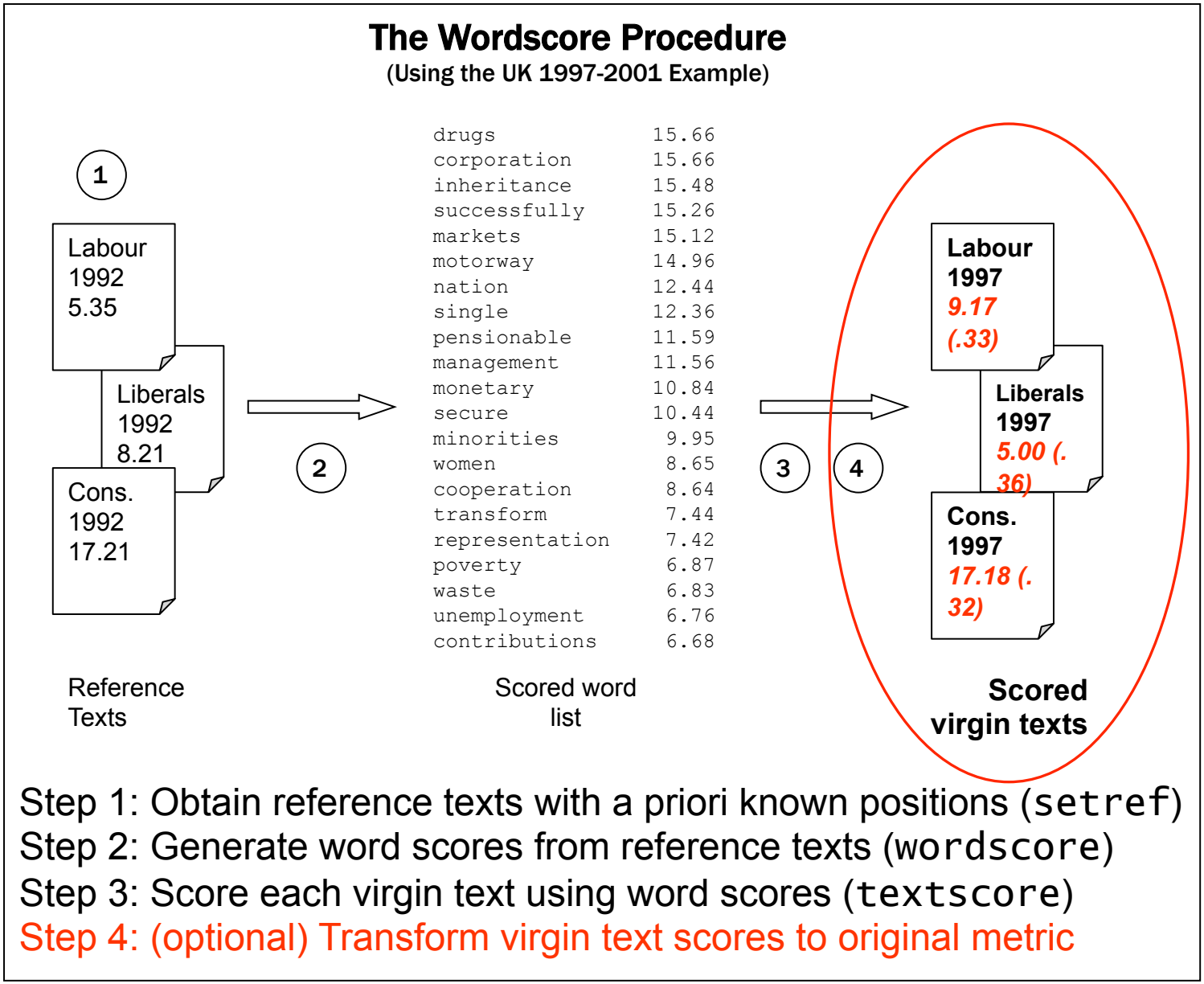
Step 4: (optional) Transform virgin text scores to original metric

Wordscores Procedure



Step 1: Obtain reference texts with a priori known positions (`setref`)
Step 2: Generate word scores from reference texts (`wordscore`)
Step 3: Score each virgin text using word scores (`textscore`)
Step 4: (optional) Transform virgin text scores to original metric

Wordscores Procedure



Step 1: Obtain reference texts with a priori known positions (`setref`)
Step 2: Generate word scores from reference texts (`wordscore`)
Step 3: Score each virgin text using word scores (`textscore`)
Step 4: (optional) Transform virgin text scores to original metric

Wordscores mathematically: Reference texts

- Start with a set of I *reference* texts, represented by an $I \times J$ document-term frequency matrix C_{ij}, where i indexes the document and j indexes the J total word types

Wordscores mathematically: Reference texts

- Start with a set of I *reference* texts, represented by an $I \times J$ document-term frequency matrix C_{ij}, where i indexes the document and j indexes the J total word types

Wordscores mathematically: Reference texts

- Start with a set of I *reference* texts, represented by an $I \times J$ document-term frequency matrix C_{ij}, where i indexes the document and j indexes the J total word types
- Each text will have an associated "score" a_i, which is a single number locating this text on a single dimension of difference

Wordscores mathematically: Reference texts

- Start with a set of I *reference* texts, represented by an $I \times J$ document-term frequency matrix C_{ij}, where i indexes the document and j indexes the J total word types
- Each text will have an associated "score" a_i, which is a single number locating this text on a single dimension of difference
  - This can be on a scale metric, such as 1–20
  - Can use arbitrary endpoints, such as -1, 1

Wordscores mathematically: Reference texts

- Start with a set of I *reference* texts, represented by an $I \times J$ document-term frequency matrix C_{ij}, where i indexes the document and j indexes the J total word types
- Each text will have an associated "score" a_i, which is a single number locating this text on a single dimension of difference
  - This can be on a scale metric, such as 1–20
  - Can use arbitrary endpoints, such as -1, 1
- We *normalize* the document-term frequency matrix within each document by converting C_{ij} into a *relative* document-term frequency matrix (within document), by dividing C_{ij} by its word total marginals:

$$F_{ij} = \frac{C_{ij}}{C_{i\cdot}} \tag{1}$$

where $C_{i\cdot} = \sum_{j=1}^{J} C_{ij}$

Wordscores mathematically: Word scores

Wordscores mathematically: Word scores

- Compute an $I \times J$ matrix of relative document probabilities P_{ij} for each word in each reference text, as

$$P_{ij} = \frac{F_{ij}}{\sum_{i=1}^{I} F_{ij}} \tag{2}$$

Wordscores mathematically: Word scores

- Compute an $I \times J$ matrix of relative document probabilities P_{ij} for each word in each reference text, as

$$P_{ij} = \frac{F_{ij}}{\sum_{i=1}^{I} F_{ij}} \tag{2}$$

- This tells us the probability that given the observation of a specific word j, that we are reading a text of a certain reference document i

Wordscores mathematically: Word scores (example)

- Assume we have two reference texts, A and B
- The word "choice" is used 10 times per 1,000 words in Text A and 30 times per 1,000 words in Text B
- So $F_{i\ \text{"choice"}} = \{.010, .030\}$
- If we know only that we are reading the word choice in one of the two reference texts, then probability is 0.25 that we are reading Text A, and 0.75 that we are reading Text B

$$P_{A\ \text{"choice"}} = \frac{.010}{(.010 + .030)} = 0.25 \tag{3}$$

$$P_{B\ \text{"choice"}} = \frac{.030}{(.010 + .030)} = 0.75 \tag{4}$$

Wordscores mathematically: Word scores

Wordscores mathematically: Word scores

- Compute a J-length "score" vector S for each word j as the average of each document i's scores a_i, weighted by each word's P_{ij}:

$$S_j = \sum_{i=1}^{I} a_i P_{ij} \tag{5}$$

Wordscores mathematically: Word scores

- Compute a J-length "score" vector S for each word j as the average of each document i's scores a_i, weighted by each word's P_{ij}:

$$S_j = \sum_{i=1}^{I} a_i P_{ij} \tag{5}$$

- In matrix algebra, $\underset{1\times J}{S} = \underset{1\times I}{a} \cdot \underset{I\times J}{P}$
- This procedure will yield a single "score" for every word that reflects the balance of the scores of the reference documents, weighted by the relative document frequency of its normalized term frequency

Wordscores mathematically: Word scores

Wordscores mathematically: Word scores

- Continuing with our example:
  - We "know" (from independent sources) that Reference Text A has a position of -1.0, and Reference Text B has a position of $+1.0$
  - The score of the word choice is then $0.25(-1.0) + 0.75(1.0) = -0.25 + 0.75 = +0.50$

Wordscores mathematically: Scoring "virgin" texts

- Here the objective is to obtain a single score for any new text, relative to the reference texts

Wordscores mathematically: Scoring "virgin" texts

- Here the objective is to obtain a single score for any new text, relative to the reference texts
- We do this by taking the mean of the scores of its words, weighted by their term frequency

Wordscores mathematically: Scoring "virgin" texts

- Here the objective is to obtain a single score for any new text, relative to the reference texts
- We do this by taking the mean of the scores of its words, weighted by their term frequency
- So the score v_k of a virgin document k consisting of the j word types is:

$$v_k = \sum_j (F_{kj} \cdot s_j) \tag{6}$$

where $F_{kj} = \frac{C_{kj}}{C_{k.}}$ as in the reference document relative word frequencies

Wordscores mathematically: Scoring "virgin" texts

- Here the objective is to obtain a single score for any new text, relative to the reference texts
- We do this by taking the mean of the scores of its words, weighted by their term frequency
- So the score v_k of a virgin document k consisting of the j word types is:

$$v_k = \sum_j (F_{kj} \cdot s_j) \tag{6}$$

where $F_{kj} = \frac{C_{kj}}{C_{k\cdot}}$ as in the reference document relative word frequencies

- Note that new words outside of the set J may appear in the K virgin documents — these are simply ignored (because we have no information on their scores)

Wordscores mathematically: Scoring "virgin" texts

- Here the objective is to obtain a single score for any new text, relative to the reference texts
- We do this by taking the mean of the scores of its words, weighted by their term frequency
- So the score v_k of a virgin document k consisting of the j word types is:

$$v_k = \sum_j (F_{kj} \cdot s_j) \tag{6}$$

where $F_{kj} = \frac{C_{kj}}{C_{k.}}$ as in the reference document relative word frequencies

- Note that new words outside of the set J may appear in the K virgin documents — these are simply ignored (because we have no information on their scores)
- Note also that nothing prohibits reference documents from also being scored as virgin documents

Wordscores mathematically: Rescaling raw text scores

- Because of overlapping or non-discriminating words, the raw text scores will be dragged to the interior of the reference scores (we will see this shortly in the results)

Wordscores mathematically: Rescaling raw text scores

- Because of overlapping or non-discriminating words, the raw text scores will be dragged to the interior of the reference scores (we will see this shortly in the results)
- Some procedures can be applied to rescale them, either to a unit normal metric or to a more "natural" metric

Wordscores mathematically: Rescaling raw text scores

- Because of overlapping or non-discriminating words, the raw text scores will be dragged to the interior of the reference scores (we will see this shortly in the results)
- Some procedures can be applied to rescale them, either to a unit normal metric or to a more "natural" metric
- Martin and Vanberg (2008) have proposed alternatives to the LBG (2003) rescaling

Computing confidence intervals

- The score v_k of any text represents a weighted mean

Computing confidence intervals

- The score v_k of any text represents a weighted mean
- LBG (2003) used this logic to develop a standard error of this mean using a *weighted variance* of the scores in the virgin text

Computing confidence intervals

- The score v_k of any text represents a weighted mean
- LBG (2003) used this logic to develop a standard error of this mean using a *weighted variance* of the scores in the virgin text
- Given some assumptions about the scores being fixed (and the words being conditionally independent), this yields approximately normally distributed errors for each v_k

Computing confidence intervals

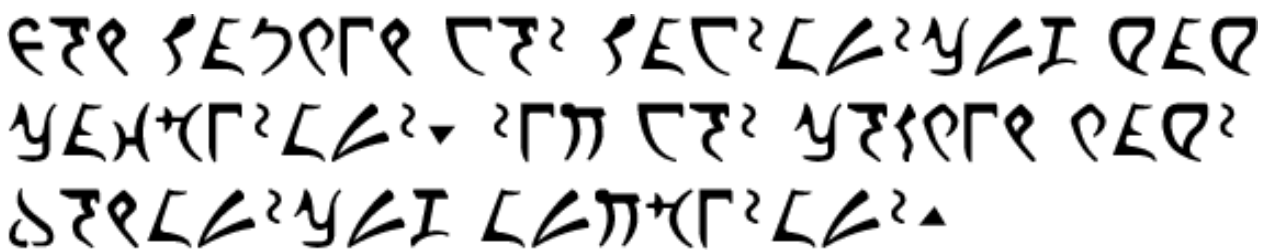
- The score v_k of any text represents a weighted mean
- LBG (2003) used this logic to develop a standard error of this mean using a *weighted variance* of the scores in the virgin text
- Given some assumptions about the scores being fixed (and the words being conditionally independent), this yields approximately normally distributed errors for each v_k
- An alternative would be to bootstrap the textual data prior to constructing C_{ij} and C_{kj} — see Lowe and Benoit (2012)

Pros and Cons of the Wordscores approach

- Fully automated technique with minimal human intervention or judgment calls – only with regard to reference text selection
- Language-blind: all we need to know are reference scores

Pros and Cons of the Wordscores approach

- Fully automated technique with minimal human intervention or judgment calls – only with regard to reference text selection
- Language-blind: all we need to know are reference scores
- Could potentially work on texts like this:



(See `http://www.kli.org`)

Pros and Cons of the Wordscores approach

- Estimates unknown positions on a priori scales – hence no inductive scaling with a posteriori interpretation of unknown policy space

Pros and Cons of the Wordscores approach

- Estimates unknown positions on a priori scales – hence no inductive scaling with a posteriori interpretation of unknown policy space
- Very dependent on correct identification of:
  - appropriate reference texts
  - appropriate reference scores

Suggestions for choosing reference texts

- Texts need to contain information representing a clearly dimensional position

Suggestions for choosing reference texts

- Texts need to contain information representing a clearly dimensional position
- Dimension must be known a priori. Sources might include:
  - Survey scores or manifesto scores
  - Arbitrarily defined scales (e.g. -1.0 and 1.0)

Suggestions for choosing reference texts

- Texts need to contain information representing a clearly dimensional position
- Dimension must be known a priori. Sources might include:
  - Survey scores or manifesto scores
  - Arbitrarily defined scales (e.g. -1.0 and 1.0)
- Should be as discriminating as possible: extreme texts on the dimension of interest, to provide reference anchors

Suggestions for choosing reference texts

- Texts need to contain information representing a clearly dimensional position
- Dimension must be known a priori. Sources might include:
  - Survey scores or manifesto scores
  - Arbitrarily defined scales (e.g. -1.0 and 1.0)
- Should be as discriminating as possible: extreme texts on the dimension of interest, to provide reference anchors
- Need to be from the same lexical universe as virgin texts

Suggestions for choosing reference texts

- Texts need to contain information representing a clearly dimensional position
- Dimension must be known a priori. Sources might include:
  - Survey scores or manifesto scores
  - Arbitrarily defined scales (e.g. -1.0 and 1.0)
- Should be as discriminating as possible: extreme texts on the dimension of interest, to provide reference anchors
- Need to be from the same lexical universe as virgin texts
- Should contain lots of words

Suggestions for choosing reference values

- Must be “known” through some trusted external source

Suggestions for choosing reference values

- Must be “known” through some trusted external source
- For any pair of reference values, all scores are simply linear rescalings, so might as well use (-1, 1)

Suggestions for choosing reference values

- Must be "known" through some trusted external source
- For any pair of reference values, all scores are simply linear rescalings, so might as well use (-1, 1)
- The "middle point" will not be the midpoint, however, since this will depend on the relative word frequency of the reference documents

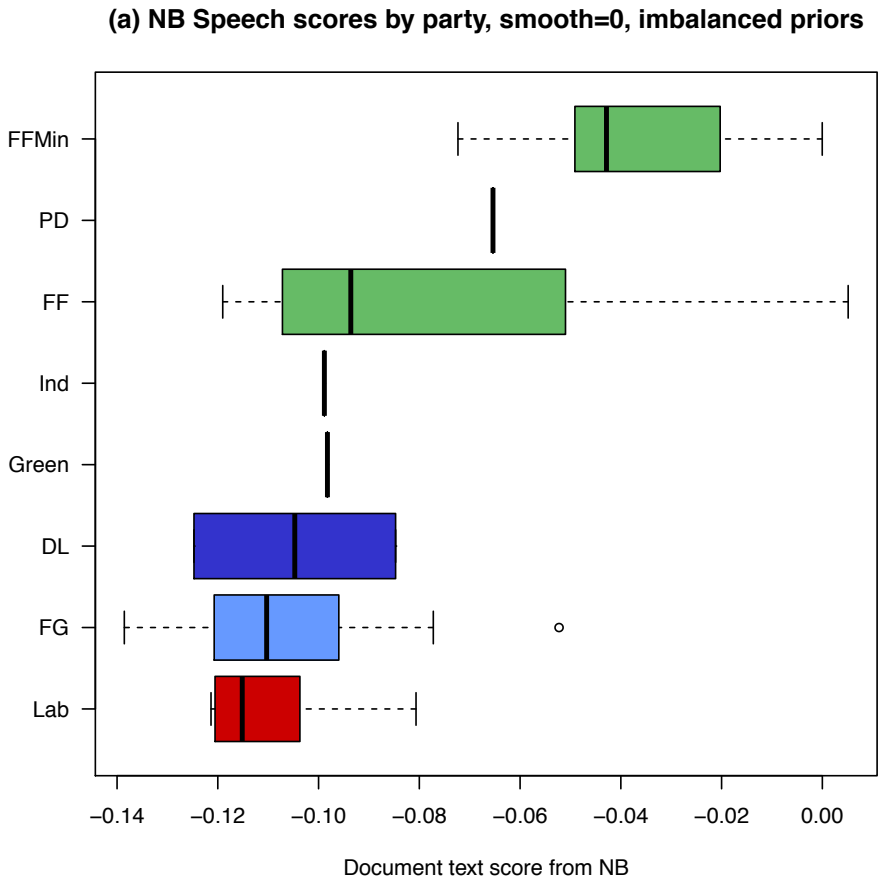
Suggestions for choosing reference values

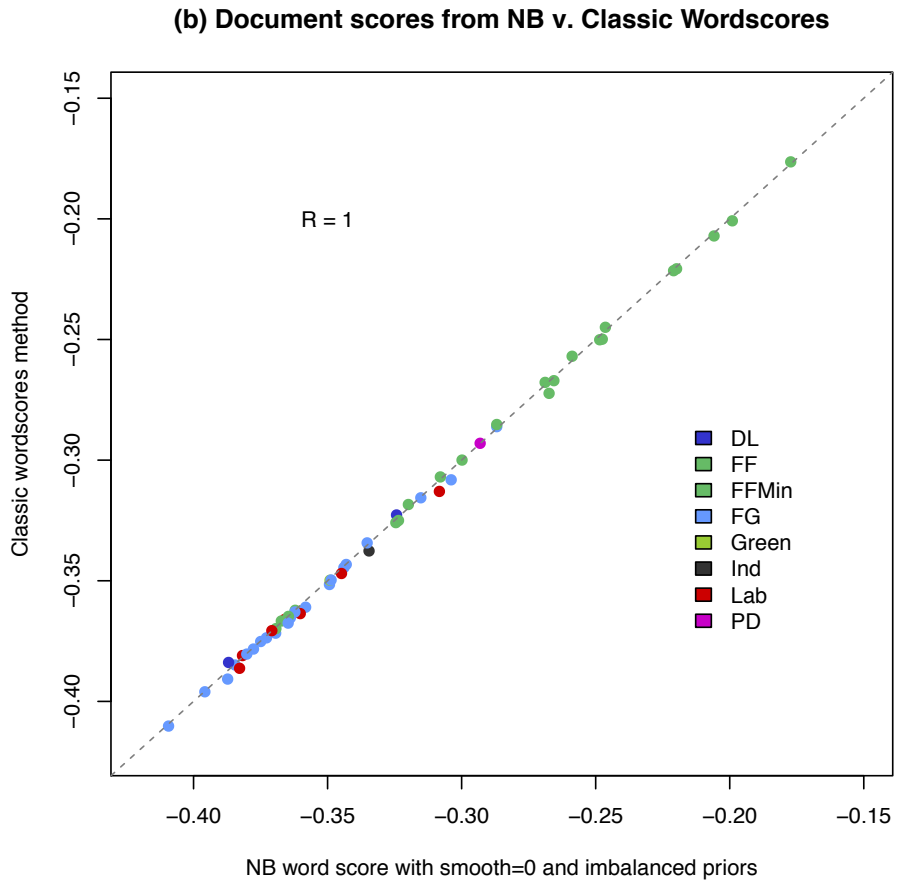
- Must be "known" through some trusted external source
- For any pair of reference values, all scores are simply linear rescalings, so might as well use (-1, 1)
- The "middle point" will not be the midpoint, however, since this will depend on the relative word frequency of the reference documents
- Reference texts if scored as virgin texts will have document scores more extreme than other virgin texts

Suggestions for choosing reference values

- Must be "known" through some trusted external source
- For any pair of reference values, all scores are simply linear rescalings, so might as well use (-1, 1)
- The "middle point" will not be the midpoint, however, since this will depend on the relative word frequency of the reference documents
- Reference texts if scored as virgin texts will have document scores more extreme than other virgin texts
- With three or more reference values, the mid-point is mapped onto a multi-dimensional simplex. The values now matter but only in relative terms (we are still investigating this fully)

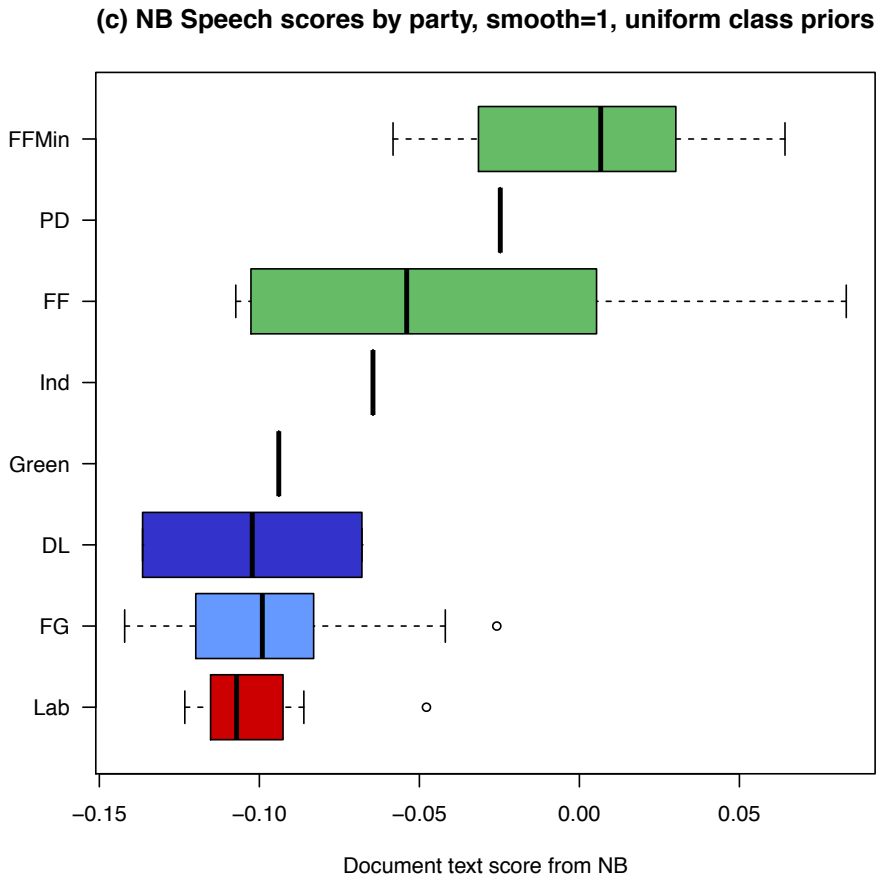
Application 1: Dail speeches from LBG (2003)

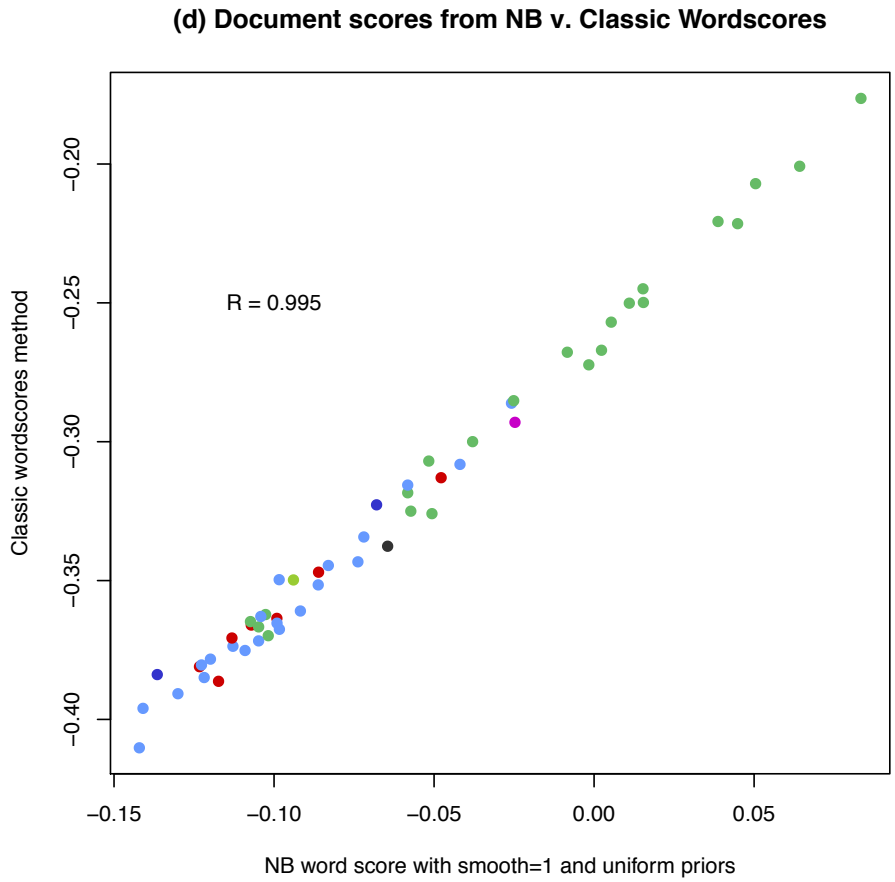




- three reference classes (Opposition, Opposition, Government) at {-1, -1, 1}
- no smoothing

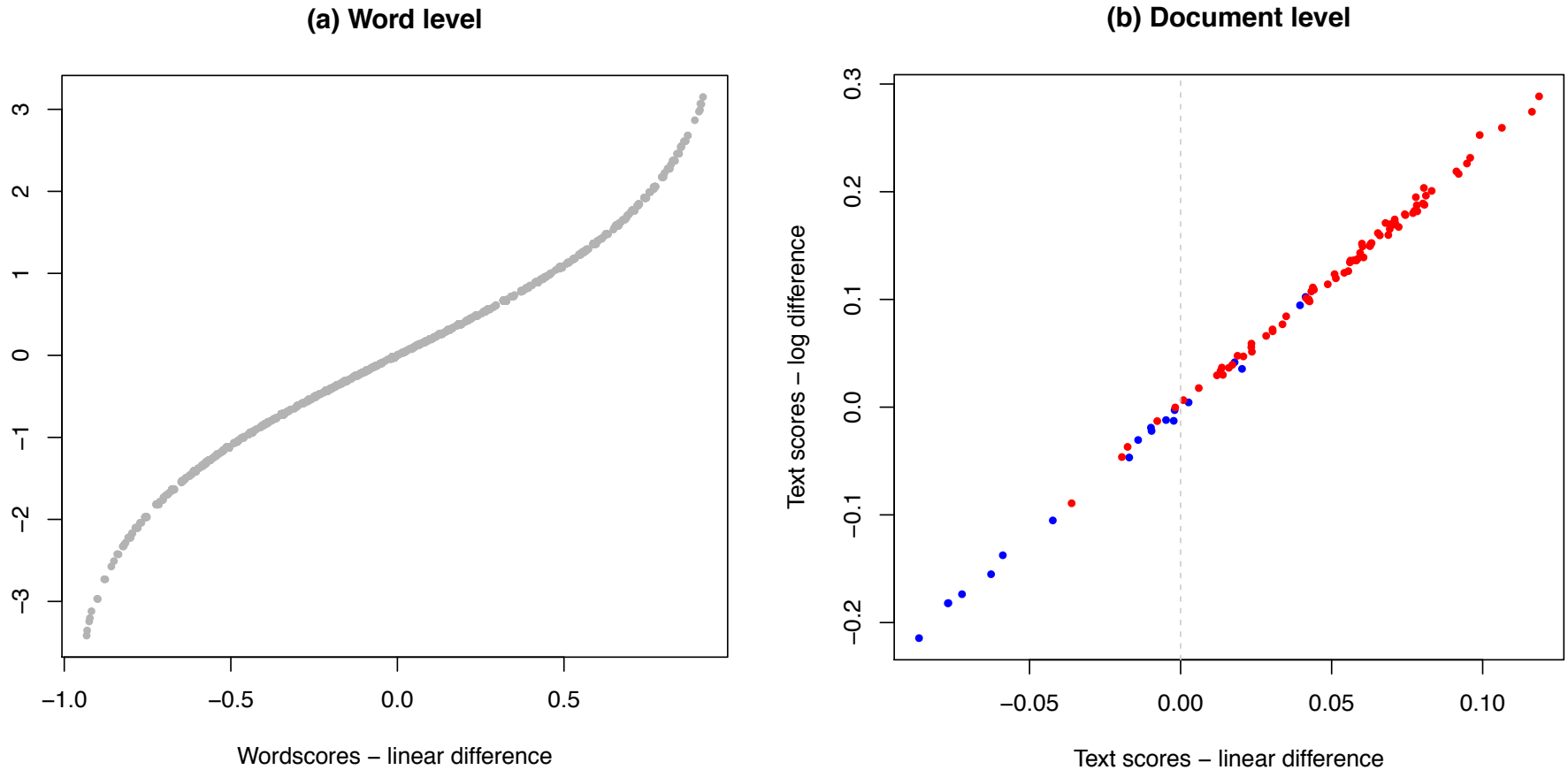
Application 1: Dail speeches from LBG (2003)





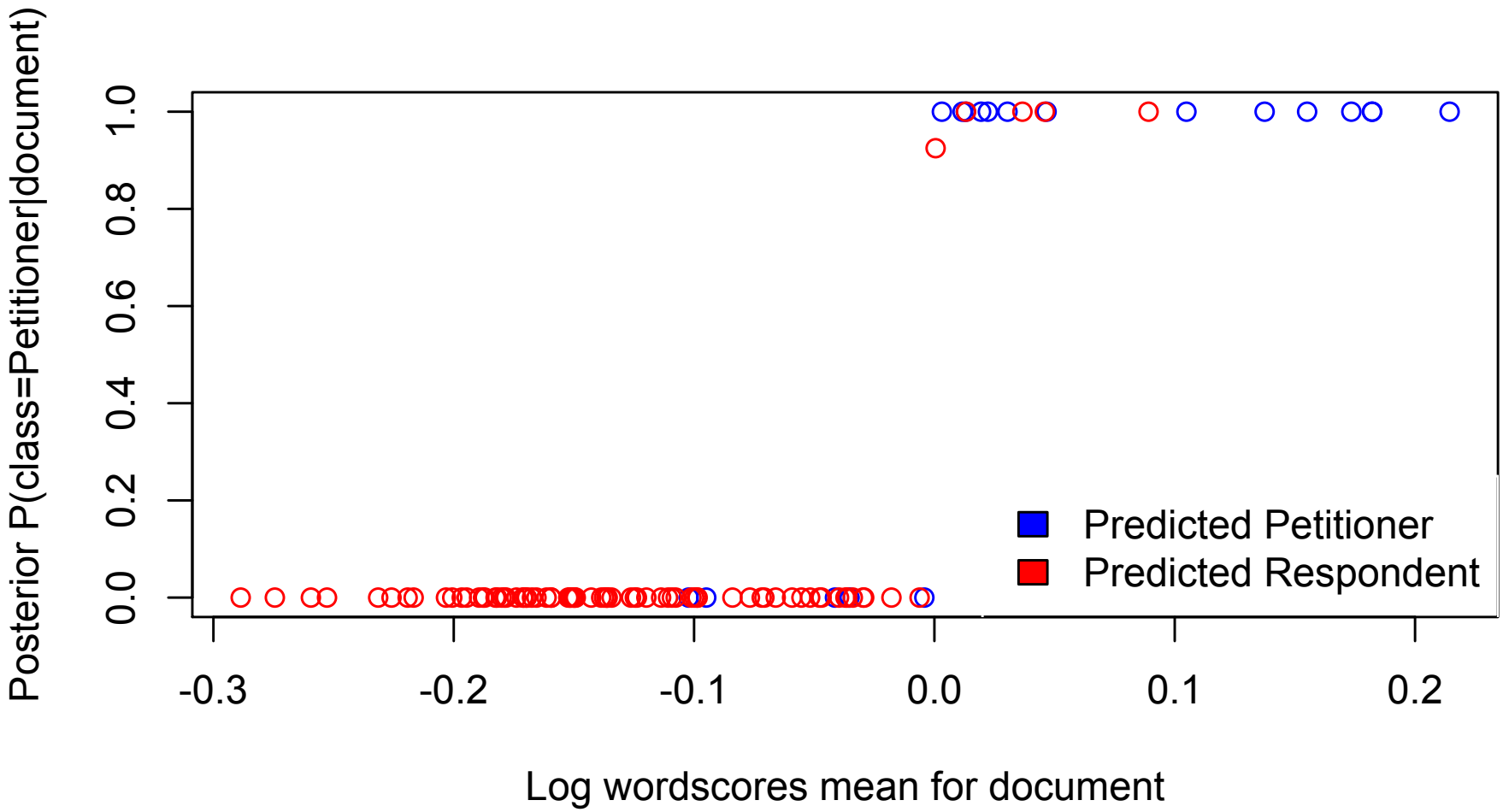
- two reference classes (Opposition+Opposition, Government) at {-1, 1}
- Laplace smoothing

Application 2: Classifying legal briefs (Evans et al 2007) Wordscores v. Bayesscore

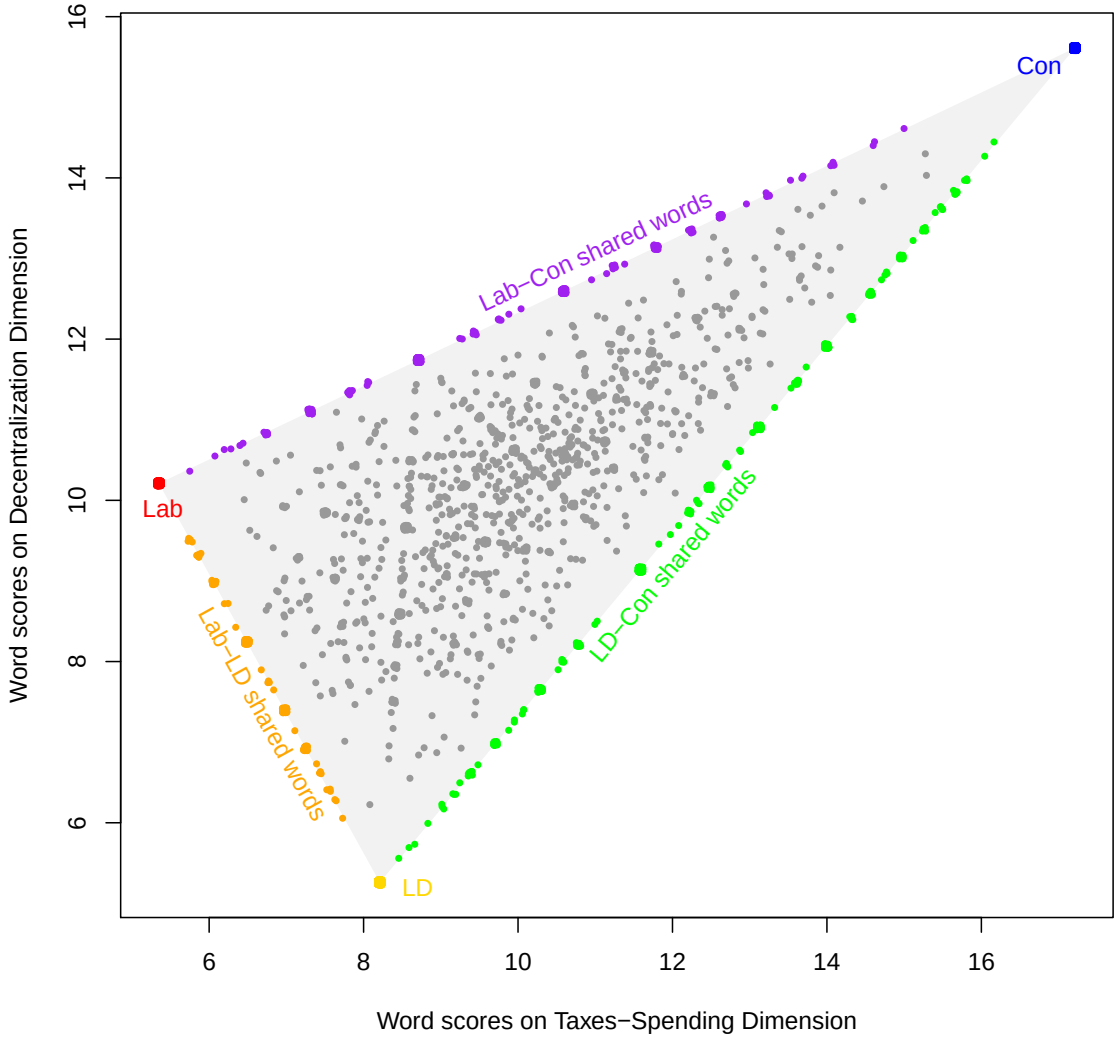


- Training set: Petitioner and Respondent litigant briefs from *Grutter/Gratz v. Bollinger* (a U.S. Supreme Court case)
- Test set: 98 amicus curiae briefs (whose P or R class is known)

Application 2: Classifying legal briefs (Evans et al 2007)
Posterior class prediction from NB versus log wordscores



Application 3: LBG's British manifestos
More than two reference classes



- x-axis: Reference scores of $\{5.35, 8.21, 17.21\}$ for Lab, LD, Conservatives
- y-axis: Reference scores of $\{10.21, 5.26, 15.61\}$

Unsupervised methods scale distance

- Text gets converted into a quantitative matrix of features
  - words, typically
  - could be dictionary entries, or parts of speech
- Language is irrelevant
- Different possible definitions of *distance*
  - see for instance `summary(pr_DB)` from `proxy` library
- Works on any quantitative matrix of features

Parametric v. non-parametric methods

- Parametric methods model feature occurrence according to some stochastic distribution, typically in the form of a measurement model
  - for instance, model words as a multi-level Bernoulli distribution, or a Poisson distribution
  - word effects and "positional" effects are unobserved parameters to be estimated
- Non-parametric methods typically based on the Singular Value Decomposition of a matrix
  - correspondence analysis
  - factor analysis
  - other (multi)dimensional scaling methods

Parametic scaling model: Model counts as Poisson

- Many dependent variables of interest may be in the form of counts of discrete events

Parametic scaling model: Model counts as Poisson

- Many dependent variables of interest may be in the form of counts of discrete events— examples:
  - international wars or conflict events
  - traffic incidents
  - deaths

Parametic scaling model: Model counts as Poisson

- Many dependent variables of interest may be in the form of counts of discrete events— examples:
  - international wars or conflict events
  - traffic incidents
  - deaths
  - word count given an underlying orientation

Parametic scaling model: Model counts as Poisson

- Many dependent variables of interest may be in the form of counts of discrete events— examples:
  - international wars or conflict events
  - traffic incidents
  - deaths
  - word count given an underlying orientation
- Characteristics: these Y are bounded between $(0, \infty)$ and take on only discrete values $0, 1, 2, \ldots, \infty$

Parametic scaling model: Model counts as Poisson

- Many dependent variables of interest may be in the form of counts of discrete events— examples:
  - international wars or conflict events
  - traffic incidents
  - deaths
  - word count given an underlying orientation
- Characteristics: these Y are bounded between $(0, \infty)$ and take on only discrete values $0, 1, 2, \ldots, \infty$
- Imagine a social system that produces events randomly during a fixed period, and at the end of this period only the total count is observed. For N periods, we have $y_1, y_2, \ldots, y_N$ observed counts

Poisson data model first principles

Poisson data model first principles

1. The probability that two events occur at precisely the same time is zero

Poisson data model first principles

1. The probability that two events occur at precisely the same time is zero
2. During each period i, the event rate occurence λ_i remains constant and is independent of all previous events during the period
   - note that this implies no *contagion* effects
   - also known as *Markov independence*

Poisson data model first principles

1. The probability that two events occur at precisely the same time is zero
2. During each period i, the event rate occurence λ_i remains constant and is independent of all previous events during the period
   - note that this implies no *contagion* effects
   - also known as *Markov independence*
3. Zero events are recorded at the start of the period

Poisson data model first principles

1. The probability that two events occur at precisely the same time is zero
2. During each period i, the event rate occurence λ_i remains constant and is independent of all previous events during the period
   - note that this implies no *contagion* effects
   - also known as *Markov independence*
3. Zero events are recorded at the start of the period
4. All observation intervals are equal over i

The Poisson distribution

$$f_{Poisson}(y_i|\lambda) = \begin{cases} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} & \forall\ \lambda > 0 \text{ and } y_i = 0, 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

The Poisson distribution

$$f_{Poisson}(y_i|\lambda) = \begin{cases} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} & \forall\ \lambda > 0 \text{ and } y_i = 0, 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

$$Pr(Y|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$$

The Poisson distribution

$$
\begin{aligned}
f_{Poisson}(y_i|\lambda) &= \begin{cases} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} & \forall\ \lambda > 0 \text{ and } y_i = 0, 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases} \\
Pr(Y|\lambda) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} \\
\lambda &= e^{X_i\beta}
\end{aligned}
$$

The Poisson distribution

$$
\begin{aligned}
f_{Poisson}(y_i|\lambda) &= \begin{cases} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} & \forall\ \lambda > 0 \text{ and } y_i = 0, 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases} \\
Pr(Y|\lambda) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} \\
\lambda &= e^{X_i\beta} \\
\mathrm{E}(y_i) &= \lambda
\end{aligned}
$$

The Poisson distribution

$$
\begin{aligned}
f_{Poisson}(y_i|\lambda) &= \begin{cases} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} & \forall\ \lambda > 0 \text{ and } y_i = 0, 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases} \\
Pr(Y|\lambda) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} \\
\lambda &= e^{X_i\beta} \\
\mathrm{E}(y_i) &= \lambda \\
\mathrm{Var}(y_i) &= \lambda
\end{aligned}
$$

Systematic component

- $\lambda_i > 0$ is only bounded from below (unlike π_i)

Systematic component

- $\lambda_i > 0$ is only bounded from below (unlike π_i)
- This implies that the effect cannot be linear

Systematic component

- $\lambda_i > 0$ is only bounded from below (unlike π_i)
- This implies that the effect cannot be linear
- Hence for the functional form we will use an exponential transformation

$$\mathrm{E}(Y_i) = \lambda_i = e^{x_i \beta}$$

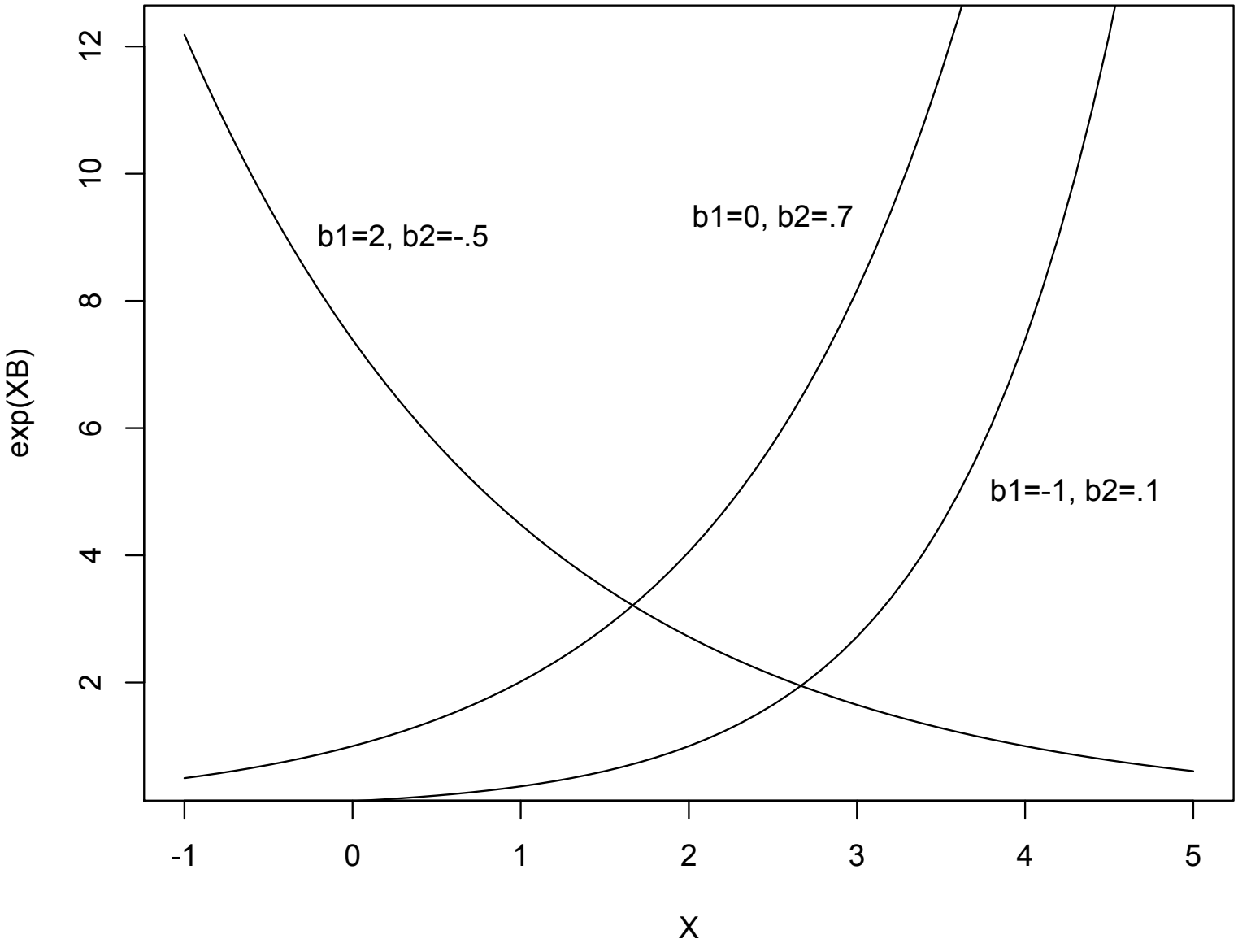
Systematic component

- $\lambda_i > 0$ is only bounded from below (unlike π_i)
- This implies that the effect cannot be linear
- Hence for the functional form we will use an exponential transformation

$$\mathrm{E}(Y_i) = \lambda_i = e^{x_i \beta}$$

- Other possibilities exist, but this is by far the most common – indeed almost universally used – functional form for event count models

Exponential link function



The Poisson scaling "wordfish" model

Data:

- Y is N (speaker) $\times$ V (word) term document matrix $V \gg N$

Model:

$$
\begin{aligned}
P(Y_i \mid \theta) &= \prod_{j=1}^{V} P(Y_{ij} \mid \theta_i) \\
Y_{ij} &\sim \text{Poisson}(\lambda_{ij}) \\
\log \lambda_{ij} &= \alpha_i + \theta_i \beta_j + \psi_j
\end{aligned}
\tag{7}
$$

Estimation:

- Easy to fit for large V (V Poisson regressions with α offsets)

Model components and notation

$$\log \lambda_{ij} = \alpha_i + \theta_i \beta_j + \psi_j$$

Element	*Meaning*
i	indexes documents
j	indexes word types
θ_i	the unobservable "position" of document i
β_j	word parameters on θ – the relationship of word j to document position
ψ_j	word "fixed effect" (function of the frequency of word j)
α_i	document "fixed effects" (a function of (log) document length to allow estimation in Poisson of an essentially multinomial process)

"Features" of the parametric scaling approach

- Standard (statistical) inference about parameters
- Uncertainty accounting for parameters
- Distributional assumptions are made explicit (as part of the data generating process motivating the choice of stochastic distribution)
  - *conditional independence*
  - *stochastic process* (e.g. $\mathrm{E}(Y_{ij}) = \mathrm{Var}(Y_{ij}) = \lambda_{ij}$)
- Permits hierarchical reparameterization (to add covariates)
- Generative model: given the estimated parameters, we could generate a document for any specified length

Some reasons why this model is wrong

- Words occur in order (unless you are Yoda: "No more training do you require. Already know you that which you need.")

Some reasons why this model is wrong

- Words occur in order (unless you are Yoda: "No more training do you require. Already know you that which you need.")
- Words occur in combinations (as collocations) "carbon tax" / "income tax" / "inhertiance tax" / "capital gains tax" /"bank tax"

Some reasons why this model is wrong

- Words occur in order (unless you are Yoda: "No more training do you require. Already know you that which you need.")
- Words occur in combinations (as collocations) "carbon tax" / "income tax" / "inhertiance tax" / "capital gains tax" /"bank tax"
- Sentences (and topics) occur in sequence (extreme serial correlation)
- Style may mean means we are likely to use synonyms – very probable.

Some reasons why this model is wrong

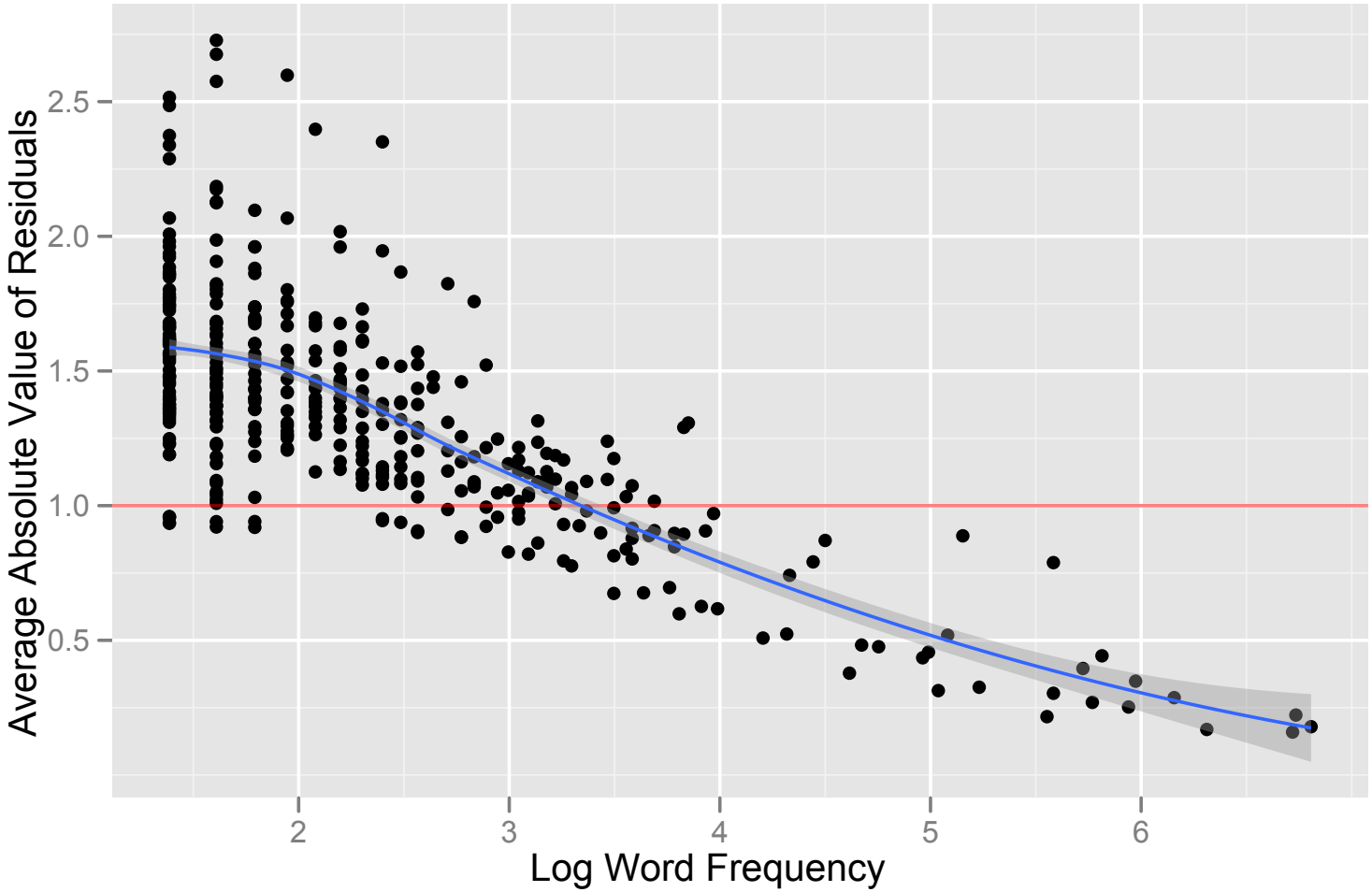
- Words occur in order (unless you are Yoda: "No more training do you require. Already know you that which you need.")
- Words occur in combinations (as collocations) "carbon tax" / "income tax" / "inhertiance tax" / "capital gains tax" /"bank tax"
- Sentences (and topics) occur in sequence (extreme serial correlation)
- Style may mean means we are likely to use synonyms – very probable. In fact it's very distinctly possible, to be expected, odds-on, plausible, imaginable; expected, anticipated, predictable, predicted, foreseeable.)
- Rhetoric may lead to repetition. ("Yes we can!") – anaphora

Assumptions of the model (cont.)

- Poisson assumes $\text{Var}(Y_{ij}) = \text{E}(Y_{ij}) = \lambda_{ij}$
- For many reasons, we are likely to encounter overdispersion or underdispersion
  - overdispersion when "informative" words tend to cluster together
  - underdispersion could (possibly) occur when words of high frequency are uninformative and have relatively low between-text variation (once length is considered)
- This should be a *word*-level parameter

Overdispersion in German manifesto data

(data taken from Slapin and Proksch 2008)



One solution: Model overdispersion

Lo, Proksch, and Slapin:

$$\text{Poisson}(\lambda) = \lim_{r \to \infty} \text{NB}\left(r, \frac{\lambda}{\lambda + r}\right)$$

$$Y_{ij} \sim \text{NB}\left(r, \frac{\lambda_{ij}}{\lambda_{ij} + r}\right)$$

where the variance inflation parameter r varies across *documents*:

$$Y_{ij} \sim \text{NB}\left(r_i, \frac{\lambda_{ij}}{\lambda_{ij} + r_i}\right)$$

Relationship to multinomial

If each feature count Y_{ij} is an independent Poisson random variable with mean μ_{ij}, then we can formulate this as the following log-linear model:

$$\log \mu_{ij} = \lambda + \alpha_i + \psi_j^* + \theta_i \beta_j^* \tag{8}$$

where the log-odds that a generated token will fall into feature category j relative to the last feature J is:

$$\log \frac{\mu_{ij}}{\mu_{iJ}} = (\psi_j^* - \psi_J^*) + \theta_i(\beta_j^* - \beta_J^*) \tag{9}$$

which is the formula for multinomial logistic

Current project: exploring relationship to 2PL IRT

Specifically, "wordfish" appears to be a version of Bock's (1972) *nominal response model*

Cf. Benoit and Däubler (in progress)

Poisson/multinomial process as a DAG

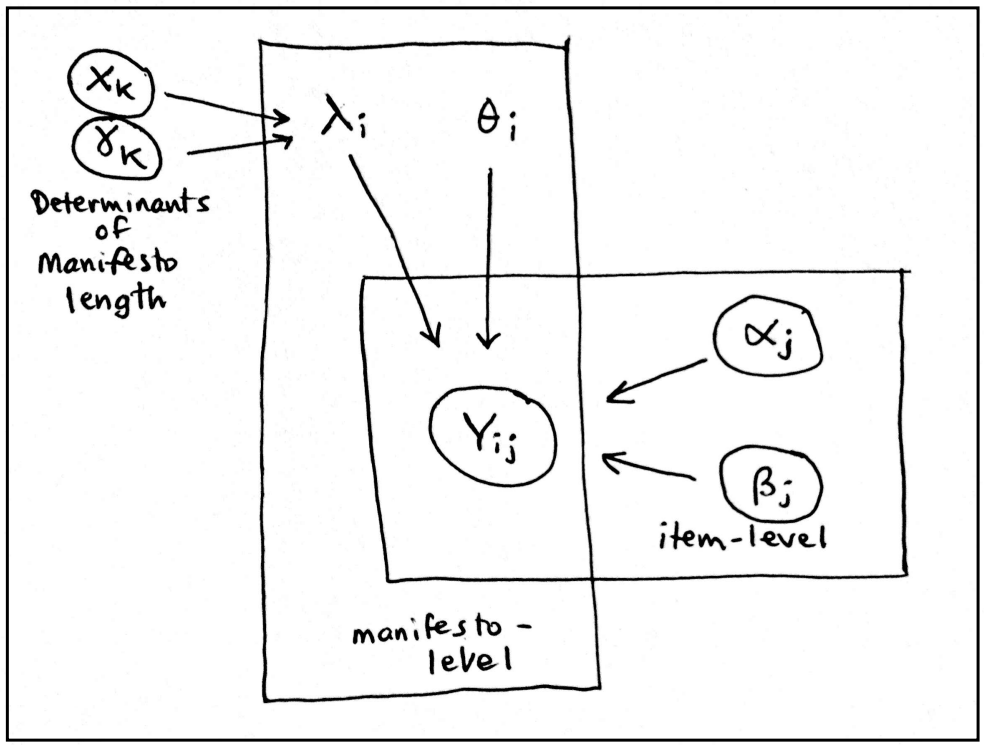


Figure 2: Directed acyclic graph of the one-dimensional Poisson IRT for document and item parameters to category counts Y_{ij}

How to estimate this model

Iterative maximimum likelihood estimation:

- If we knew Ψ and β (the word parameters) then we have a Poisson regression model
- If we knew α and θ (the party / politician / document parameters) then we have a Poisson regression model too!
- So we alternate them and hope to converge to reasonable estimates for both
- Implemented in the `austin` package as `wordfish`

An alternative is MCMC with a Bayesian formulation

Marginal maximum likelihood for wordfish

Start by guessing the parameters
Algorithm:

- Assume the current party parameters are correct and fit as a Poisson regression model
- Assume the current word parameters are correct and fit as a Poisson regression model
- Normalize θs to mean 0 and variance 1

Repeat

Identification

The *scale* and *direction* of θ is undetermined — like most models with latent variables

To identify the model in Wordfish

- Fix one α to zero to specify the left-right direction (Wordfish option 1)
- Fix the $\hat{\theta}$s to mean 0 and variance 1 to specify the scale (Wordfish option 2)
- Fix two $\hat{\theta}$s to specify the direction and scale (Wordfish option 3 and Wordscores)

Note: Fixing two reference scores does not specify the policy domain, it just identifies the model

Or: Use non-parametric methods

- Non-parametric methods are algorithmic, involving no "parameters" in the procedure that are estimated
- Hence there is no uncertainty accounting given distributional theory
- Advantage: don't have to make assumptions
- Disadvantages:
  - cannot leverage probability conclusions given distribtional assumptions and statistical theory
  - results highly fit to the data
  - not really assumption-free, if we are honest

Correspondence Analysis

- CA is like factor analysis for categorical data
- Following normalization of the marginals, it uses Singular Value Decomposition to reduce the dimensionality of the word-by-text matrix
- This allows projection of the positioning of the words as well as the texts into multi-dimensional space
- The number of dimensions – as in factor analysis – can be decided based on the eigenvalues from the SVD

Singular Value Decomposition

- A matrix $\underset{i \times j}{\mathbf{X}}$ can be represented in a dimensionality equal to its rank k as:

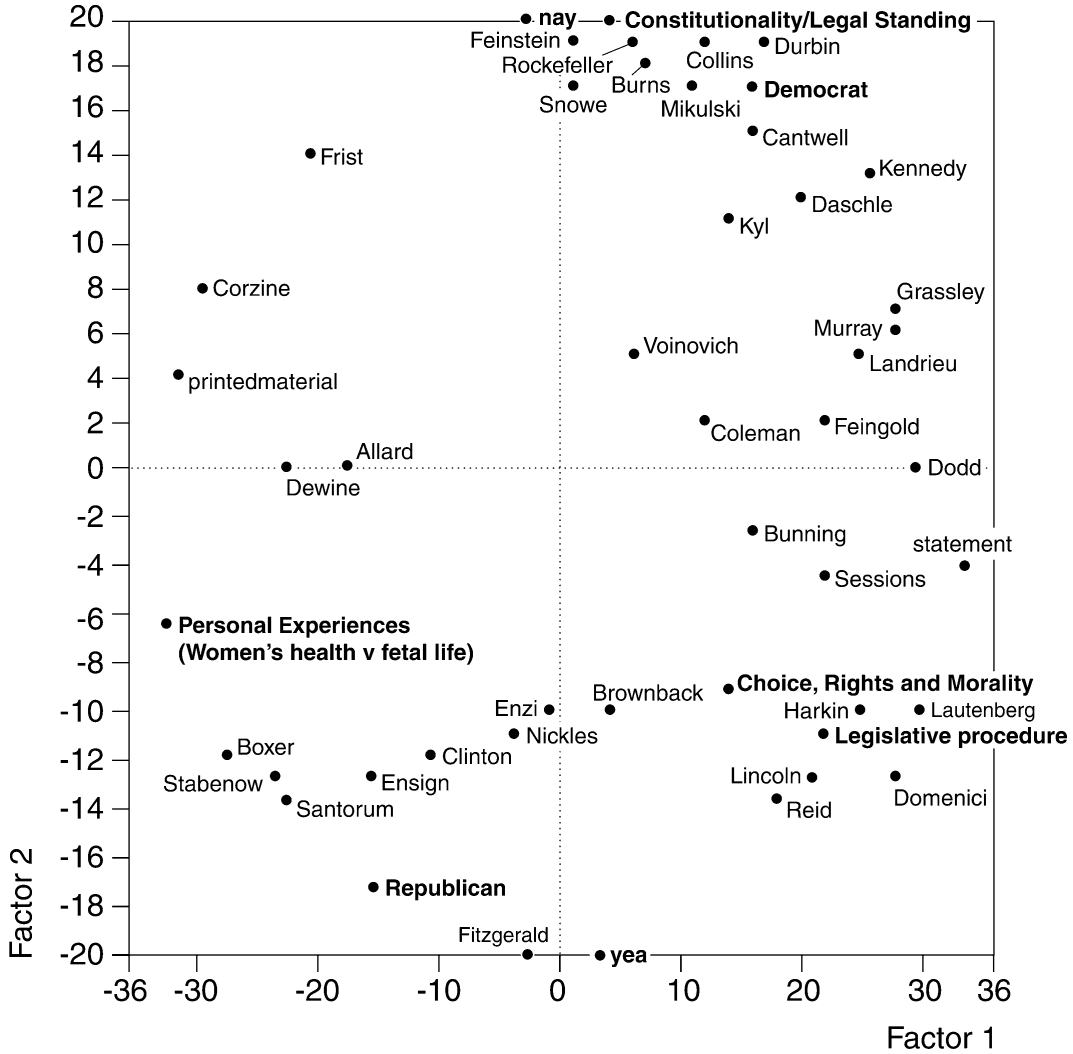
$$\underset{i \times j}{\mathbf{X}} = \underset{i \times k}{\mathbf{U}} \; \underset{k \times k}{\mathbf{d}} \; \underset{j \times k}{\mathbf{V}'} \tag{10}$$

- The **U**, **d**, and **V** matrixes "relocate" the elements of **X** onto new coordinate vectors in n-dimensional Euclidean space
- Row variables of **X** become points on the **U** column coordinates, and the column variables of **X** become points on the **V** column coordinates
- The coordinate vectors are perpendicular (*orthogonal*) to each other and are normalized to unit length

Correspondence Analysis and SVD

- Divide each value of **X** by the geometric mean of the corresponding marginal totals (square root of the product of row and column totals for each cell)
  - Conceptually similar to subtracting out the χ^2 expected cell values from the observed cell values
- Perform an SVD on this transformed matrix
  - This yields singular values **d** (with first always 1.0)
- Rescale the row (**U**) and column (**V**) vectors to obtain canonical scores (rescaled as $U_i \sqrt{f_{..}/f_{i.}}$ and $V_j \sqrt{f_{..}/f_{j.}}$)

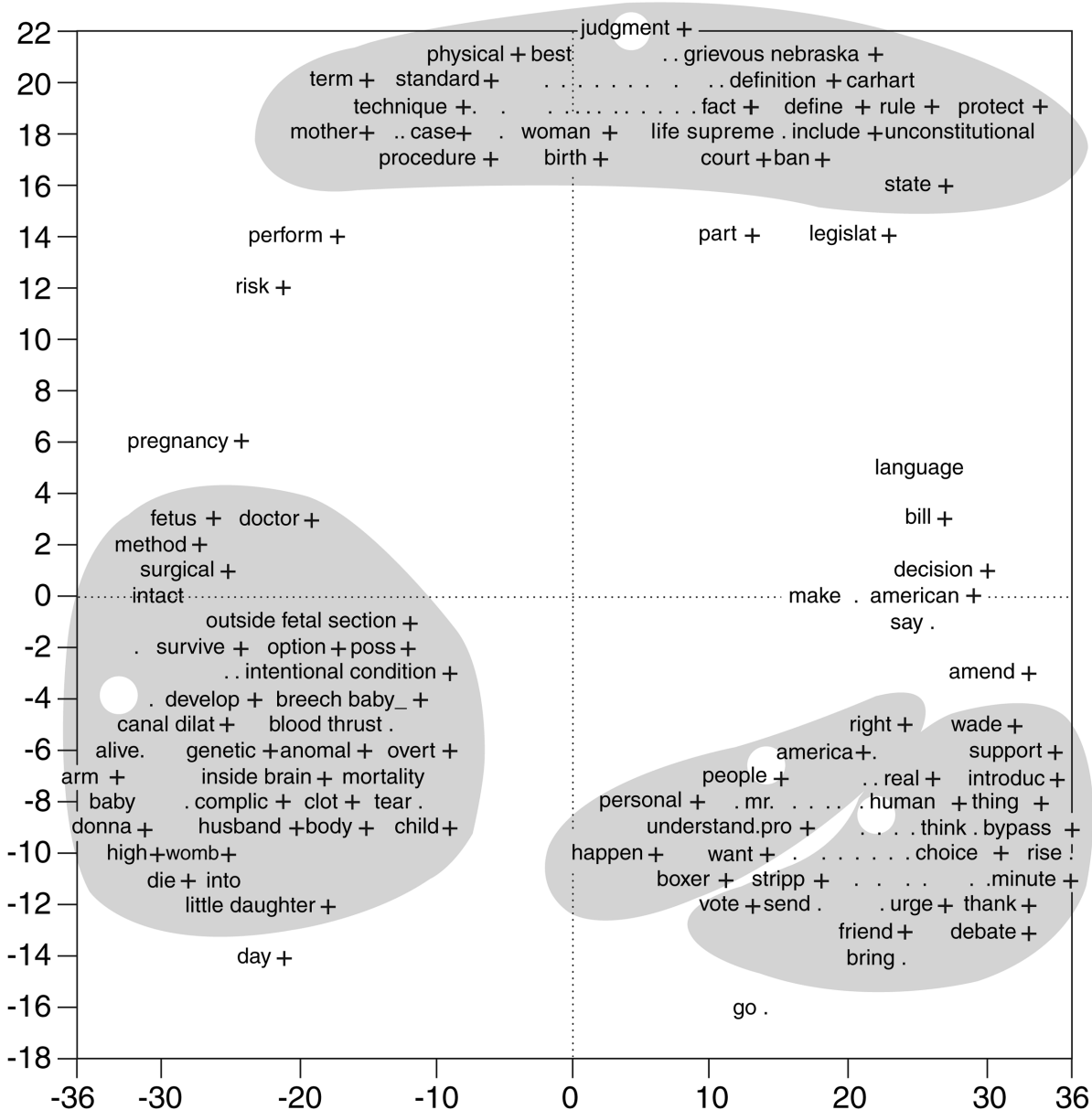
Example: Schonhardt-Bailey (2008) - speakers



	Eigenvalue	% Association	% Cumulative
Factor 1	0.30	44.4	44.4
Factor 2	0.22	32.9	77.3

Fig. 3. Correspondence analysis of classes and tags from Senate debates on Partial-Birth Abortion Ban Act

Example: Schonhardt-Bailey (2008) - words



How to get confidence intervals for CA

- There are problems with bootstrapping: (Milan and Whittaker 2004)
  - rotation of the principal components
  - inversion of singular values
  - reflection in an axis

How to account for uncertainty

How to account for uncertainty

- Ignore the problem and hope it will go away

How to account for uncertainty

- Ignore the problem and hope it will go away
  - SVD-based methods (e.g. correspondence analysis) typically do not present errors
  - and traditionally, point estimates based on other methods have not either

How to account for uncertainty

How to account for uncertainty

- Analytical derivatives
  - Using the multinomial formulation of the Poisson model, we can compute a Hessian for the log-likelihood function
  - The standard errors on the θ_i parameters can be computed from the covariance matrix from the log-likelihood estimation (square roots of the diagonal)
  - The covariance matrix is (asymptotically) the inverse of the negative of the Hessian
    (where the negative Hessian is the observed Fisher information matrix, a.ka. the second derivative of the log-likelihood evaluated at the maximum likelihood estimates)

How to account for uncertainty

- Analytical derivatives
  - Using the multinomial formulation of the Poisson model, we can compute a Hessian for the log-likelihood function
  - The standard errors on the θ_i parameters can be computed from the covariance matrix from the log-likelihood estimation (square roots of the diagonal)
  - The covariance matrix is (asymptotically) the inverse of the negative of the Hessian
    (where the negative Hessian is the observed Fisher information matrix, a.ka. the second derivative of the log-likelihood evaluated at the maximum likelihood estimates)
  - Problem: These are *too small*

How to account for uncertainty

How to account for uncertainty

- Parametric bootstrapping (Slapin and Proksch, Lewis and Poole)
  Assume the distribution of the parameters, and generate data after drawing new parameters from these distributions.
  Issues:
  - slow
  - relies heavily (twice now) on parametric assumptions
  - requires some choices to be made with respect to data generation in simulations

How to account for uncertainty

- Parametric bootstrapping (Slapin and Proksch, Lewis and Poole)
  Assume the distribution of the parameters, and generate data after drawing new parameters from these distributions.
  Issues:
  - slow
  - relies heavily (twice now) on parametric assumptions
  - requires some choices to be made with respect to data generation in simulations
- Non-parametric bootstrapping
-

How to account for uncertainty

- Parametric bootstrapping (Slapin and Proksch, Lewis and Poole)
  Assume the distribution of the parameters, and generate data after drawing new parameters from these distributions.
  Issues:
  - slow
  - relies heavily (twice now) on parametric assumptions
  - requires some choices to be made with respect to data generation in simulations
- Non-parametric bootstrapping
-
- (and yes of course) Posterior sampling from MCMC

How to account for uncertainty

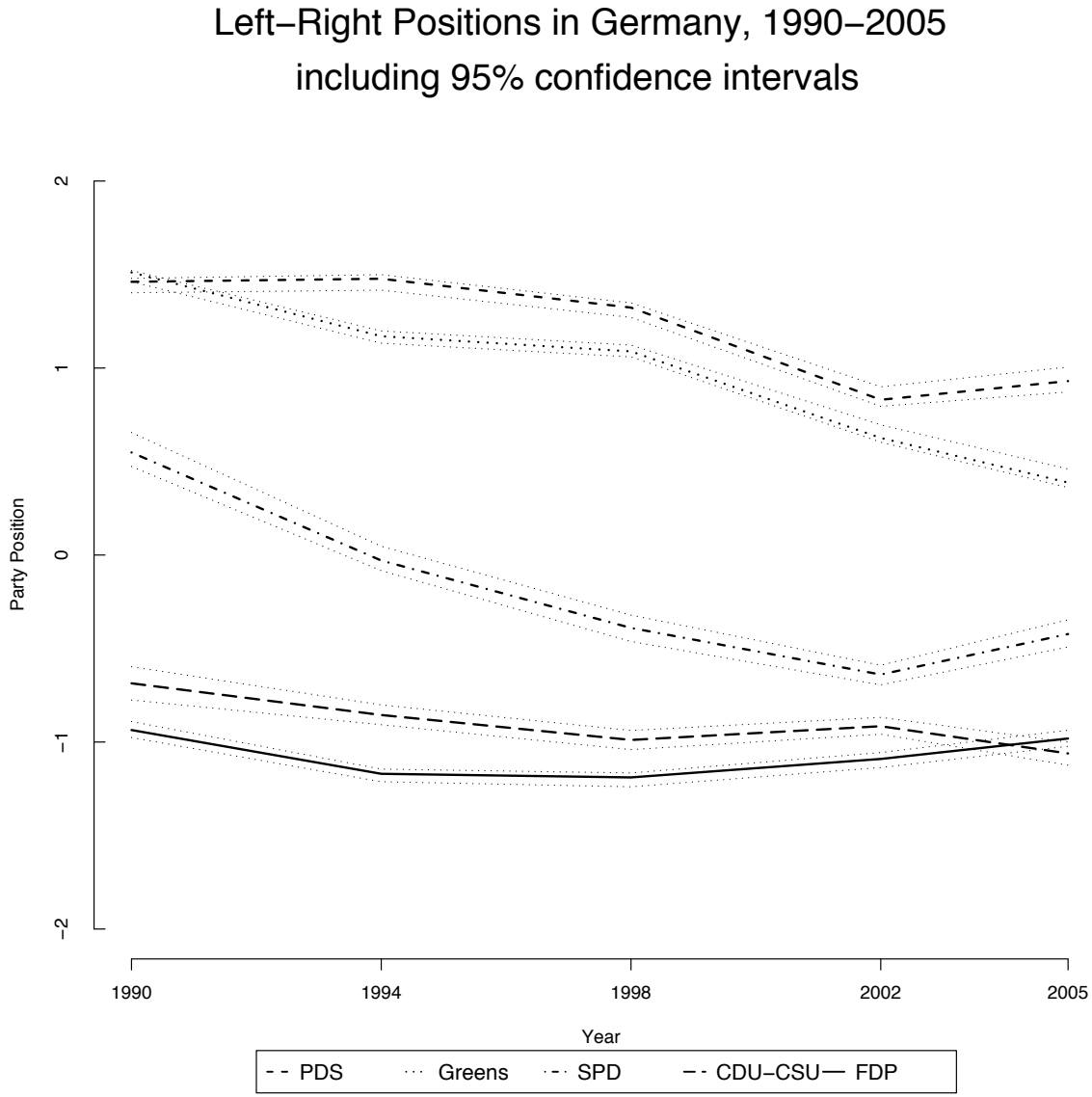
How to account for uncertainty

- Non-parametric bootstrapping
  - draw new versions of the texts, refit the model, save the parameters, average over the parameters
  - slow
  - not clear how the texts should be resampled

How to account for uncertainty

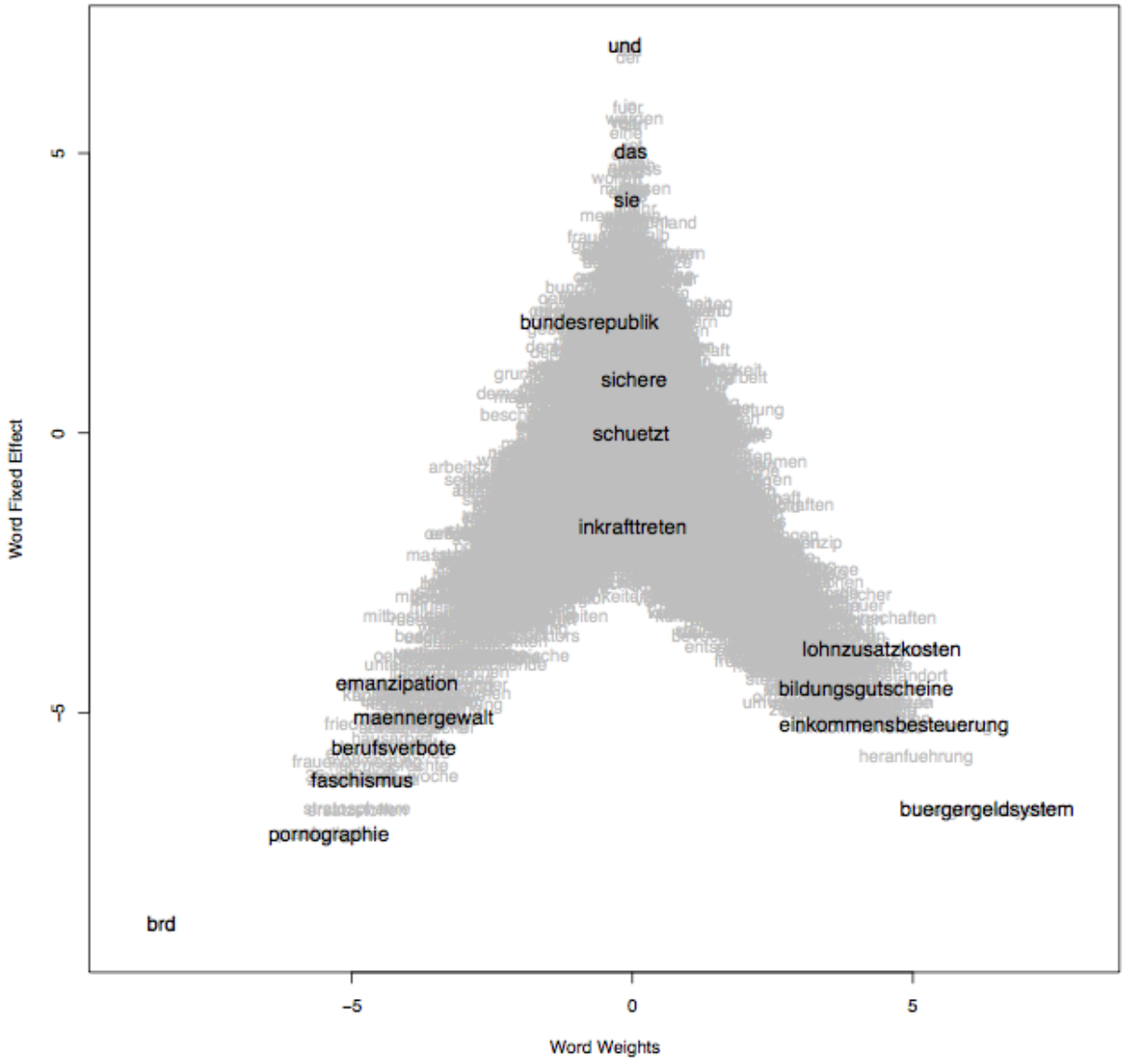
- For MCMC: from the distribution of posterior samples

Parametric Bootstrapping and analytical derivatives yield "errors" that are too small



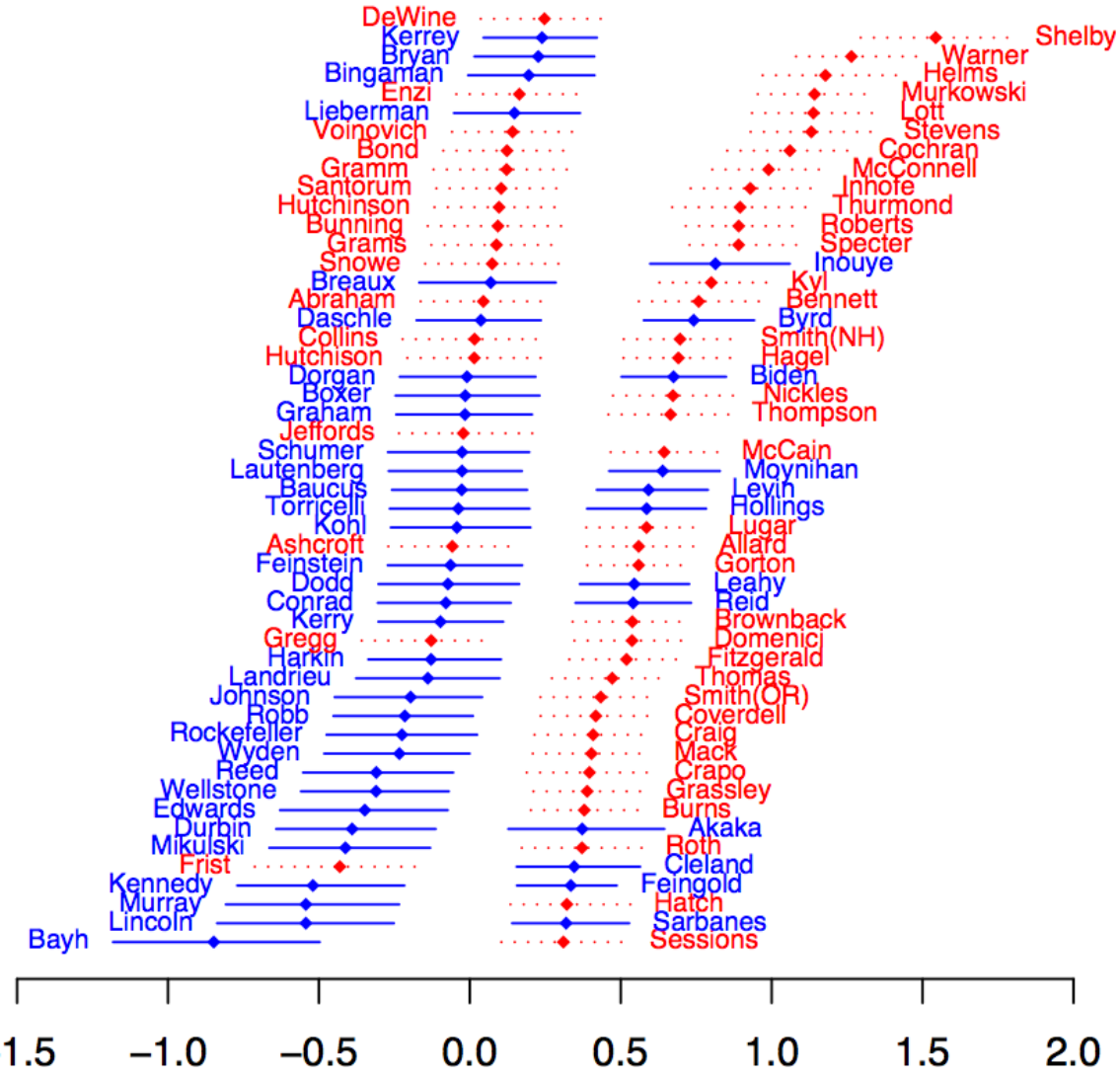
Frequency and informativeness

Ψ and β (frequency and informativeness) tend to trade-off



Plotting θ

Plotting θ (the ideal points) gives estimated positions. Here is Monroe and Maeda's (essentially identical) model of legislator positions:



Interpreting multiple dimensions

To get one dimension for each policy area, split up the document by hand and use the subparts as documents (the Slapin and Proksch method)
There is currently *no* implementation of Wordscores or Wordfish that extracts two or more dimensions at once

- But since Wordfish is a type of factor analysis model, there is no reason in principle why it could not

Interpreting scaled dimensions

- Another (better) option: compare them other known descriptive variables

Interpreting scaled dimensions

- Another (better) option: compare them other known descriptive variables
- Hopefully also *validate* the scale results with some human judgments

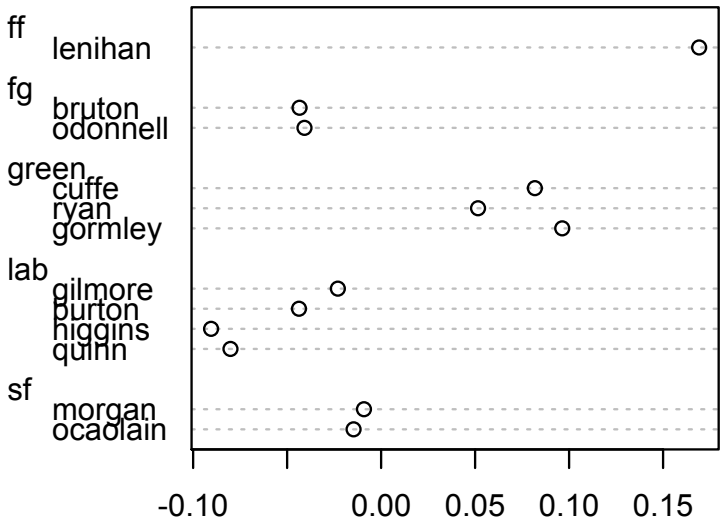
Interpreting scaled dimensions

- Another (better) option: compare them other known descriptive variables
- Hopefully also *validate* the scale results with some human judgments
- This is necessary even for single-dimensional scaling

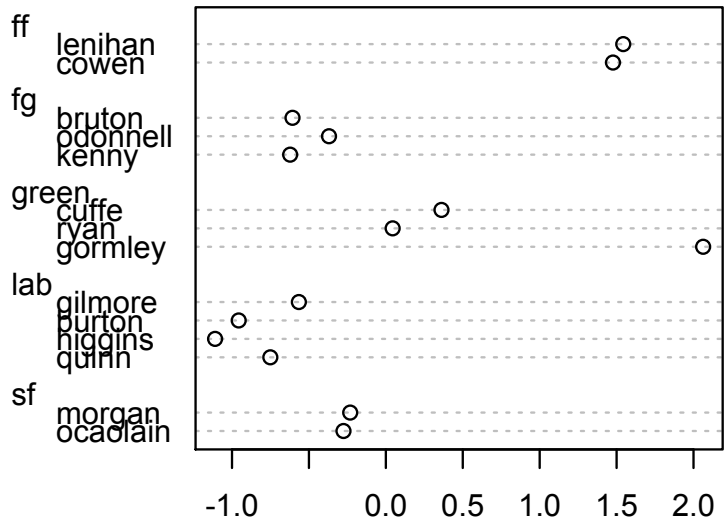
Interpreting scaled dimensions

- Another (better) option: compare them other known descriptive variables
- Hopefully also *validate* the scale results with some human judgments
- This is necessary even for single-dimensional scaling
- And just as applicable for non-parametric methods (e.g. correspondence analysis) as for the Poisson scaling model

What happens if we include irrelevant text?



Wordscores LBG Position on Budget 2009



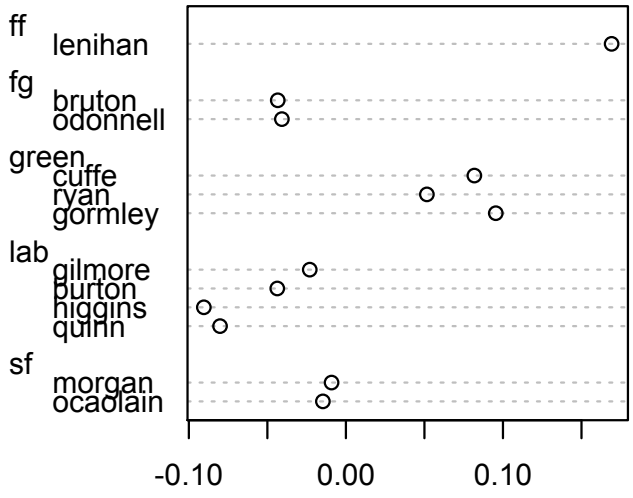
Normalized CA Position on Budget 2009

What happens if we include irrelevant text?

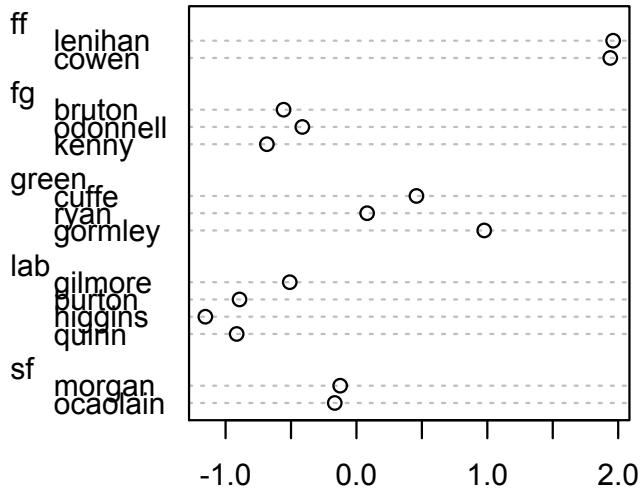
John Gormley: leader of the Green Party and Minister for the Environment, Heritage and Local Government

"As leader of the Green Party I want to take this opportunity to set out my party's position on budget 2010..."
[772 words later]
"I will now comment on some specific aspects of my Department's Estimate. I will concentrate on the principal sectors within the Department's very broad remit ..."

Without irrelevant text



Wordscores LBG Position on Budget 2009



Normalized CA Position on Budget 2009