Day 6: Text Scaling

Kenneth Benoit

CEU April 14-21 2011

April 21, 2011

Models for continuous θ

Background: Spatial politics

Methods

- Wordscores
- Wordfish

Document scaling is for continuous θ

Some spatial theory

Spatial theories of *national* voting assumes that

- Voters and politicians/parties have *preferred positions* 'ideal points' on ideological dimensions or policy spaces
- Voters support the politician/prty with the ideal point *nearest* their own
- Politicians/parties *position themselves* to maximize their vote share

Some spatial theory

Spatial theories of *parliamentary* voting assume that

- Each vote is a decision between *two* policy outcomes
- Each outcomes has a position on an ideological dimension or a policy space
- Voters choose the outcome *nearest* to their own ideal point

Unobserved ideal points / policy positions: θ

Voting 'reveals' θ (sometimes)

Spatial utility models

Measurement models for votes (Jackman, 2001; Clinton et al. 2004) connect voting choices to personal utilities and ideal points

Parliamentary voting example: Ted Kennedy on the 'Federal Marriage Amendment'

$$\begin{aligned} U(\pi_{\text{yes}}) &= -\|\theta - \pi_{\text{yes}}\|^2 + \epsilon_{\text{yes}} \\ U(\pi_{\text{no}}) &= -\|\theta - \pi_{\text{no}}\|^2 + \epsilon_{\text{no}} \end{aligned}$$

- θ is Kennedy's ideal point
- π_{yes} is the policy outcome of the FMA passing (vote yes)
- π_{no} is the policy outcome of the FMA failing (vote no)

Votes 'yes' when $U(\pi_{\text{yes}}) > U(\pi_{\text{no}})$

Spatial utility models and voting

What is the probability that Ted votes yes?

$$
\begin{aligned}
P(\text{Ted votes yes}) &= P(U(\pi_{\text{yes}}) > U(\pi_{\text{no}})) \\
&= P(\epsilon_{\text{no}} - \epsilon_{\text{yes}} < \|\theta - \pi_{\text{no}}\|^2 - \|\theta - \pi_{\text{yes}}\|^2) \\
&= P(\epsilon_{\text{no}} - \epsilon_{\text{yes}} < 2(\pi_{\text{yes}} - \pi_{\text{no}})\theta + \pi_{\text{no}}^2 - \pi_{\text{yes}}^2) \\
\text{logit } P(\text{Ted votes yes}) &= \beta\theta + \alpha
\end{aligned}
$$

Only the 'cut point' or separating hyperplane *between* π_{yes} and π_{no} matters

This is logistic regression model with explanatory variable θ

Spatial voting models

This is a simple measurement model

There is some distribution of ideal points in the population (the legislature)

$$P(\theta) \;=\; \text{Normal}(0, 1)$$

Votes are conditionally independent given ideal point

$$P(\text{vote}_1, \ldots, \text{vote}_K \mid \theta) \;=\; \prod_j P(\text{vote}_j \mid \theta)$$

Probability of voting yes is monotonic in the *difference* between policy outcomes

$$P(\text{yes}) \;=\; \text{Logit}^{-1}(\beta\theta + \alpha)$$

Two problems inferring politicians' ideal points from votes

When voting probability depends on *more* than θ

- Example: party discipline (Mclean and Spirling)

Vote contents and timing are not random or representative

- Example: institutional constraint or partisan influence on whether and how issues are put to vote (Carrubba et al.)

Avoid these problems by inferring positions from text

Alternative options:

- expert survey
- text

Data sources

What we can learn from what...

	Parties	Legislators	Voters
Surveys	yes	sometimes	yes
Votes	sometimes	sometimes	sometimes
Words	yes	yes	sometimes

The relationship between policy preferences and words is less clear than for votes

Need to be clear what we are assuming in our measurement models

Data sources: manifestos

Advantages

- known sample selection mechanism
- approximately uniform policy coverage
- representative
- known audience
- lots of words

Disadvantages

- uninformative about within-party variation
- so uninformative about individual legislators
- few documents

Data sources: speeches

Advantages

- informative about individual legislators
- more focused policy content
- lots of 'documents'

Disadvantages

- variable (institutionally structured) sample selection
- possibly unrepresentative
- variable audience
- few words

Wordscores

Wordscores (Laver et al. 2003) is an automated procedure for estimating policy positions from 'virgin documents'

Assumptions

- Every word has a policy position called a *score* π_i
- The score of a document θ is the *average* of the scores of its words
- R documents with *known* scores $\theta_1 \ldots \theta_R$ on a *known* dimension are available

Procedure

- Compute V wordscores $\pi_1 \ldots \pi_V$ from the reference documents
- Estimate virgin document scores from estimated wordscores

Wordscores

Algorithmic exposition (from Laver et al. 2003):

Estimating the probability of seeing a word *given* that we are reading a particular document

Use the word count matrix

$$P(w_i \mid d_j) = \frac{c(\text{word i in j})}{c(\text{all words in j})} \quad = \quad \frac{\mathbf{C}_{ij}}{\sum_i \mathbf{C}_{ij}}$$

If we just see w_i, what is the probability we are in document 1? document 2? document R?

Estimating Probabilities

Use the word frequency matrix $\mathbf{C}$:

	uklab92,	uklab97,	uklab01,	uklab05, ...
farm,	1	0	1	0
farmer,	0	0	0	1
farmers,	1	0	2	1
farming,	2	0	7	0
...				

If $i =$ farm and $j =$ uklab01, then $\mathbf{C}_{ij} = 1$ and

$$P(W = \text{'farm'} \mid d = \text{'uklab01'}) \;=\; 1/10 \;=\; 0.1$$

if these were the only words in the manifesto

Wordscores

Then use this to compute the probability of each document given we see a particular word

By Bayes theorem

$$P(d_j \mid w_i) = \frac{P(w_i \mid d_j)\ P(d_j)}{\sum_j P(w_i \mid d_j)\ P(d_j)}$$

What is $P(d_j)$? Assume it is uniform $(1/R)$. Then

$$P(d_j \mid w_i) = \frac{P(w_i \mid d_j)}{\sum_j P(w_i \mid d_j)}$$

Wordscores

Wordscores are a weighted average of the document scores

$$\hat{\pi}_i = \sum_j^R \theta_j \, P(d_j \mid w_i)$$

Words that are more likely to turn up in an ideologically left document get a leftish score

Shrinkage

To score new documents, take the average of the scores of the words it contains

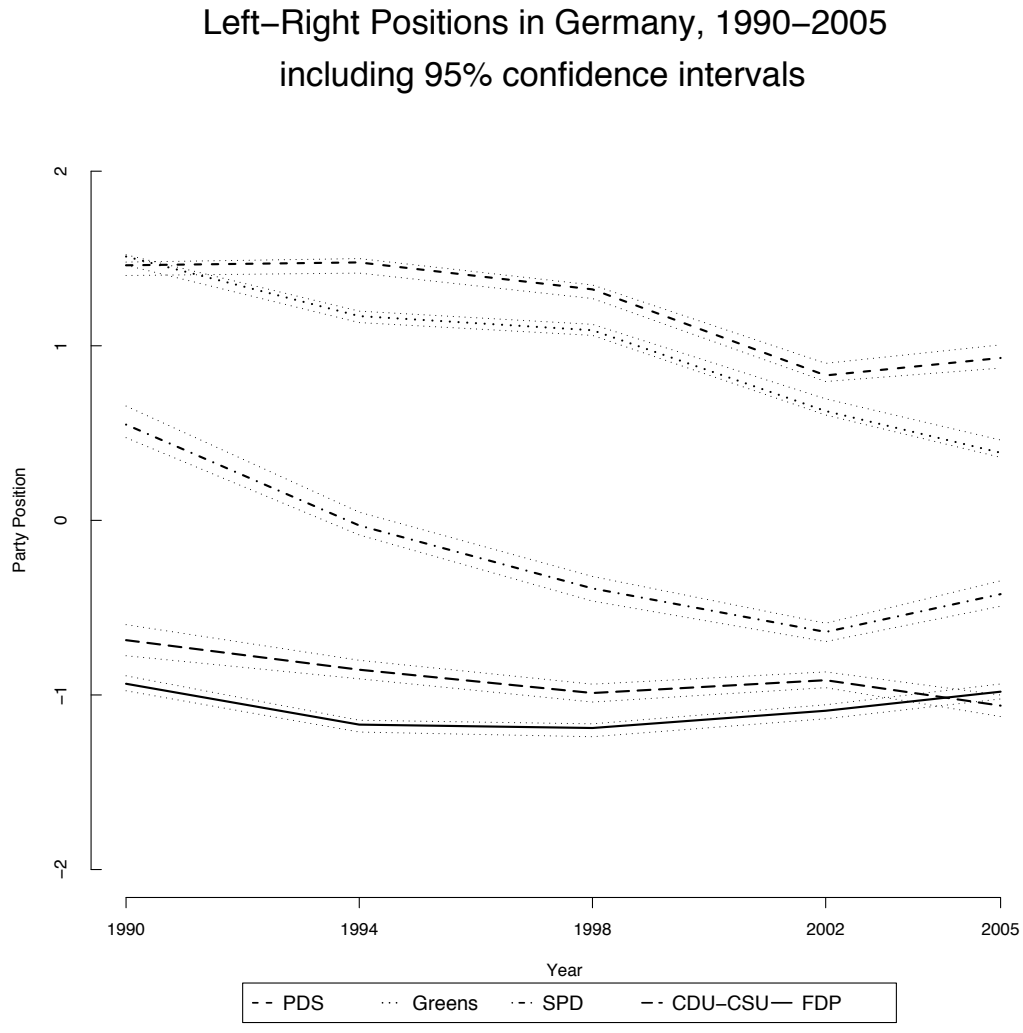
$$\begin{aligned}\hat{\theta}_j &= \hat{\pi}^{(1)} + \hat{\pi}^{(2)} + \hat{\pi}^{(3)} + \ldots + \hat{\pi}^{(N)} \\ &= \sum_{i}^{V} \hat{\pi}_i \; P(w_i \mid d_j)\end{aligned}$$

The more leftish words in a document, the further left its score

Note the *symmetry*: wordscores are a weighted average of document scores; document scores are a weighted average of wordscores

Wordfish

Wordfish is a statistical model for inferring policy positions θ from words



As measurement model

Assumptions about $P(W_1 \ldots W_V \mid \theta)$

$$\log E(W_i \mid \theta_j) \;=\; \alpha_j \;+\; \psi_i \;+\; \beta_i \theta_j$$

α_j is a constant term controlling for document length (hence it's associated with the party or politician)

The *sign* of β_i represents the *ideological direction* of W_i

The *magnitude* of β_i represents the *sensitivity* of the word to ideological differences among speakers or parties

Ψ is a constant term for the word (larger for high frequency words).

Maximum Likelihood fitting algorithm

A form of Expectation Maximization:

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

Recap: Wordfish

Start by *guessing* the parameters

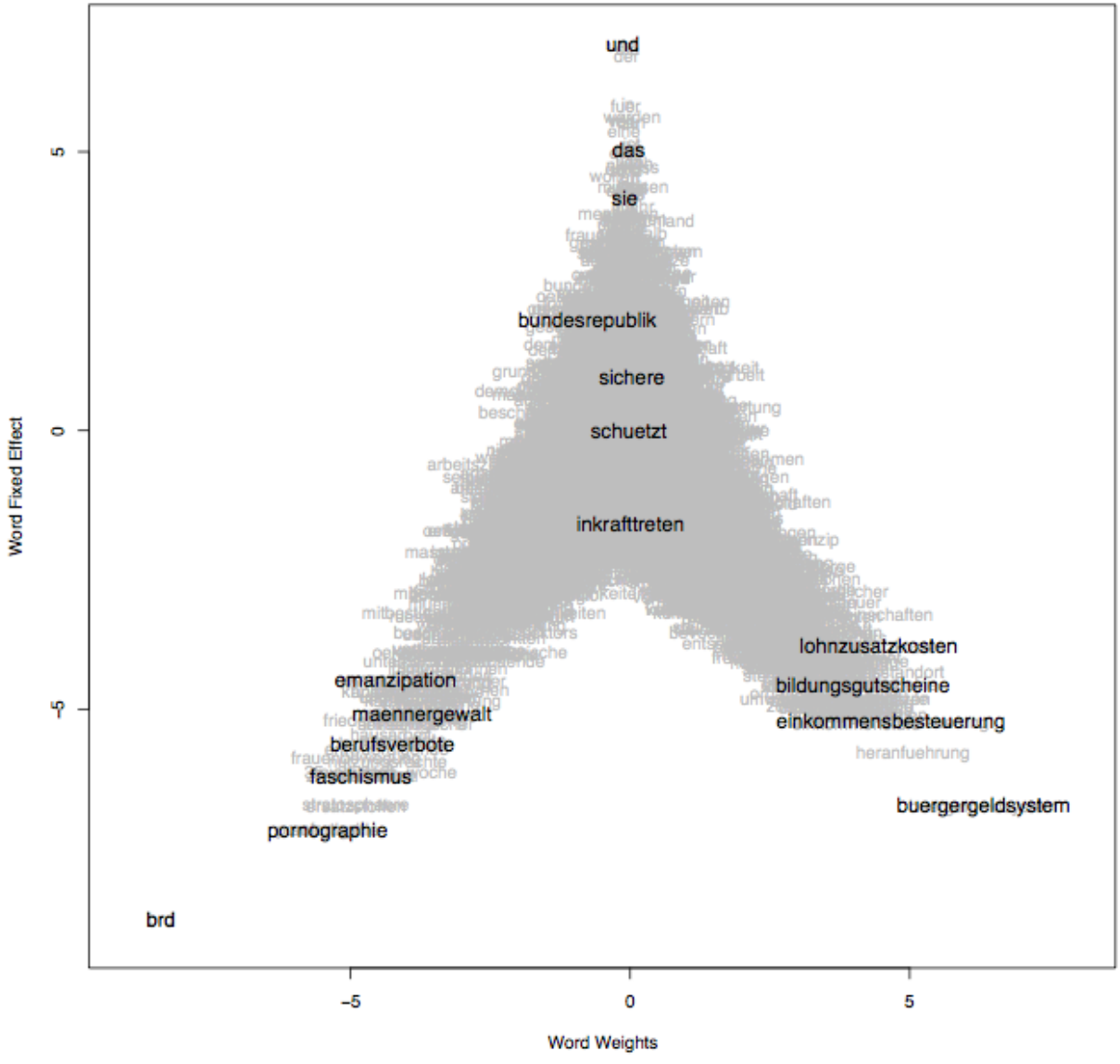
Algorithm:

- 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

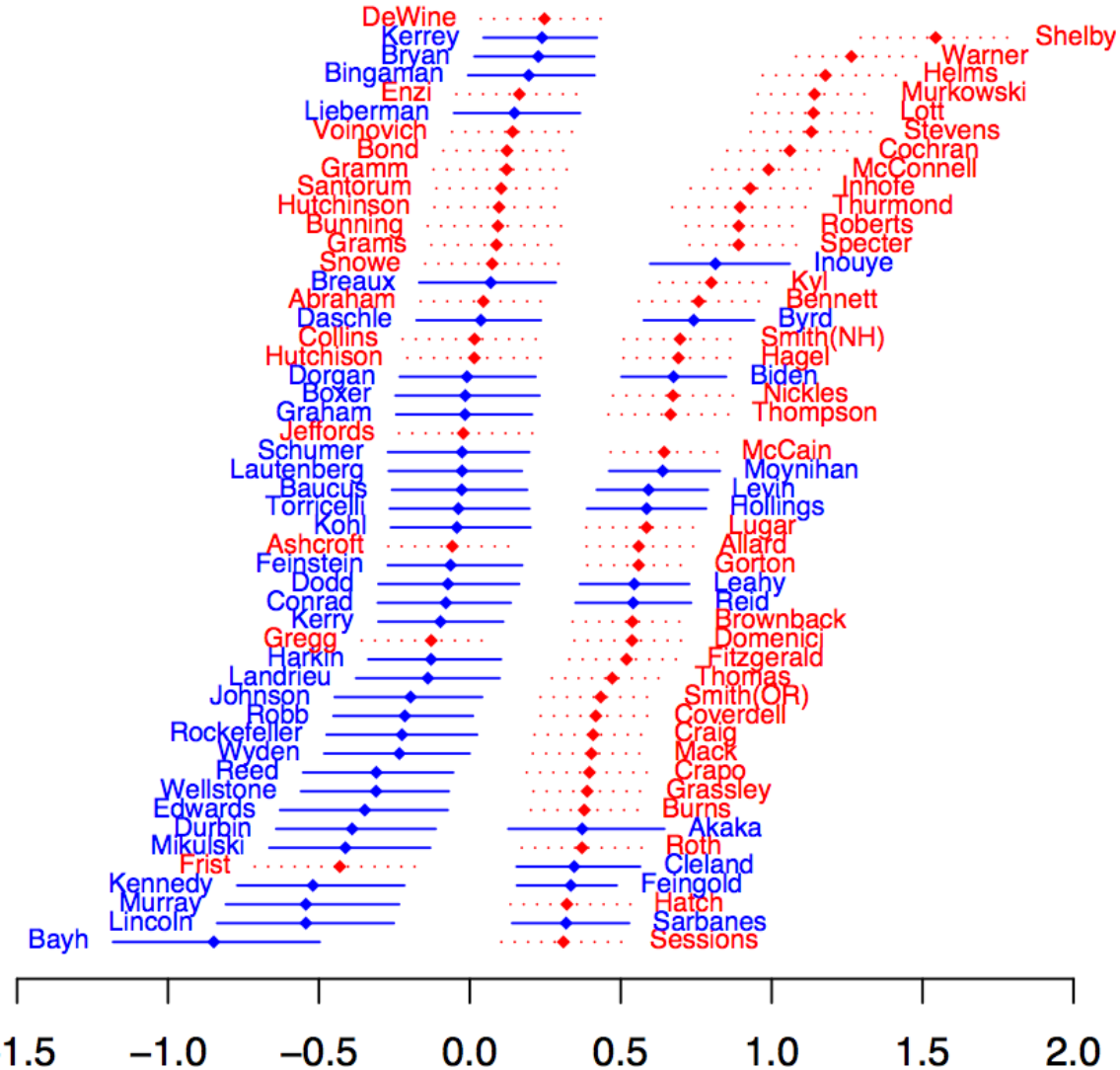
Frequency and informativeness

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:



Wordscores and Wordfish as measurement models

Wordfish assumes that

$$P(\theta) = Normal(0, 1)$$

and that $P(W_i \mid \theta)$ depends on

- Word parameters: β and ψ
- Document / party / politician parameters: θ and α

Wordscores and Wordfish as measurement models

Wordfish estimates of θ *control* for

- different document lengths (α)
- different word frequencies (ψ) different levels of ideological relevance of words (β).

But there are no wordscores!

Words do not have an ideological position themselves, only a sensitivity to the speaker's ideological position

Hot off the press

Wordscores makes no explicit assumption about $P(\theta)$ except that it is continuous

We infer that $P(W_i \mid \theta)$ depends on

- Wordscores: π
- Document scores: θ

Hence θ estimates do *not* control for

- different word frequencies
- different levels of ideological relevance of words

Dimensions

How to interpret $\hat{\theta}$s substantively?

One option is to *regress* them other known descriptive variables

Example European Parliament speeches (Proksch and Slapin)

- Inferred ideal points seem to reflect party positions on EU integration better than national left-right party placements

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)

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

Dimensions

How infer more than one dimension?

This is two questions:

- How to get two dimensions (for all policy areas) at the same time?
- How to get one dimension for each policy area?

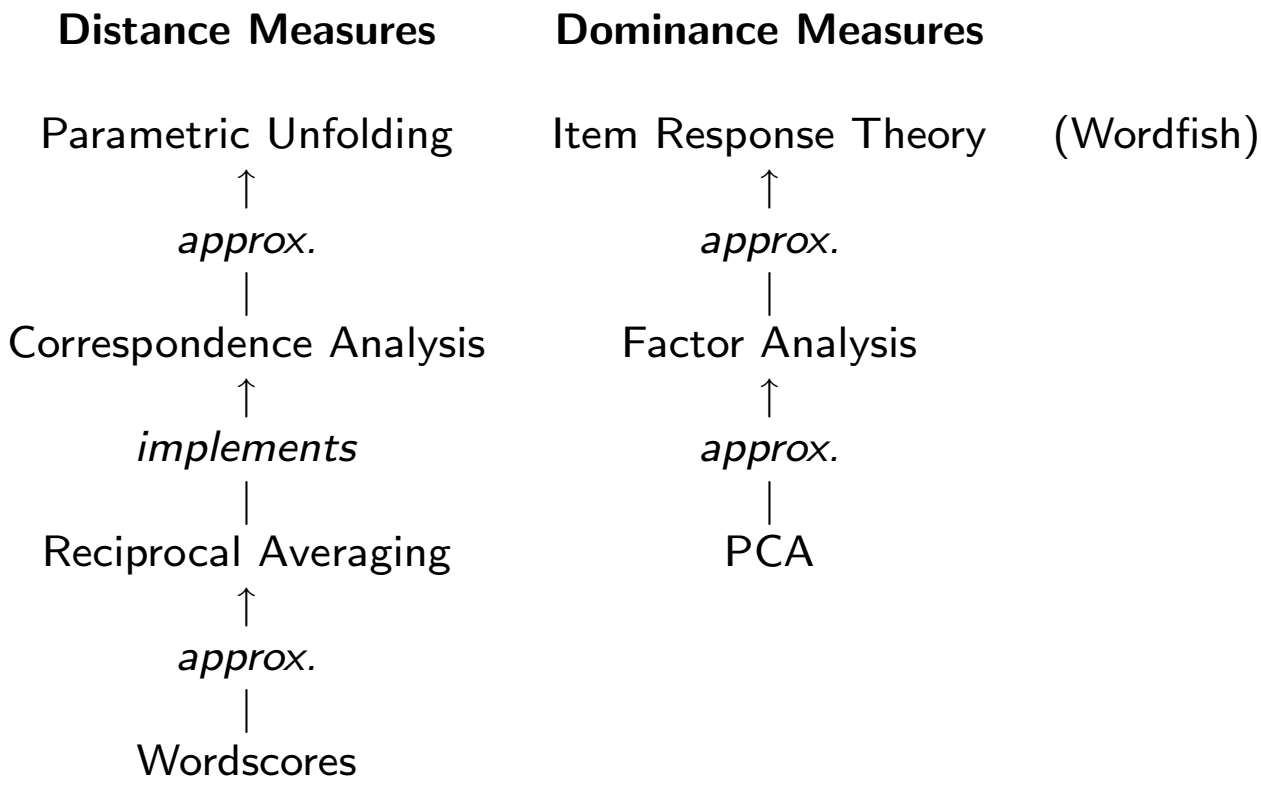
Dimensions

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

Measurement models again



Document Scaling Software

Software for Wordscores and Wordfish is available for R (and Stata for Wordscores)

Currently: the austin library written by Will Lowe

The Poisson scaling "wordfish" model

Data:

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

Model:

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

Estimation:

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

Model components and notation

Element	*Meaning*
i	indexes the targets of interest (political actors)
N	number of political actors
j	indexes word types
V	total number of word types
θ_i	the unobservable political position of actor i
β_j	word parameters on θ – the "ideological" direction of word j
ψ_j	word "fixed effect" (function of the frequency of word j)
α_i	actor "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 laid nakedly bare for inspection
  - *conditional independence*
  - *stochastic process* (e.g. $\mathrm{E}(Y_{ij}) = \mathrm{Var}(Y_{ij}) = \lambda_{ij}$)
- Permits hierarchical reparameterization (to add covariates)
- Prediction: in particular, out of sample prediction

Problems laid bare I: Conditional (non-)independence

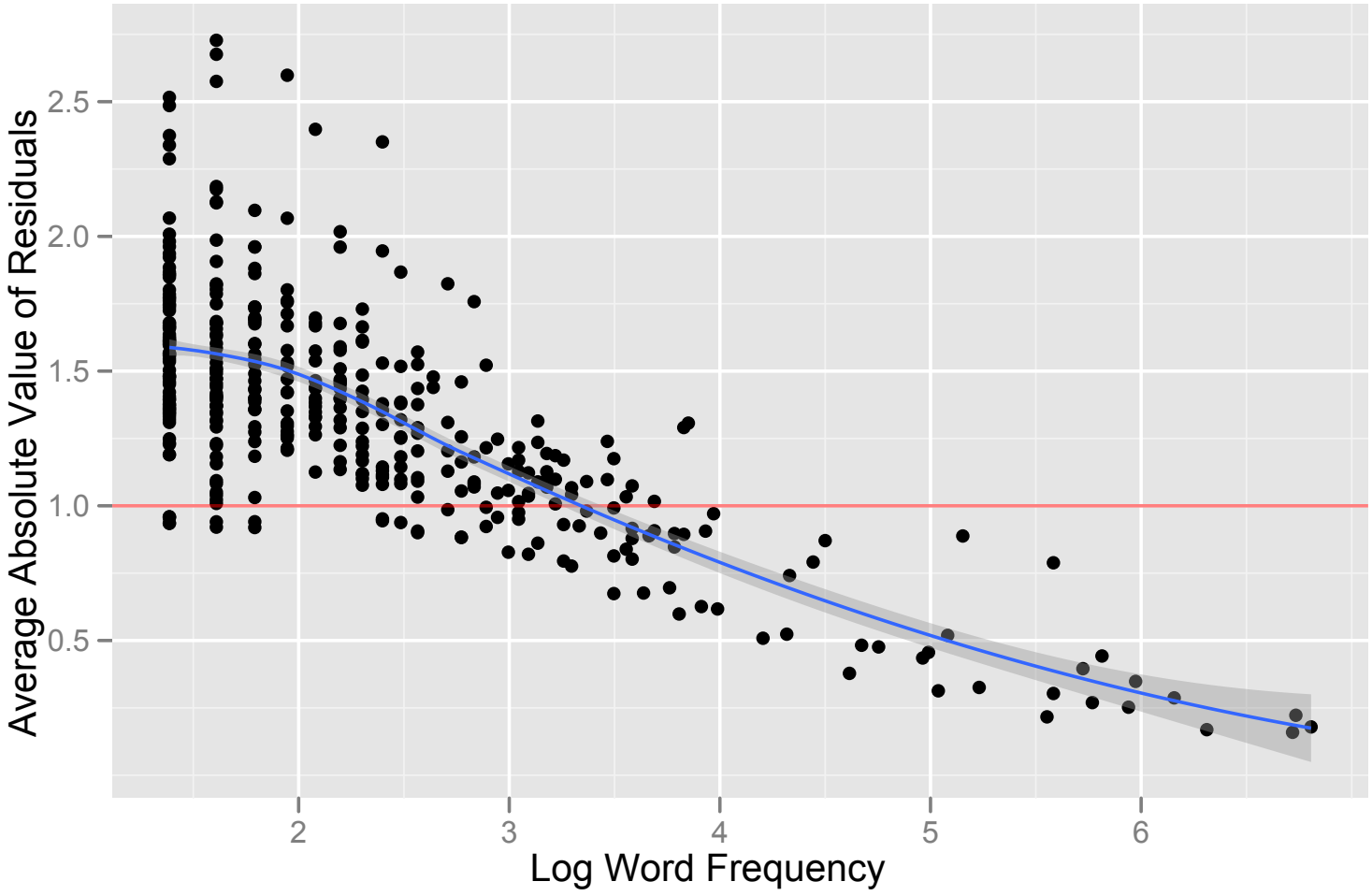
- Words occur in order
  In occur words order.
  Occur order words in.
  "No more training do you require. Already know you that which you need." (Yoda)
- Words occur in combinations
  "carbon tax" / "income tax" / "inhertiance tax" / "capital gains tax" /"bank tax"
- Sentences (and topics) occur in 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

Problems laid bare II: Parametric (stochastic) model

- 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

(from Slapin and Proksch 2008)



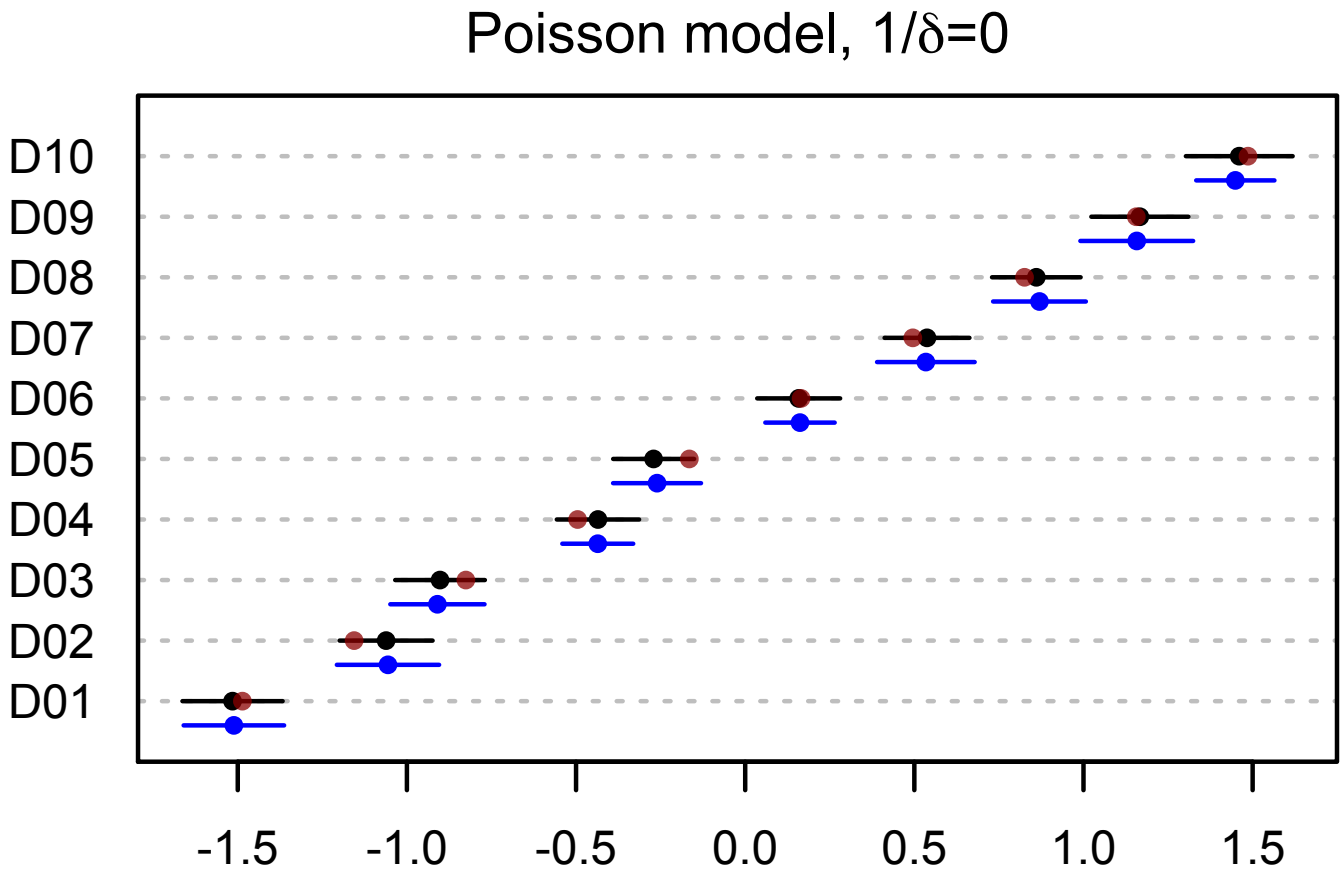
How to account for uncertainty?

- Don't. (SVD-like methods, e.g. correspondence analysis)
- Analytical derivatives
- Parametric bootstrapping (Slapin and Proksch, Lewis and Poole)
- Non-parametric bootstrapping
- (and yes of course) Posterior sampling from MCMC

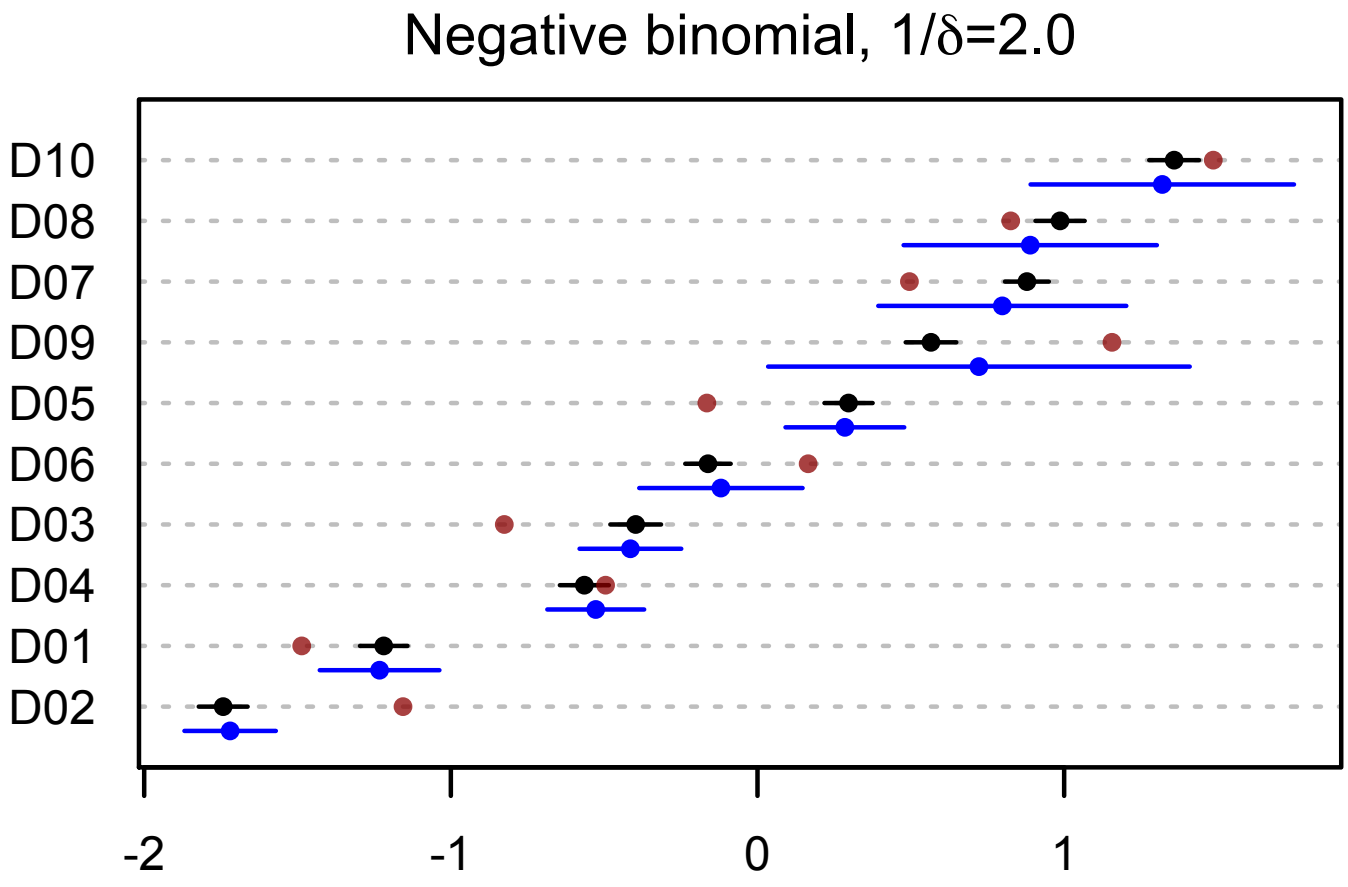
Steps forward

- Diagnose (and ultimately treat) the issue of whether a separate variance parameter is needed
- Diagnose (and treat) violations of conditional independence
- Explore non-parametric methods to estimate uncertainty

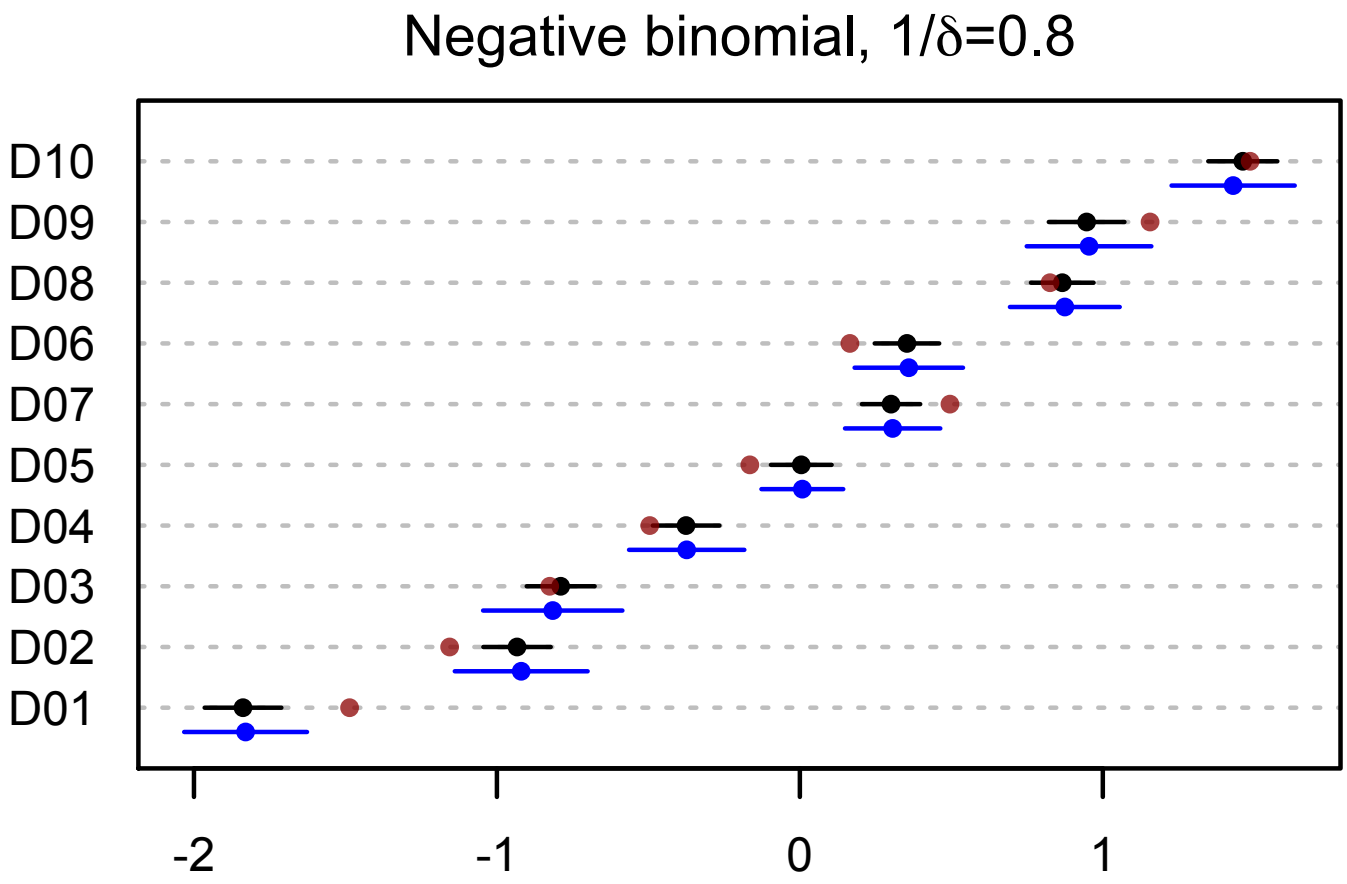
Diagnosis I: Estimations on simulated texts



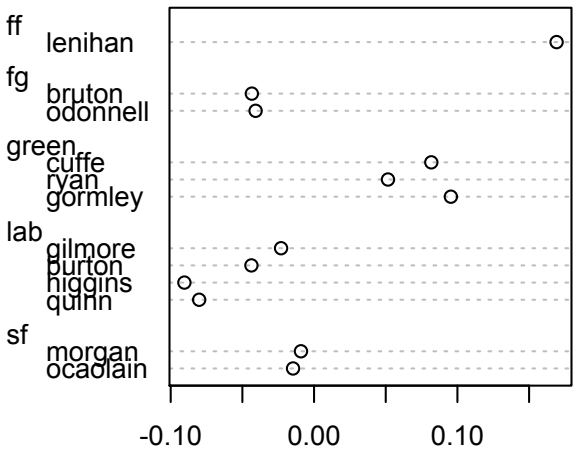
Diagnosis I: Estimations on simulated texts



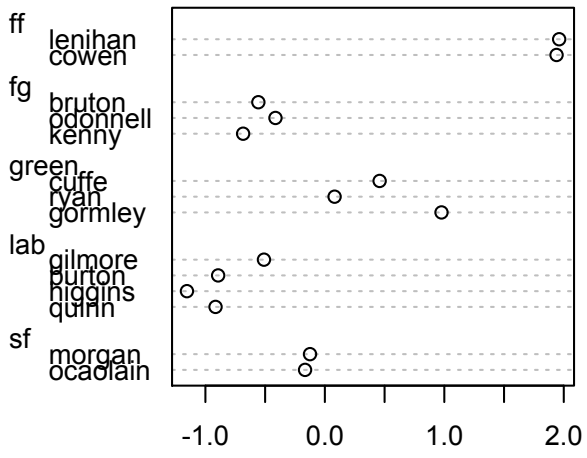
Diagnosis I: Estimations on simulated texts



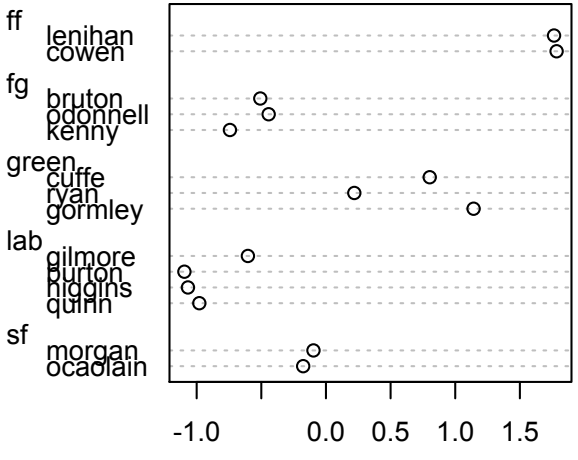
Diagnosis 2: Irish Budget debate of 2009



Wordscores LBG Position on Budget 2009



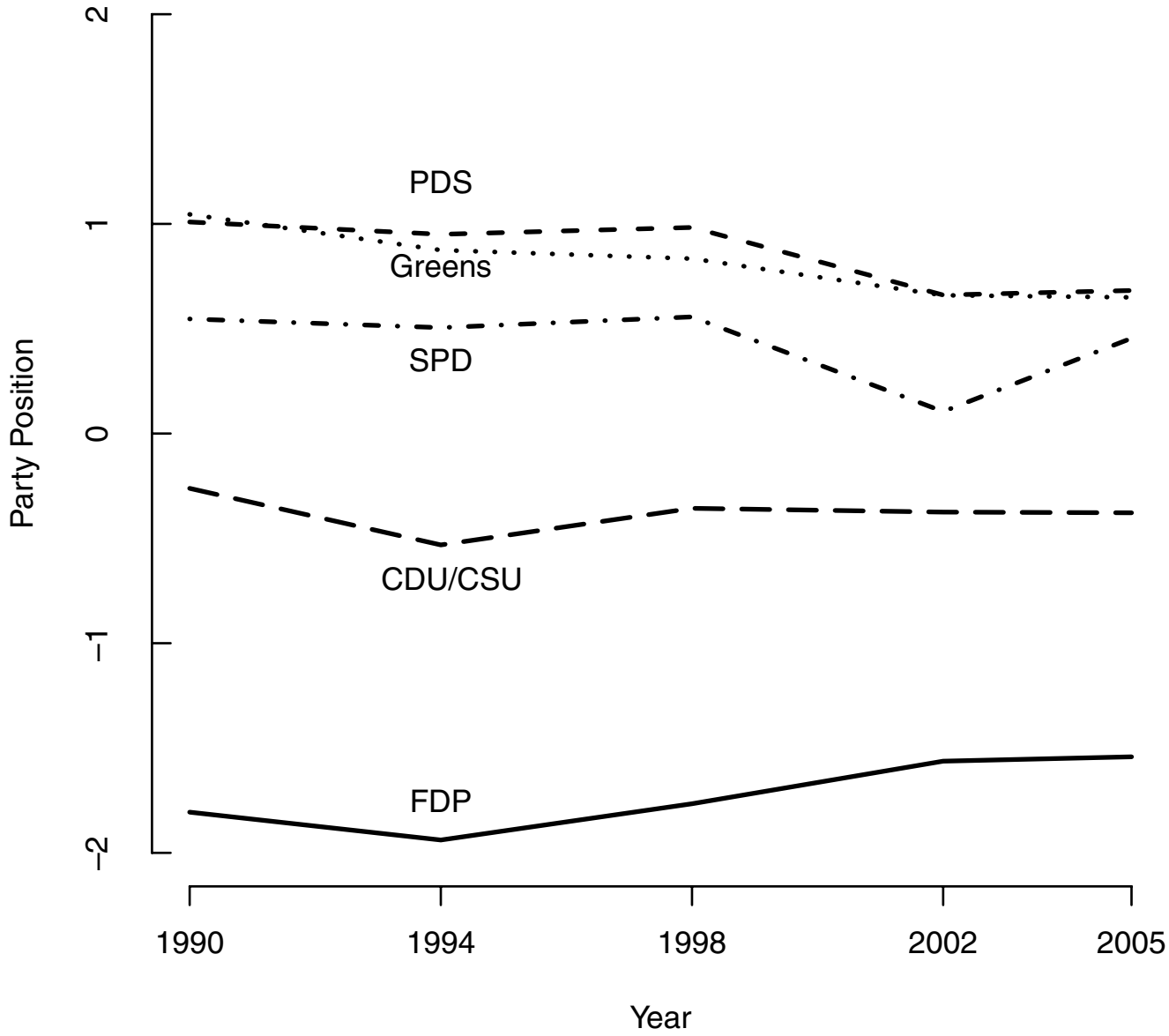
Normalized CA Position on Budget 2009



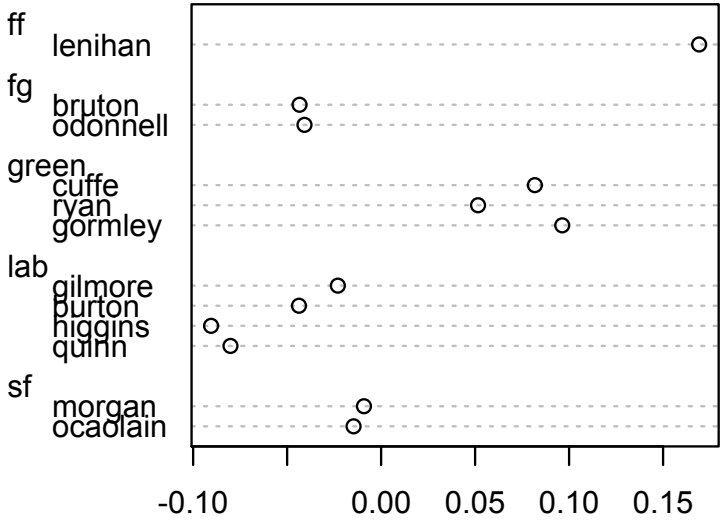
Classic Wordfish Position on Budget 2009

Diagnosis 3: German party manifestos (economic sections)

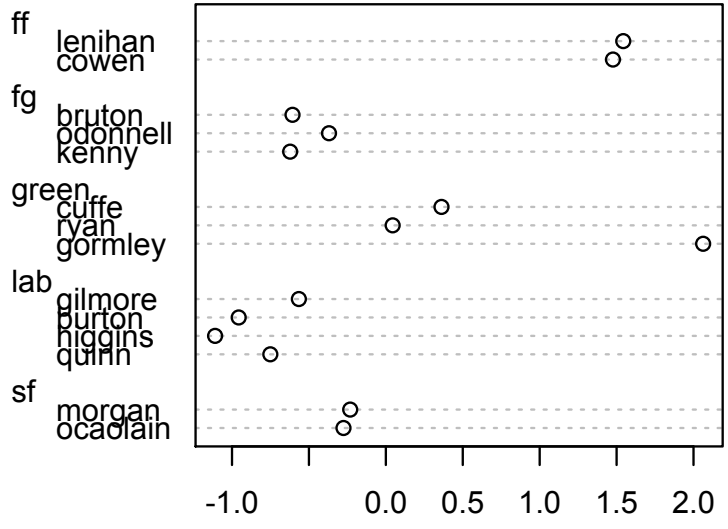
(Slapin and Proksch 2008)



Diagnosis 4: What happens if we include irrelevant text?



Wordscores LBG Position on Budget 2009



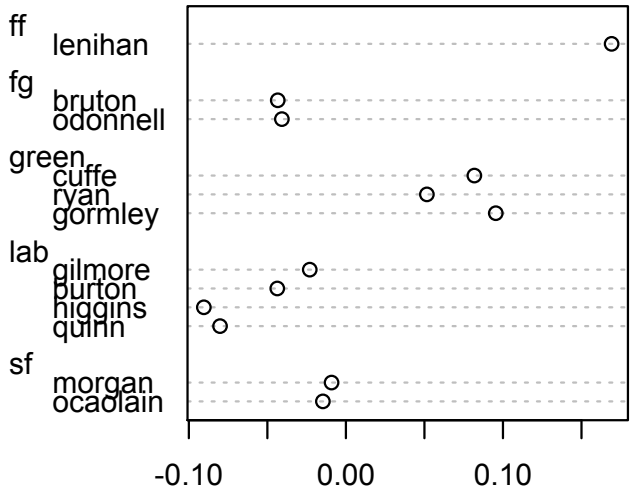
Normalized CA Position on Budget 2009

Diagnosis 4: What happens if we include irrelevant text?

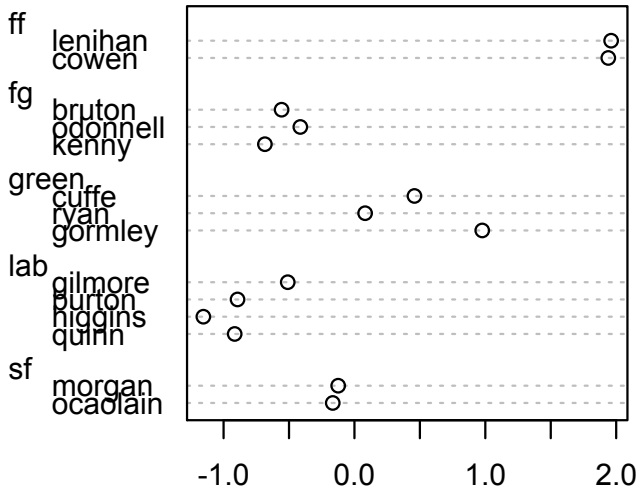
John Gormley: leader of the Green Party and Minister for the Environment, Heritage and Local Government

"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 ..."

Diagnosis 4: Without irrelevant text



Wordscores LBG Position on Budget 2009



Normalized CA Position on Budget 2009

The Way Forward

- Parametric Poisson model with variance parameter ("negative binomial" with parameter for over- or under-dispersion at the *word* level, could use CML
- Block Bootstrap resampling schemes
  - text unit blocks (sentences, paragraphs)
  - fixed length blocks
  - variable length blocks
  - could be overlapping or adjacent
- More detailed investigation of feasible methods for characterizing fundamental uncertainty from non-parametric scaling models (CA and others based on SVD)