Day 10: Parametric Models for Text Scaling

Kenneth Benoit

Essex Summer School 2012

July 20, 2012

When dependent variables are counts

- Many dependent variables of interest may be in the form of counts of discrete events— examples:
  - international wars or conflict events
  - the number of coups d'état
  - 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

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

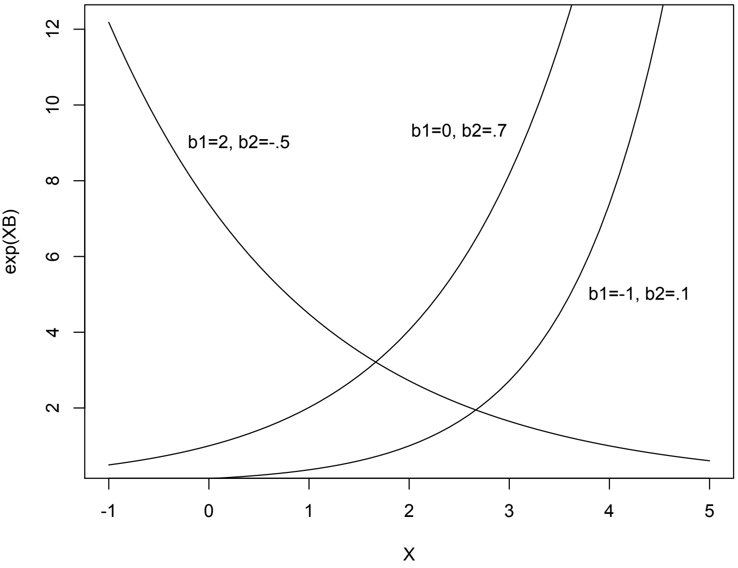
The Poisson distribution

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

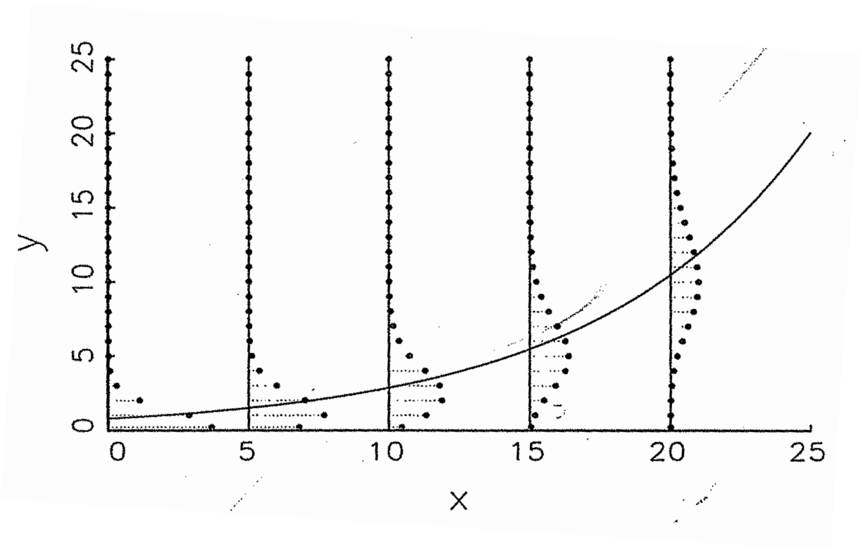
Systematic component

- $\lambda_i > 0$ is only bounded from below (unlike π_i)
- This implies that the effect cannot be linear
- Hence for the functional form we will use an exponential transformation
$$E(Y_i) = \lambda_i = e^{X_i \beta}$$
- Other possibilities exist, but this is by far the most common – indeed almost universally used – functional form for event count models

Exponential link function



Exponential link function



Likelihood for Poisson

$$\begin{aligned}
\mathrm{L}(\lambda|y) &= \prod_{i=1}^{N} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \\
\ln \mathrm{L}(\lambda|y) &= \sum_{i=1}^{N} \ln\left[\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}\right] \\
&= \sum_{i=1}^{N} \left\{\ln e^{-\lambda_i} + \ln(\lambda_i^{y_i}) + \ln\left(\frac{1}{y_i!}\right)\right\} \\
&= \sum_{i=1}^{N} \{-\lambda_i + y_i \ln(\lambda_i) - \ln(y_i!)\} \\
&= \sum_{i=1}^{N} \left\{-e^{X_i\beta} + y_i \ln e^{X_i\beta} - \ln y_i!\right\} \\
&\propto \sum_{i=1}^{N} \left\{-e^{X_i\beta} + y_i X_i\beta - dropped\right\} \\
\ln L(\beta|y) &\propto \sum_{i=1}^{N} \left\{X_i\beta y_i - e^{X_i\beta}\right\}
\end{aligned}$$

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

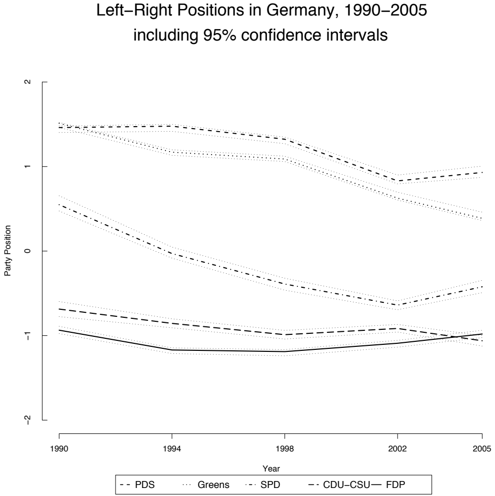
$$P(\text{vote}_1, \ldots, \text{vote}_K \mid \theta) \; = \; \prod_j P(\text{vote}_j \mid \theta)$$

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

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

Poisson scaling models for text

Poisson scaling models for text (aka "wordfish") is a statistical model for inferring policy positions θ from words



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)

How to estimate this model

Maximimum likelihood estimation using (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

The iterative (conditional) maximum likelihood estimation

Start by *guessing* the parameters

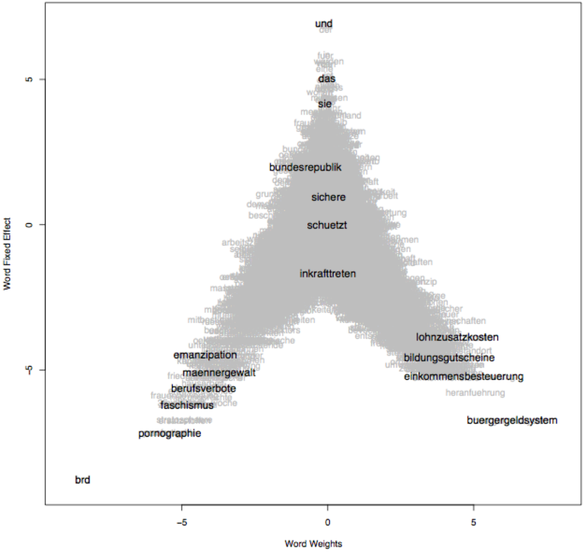
Algorithm:

- 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

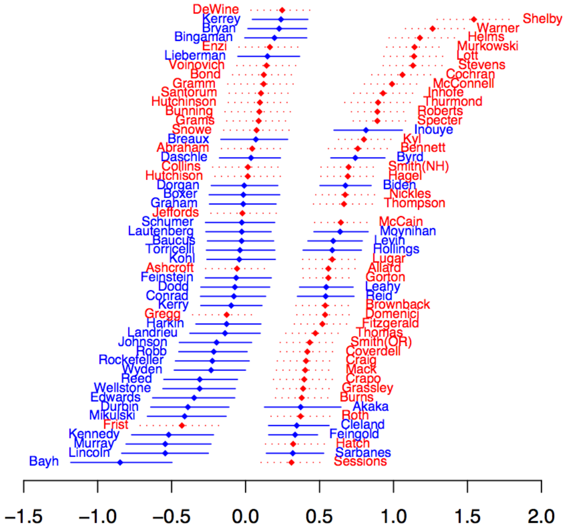
Frequency and informativeness

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

Wordscores and Wordfish as measurement models

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?

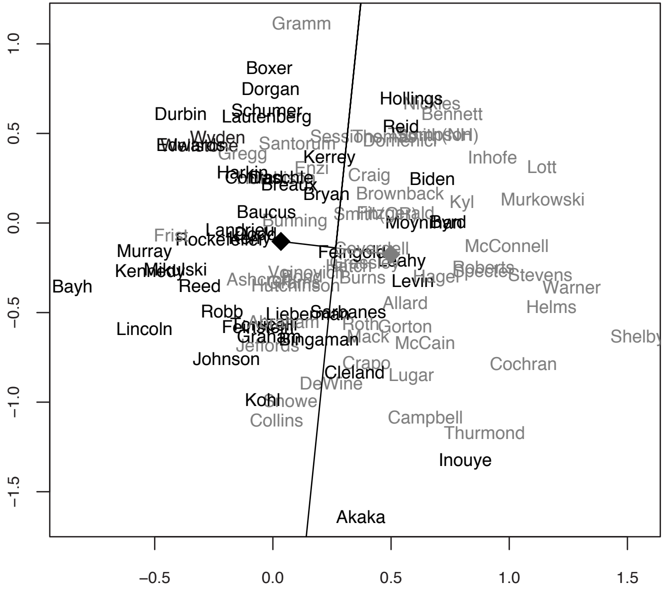
Dimensions

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

The hazards of ex-post interpretation illustrated



“Features” of the parametric scaling approach

- Standard (statistical) inference about parameters
- Uncertainty accounting for parameters
- Distributional assumptions are laid 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 to solve I: Conditional (non-)independence

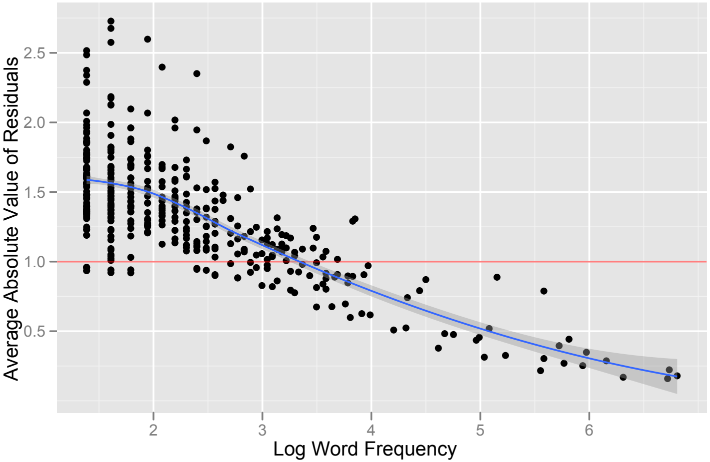
- Words occur in order
  In occur words order.
  Occur order words in.
  "No more training do you require. Already know you that which you need." (Yoda)
- Words occur in combinations
  "carbon tax" / "income tax" / "inhertiance tax" / "capital gains tax" /"bank tax"
- Sentences (and topics) occur in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



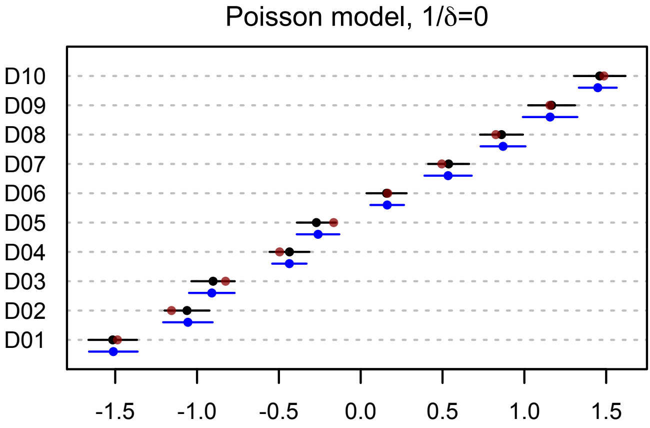
How to account for uncertainty?

- 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

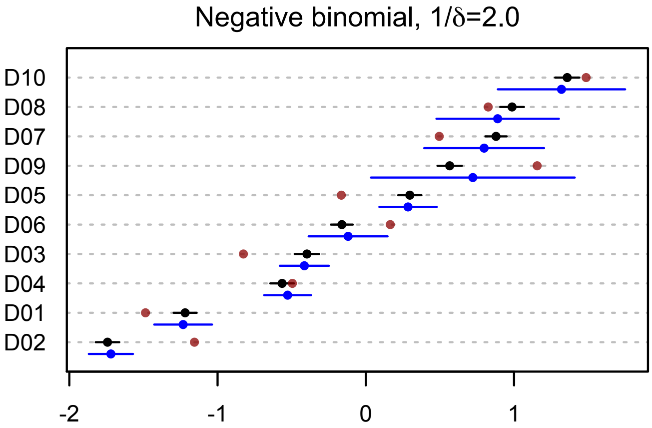
Steps forward

- 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

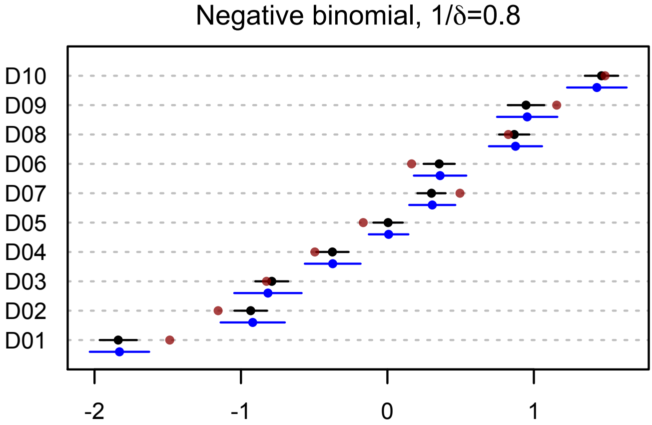
Diagnosis I: Estimations on simulated texts



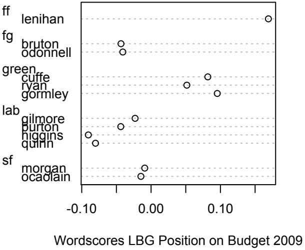
Diagnosis I: Estimations on simulated texts

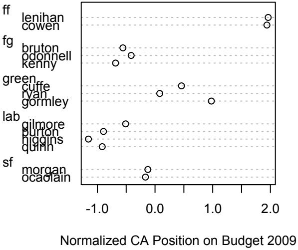


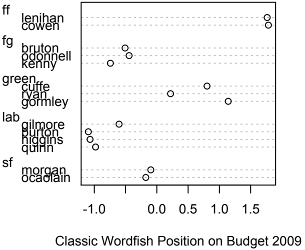
Diagnosis I: Estimations on simulated texts



Diagnosis 2: Irish Budget debate of 2009

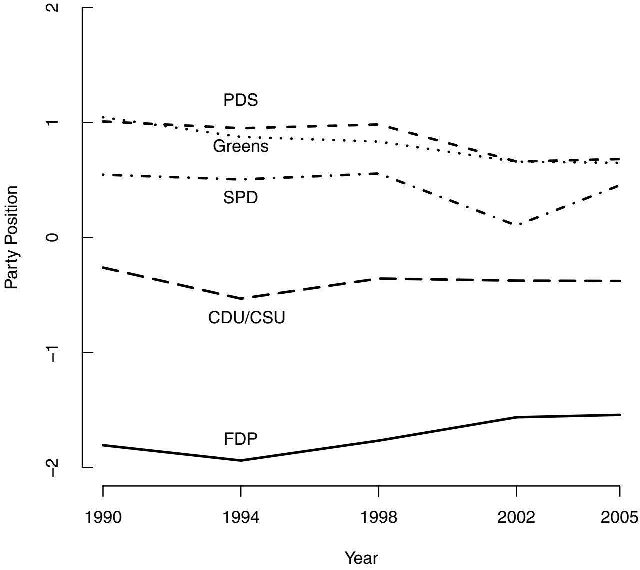




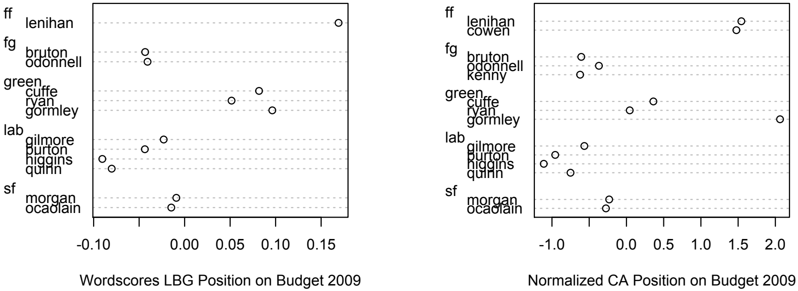


Diagnosis 3: German party manifestos (economic sections)

(Slapin and Proksch 2008)



Diagnosis 4: What happens if we include irrelevant text?

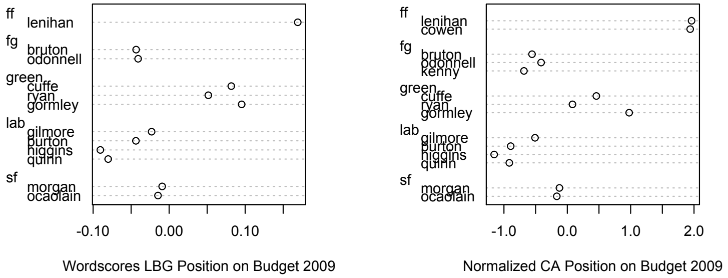


Diagnosis 4: What happens if we include irrelevant text?

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



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)

The Negative Binomial model

- Generalize the Poisson model to:

$$f_{nb}(y_i|\lambda_i, \sigma^2) \text{ where :}$$

  - σ^2 is the variability (a new parameter v. Poisson)
  - λ_i is the expected number of events for i
  - λ is the average of individual λ_is
- Here we have dropped Poisson assumption that $\lambda_i = \lambda \; \forall \; i$
- New assumption: Assume that λ_i is a random variable following a *gamma* distribution (takes on only non-negative numbers)
- For the NB model, $\text{Var}(Y_i) = \lambda_i \sigma^2$ for $\lambda_i > 0$ and $\sigma^2 > 0$

The Negative Binomial model cont.

- For the NB model, $\text{Var}(Y_i) = \lambda_i \sigma^2$ for $\lambda_i > 0$ and $\sigma^2 > 0$
- How to interpret σ^2 in the negative binomial
  - when $\sigma^2 = 1.0$, negative binomial $\equiv$ Poisson
  - when $\sigma^2 > 1$, then it means there is overdispersion in Y_i caused by correlated events, or heterogenous λ_i
  - when $\sigma^2 < 1$ it means something strange is going on
- When $\sigma^2 \neq 1$, then Poisson results will be inefficient and standard errors inconsistent
- Functional form: same as Poisson

$$\text{E}(y_i) = \lambda$$

- Variance of λ is now:

$$\text{Var}(y_i) = \lambda_i \sigma^2 = e^{X_i \beta} \sigma^2$$

Problems to Solve III: Integrating 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

Correspondence Analysis contd.

- There are also 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?

- 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

Methods of uncertainty accounting in text scaling

Uncertainty accounting	MCMC (multinomial+)	Conditional ML (Poisson)	SVD-based (CA)	Algorithmic (Wordscores)
Posterior sampling	$\surd$			
Analytical		$\surd$	??	?
Parametric bootstrap		$\surd$		
Non-parametric BS		$\surd$	?	$\surd$

Data-driven versus parametric methods

None	Analytic	Non-parametric bootstrap	Parametric bootstrap	Analytic

DATA-DRIVEN ⟷ **MODEL-DRIVEN**

Correspondence Analysis
Wordscores

Poisson scaling
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