Day 10: Additional Scaling Issues

Kenneth Benoit

Essex Summer School 2011

July 22, 2011

Problems to solve I: Conditional (non-)independence

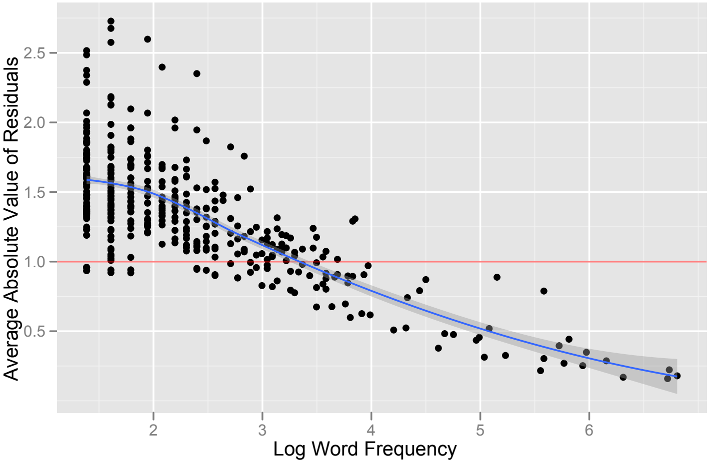
- Words occur in order
  In occur words order.
  Occur order words in.
  "No more training do you require. Already know you that which you need." (Yoda)
- Words occur in combinations
  "carbon tax" / "income tax" / "inhertiance tax" / "capital gains tax" /"bank tax"
- Sentences (and topics) occur in 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 $\mathrm{Var}(Y_{ij}) = \mathrm{E}(Y_{ij}) = \lambda_{ij}$
- 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)



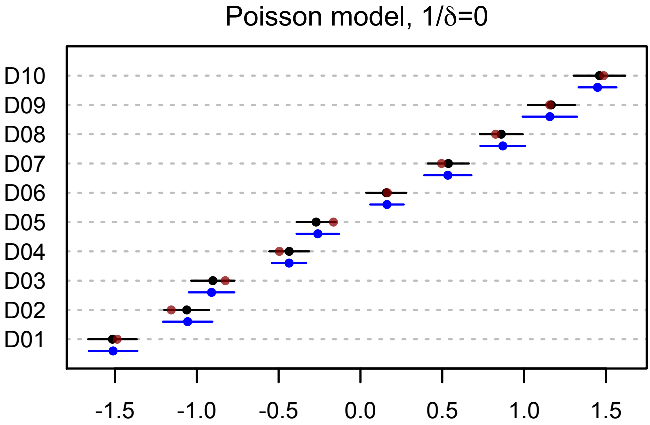
How to account for uncertainty?

- 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

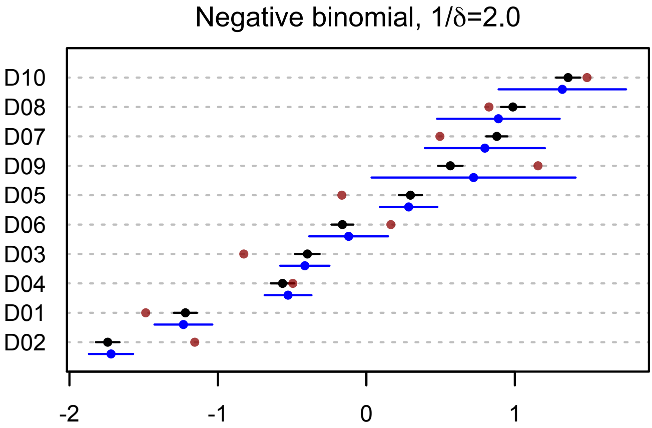
Steps forward

- 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

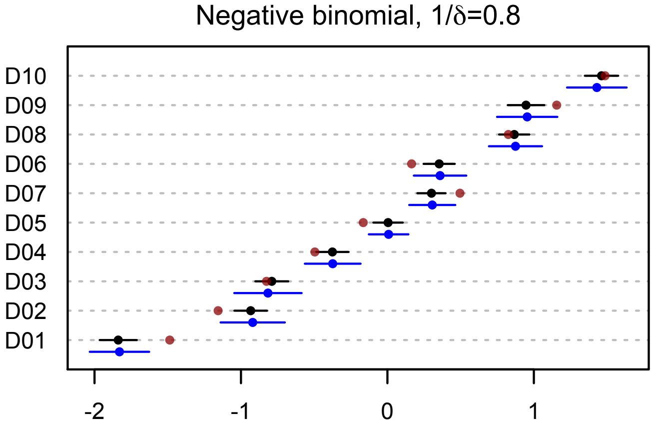
Diagnosis I: Estimations on simulated texts



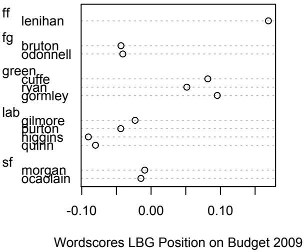
Diagnosis I: Estimations on simulated texts

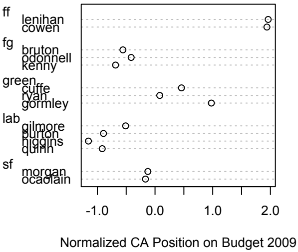


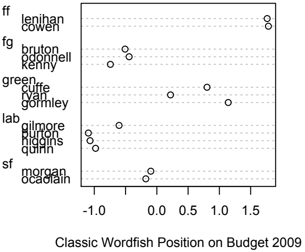
Diagnosis I: Estimations on simulated texts



Diagnosis 2: Irish Budget debate of 2009

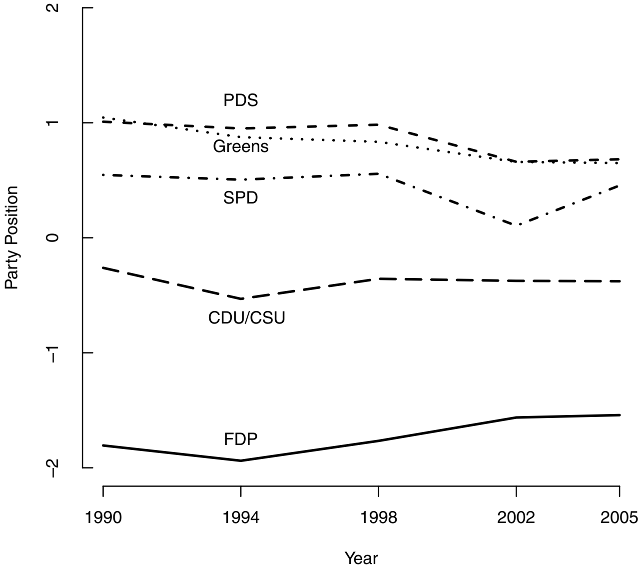




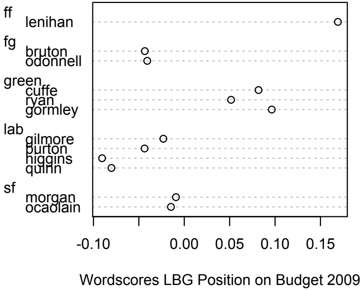


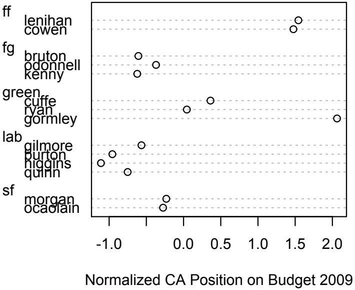
Diagnosis 3: German party manifestos (economic sections)

(Slapin and Proksch 2008)



Diagnosis 4: What happens if we include irrelevant text?



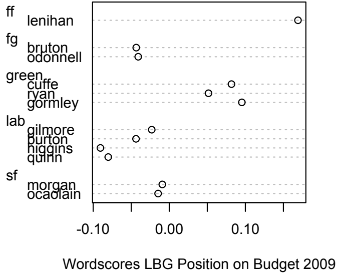


Diagnosis 4: What happens if we include irrelevant text?

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

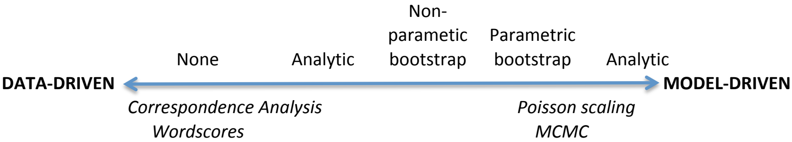
How to account for uncertainty?

- 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		$\surd$	?	$\surd$

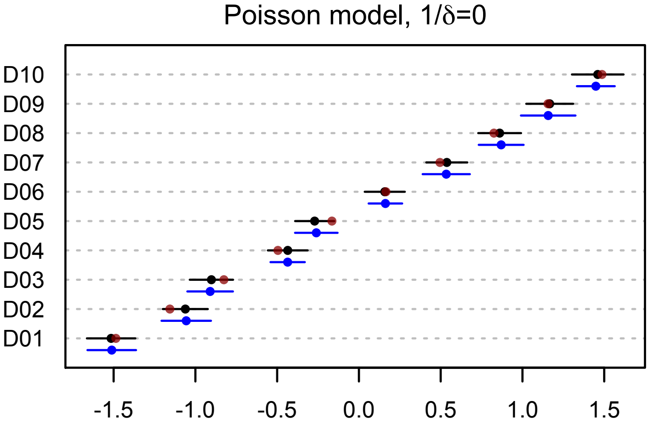
Data-driven versus parametric methods



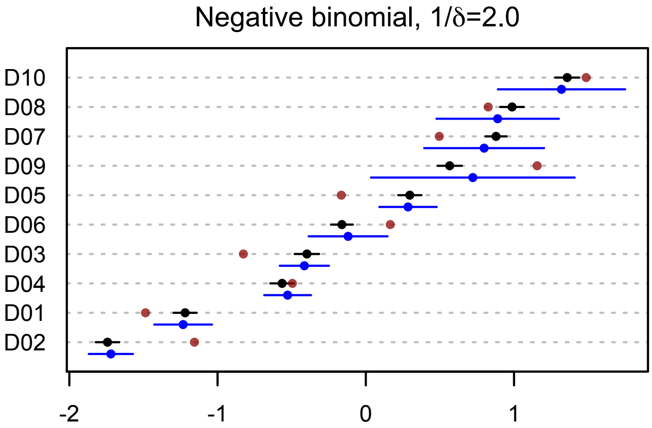
Steps forward

- 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

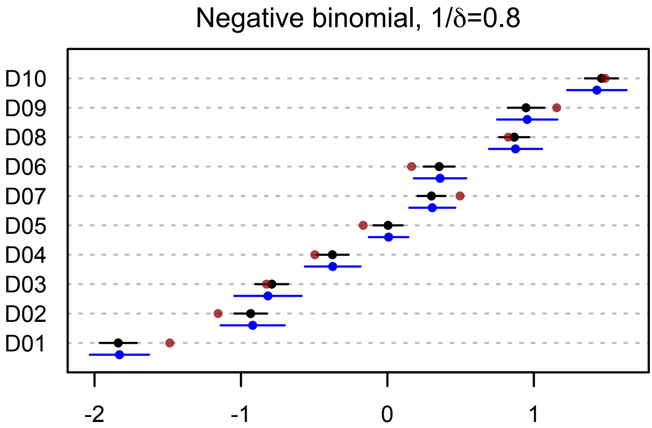
Diagnosis I: Estimations on simulated texts



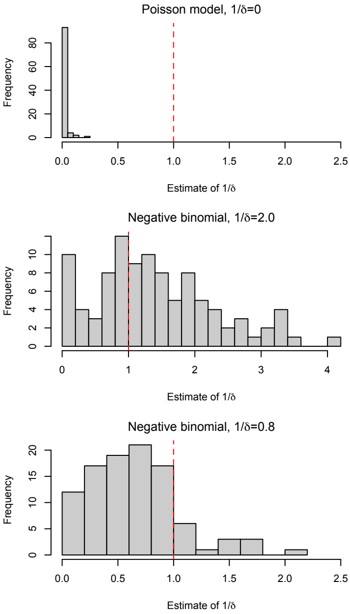
Diagnosis I: Estimations on simulated texts



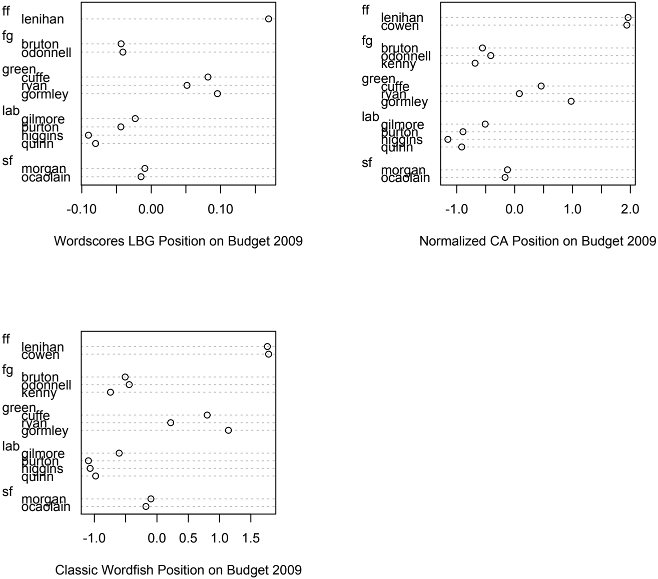
Diagnosis I: Estimations on simulated texts



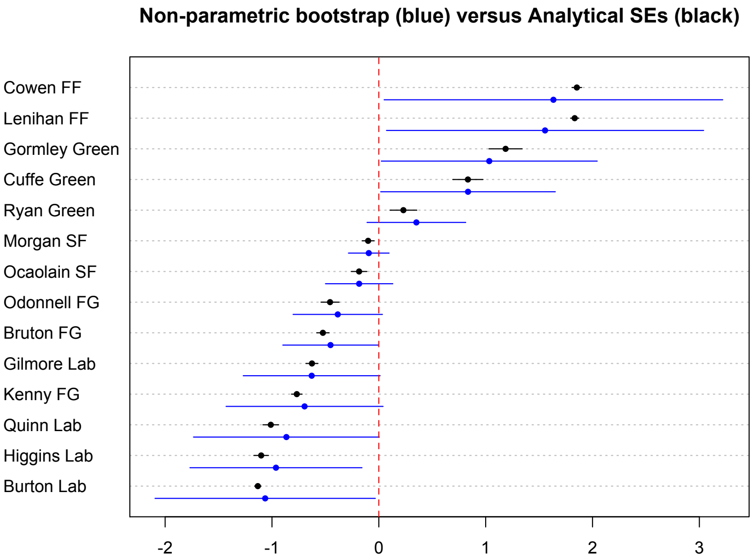
Simulated text results



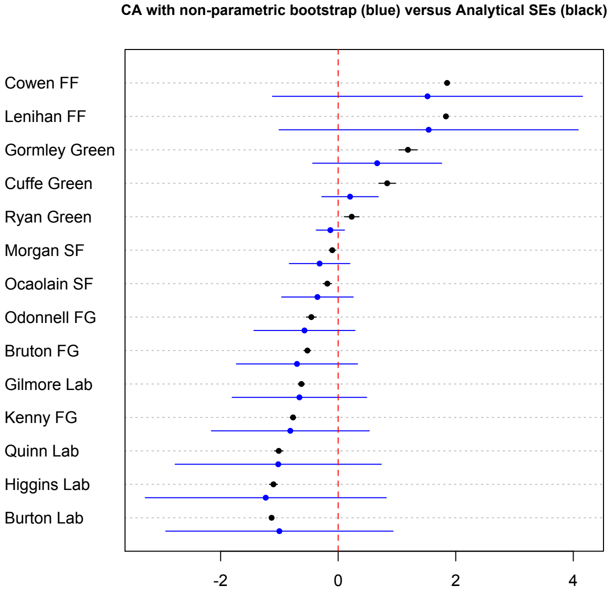
Diagnosis 2: Irish Budget debate of 2009



Budget debates: Analytical SEs

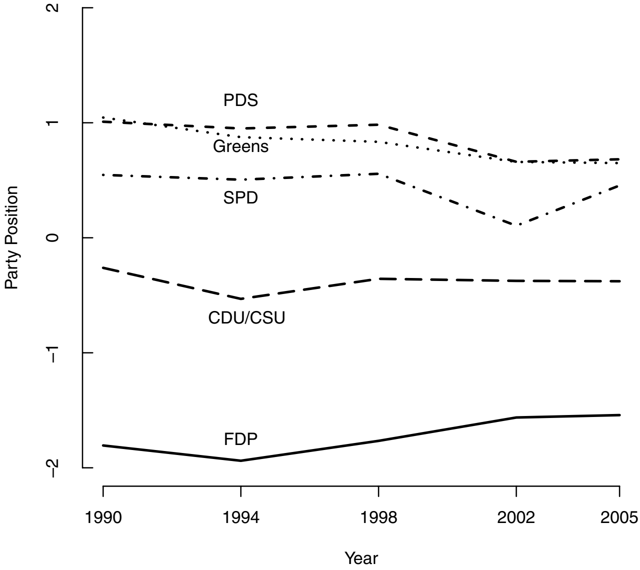


Budget debates: Bootstrapped SEs on CA

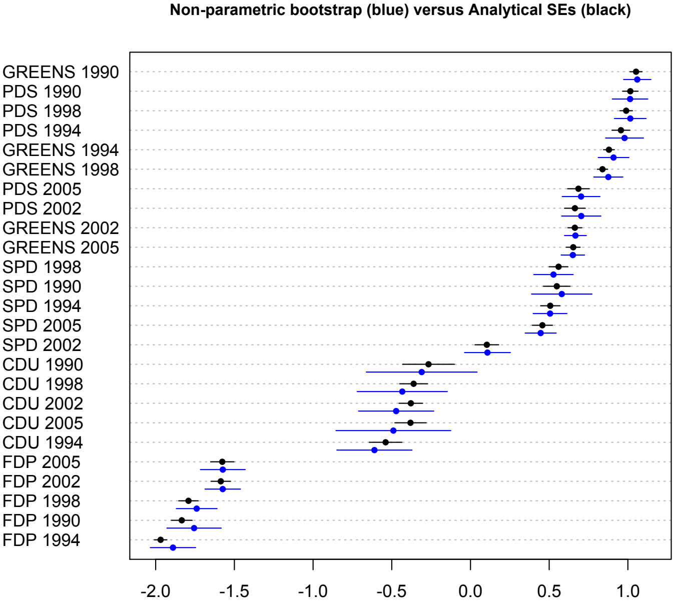


Diagnosis 3: German party manifestos (economic sections)

(Slapin and Proksch 2008)



German manifestos: Poisson Scaled Analytical SEs



German manifestos: Non-parametric bootstrap on CA