Unsupervised Methods for Scaling Texts

MY560 Workshop in Advanced Quantitative Analysis

Kenneth Benoit

June 11, 2013

Supervised v. unsupervised methods compared

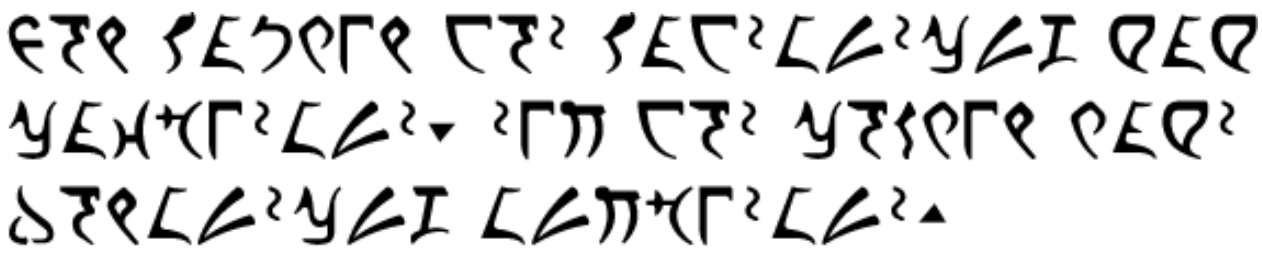
- The goal (in text analysis) is to differentiate *documents* from one another, treating them as "bags of words"
- Different approaches:
  - *Supervised methods* require a training set that exmplify constrasting classes, identified by the researcher
  - *Unsupervised methods* scale documents based on 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
- 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)

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
- Could potentially work on texts like this:



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

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— 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

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

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

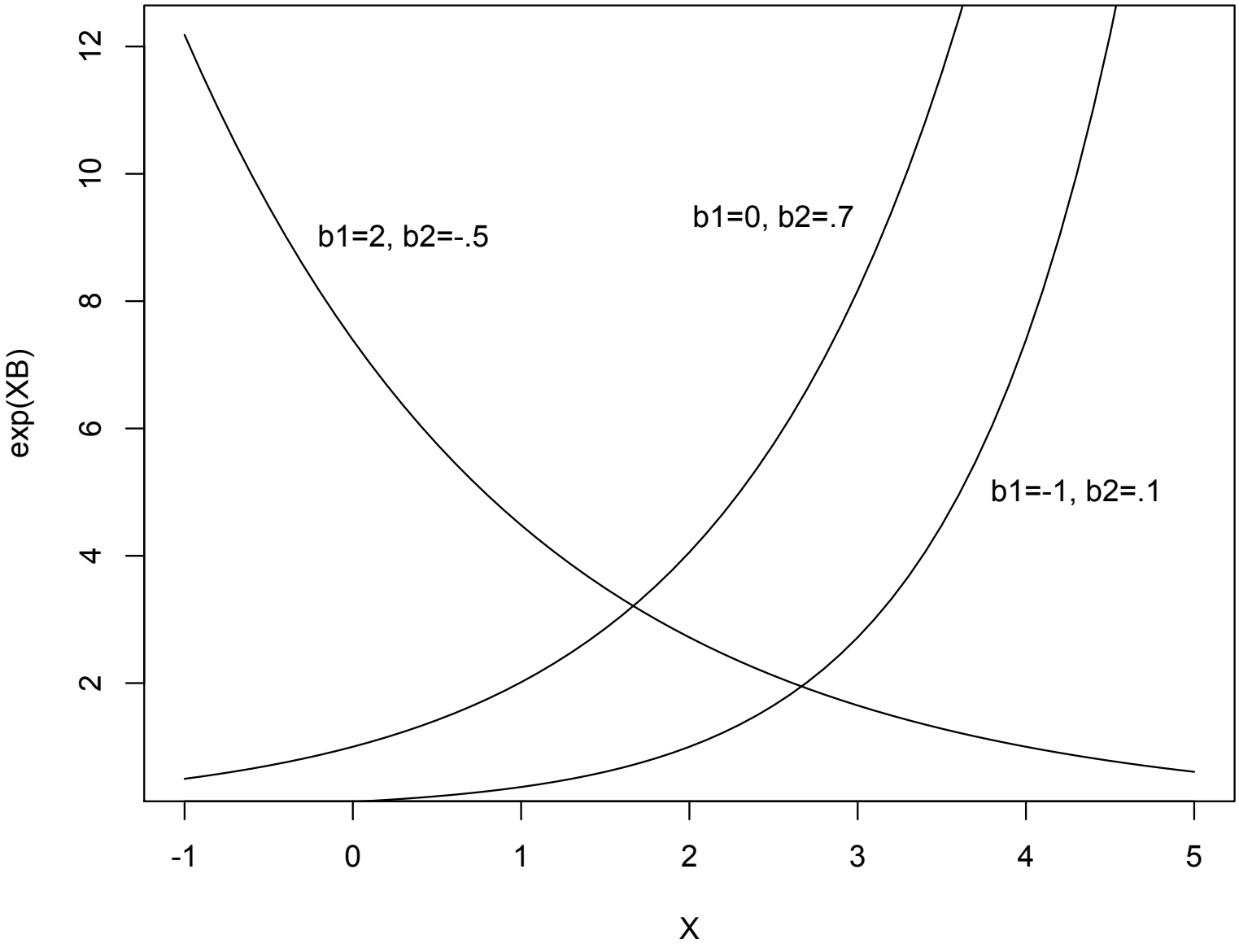
Systematic component

- $\lambda_i > 0$ 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

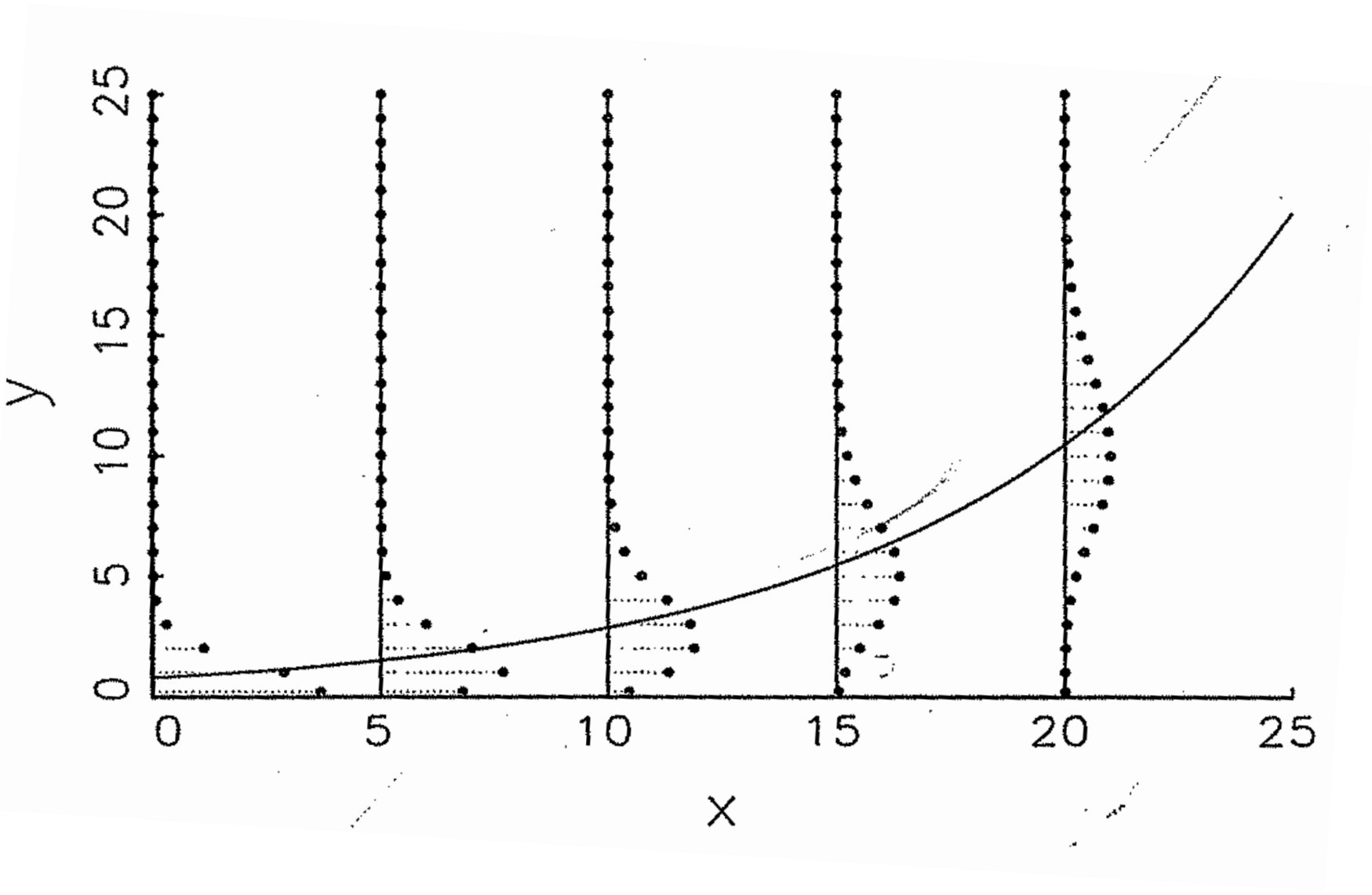
$$\mathrm{E}(Y_i) = \lambda_i = e^{x_i \beta}$$

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

Exponential link function



Exponential link function



Likelihood for Poisson

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

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

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

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

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!

"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

Assumptions of the model

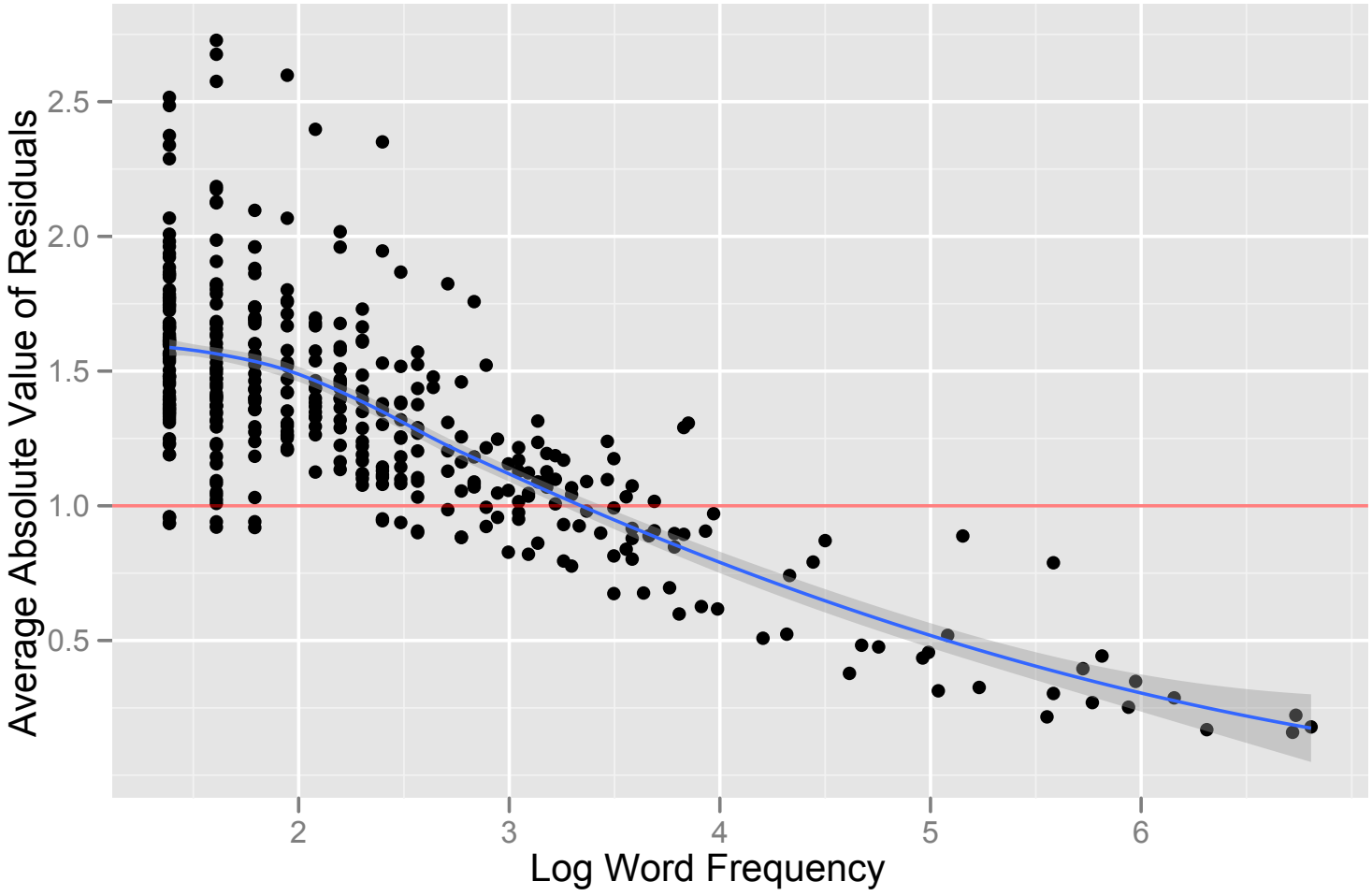
- Words occur in order
  In occur words order.
  Occur order words in.
  "No more training do you require. Already know you that which you need." (Yoda)
- Words occur in combinations
  "carbon tax" / "income tax" / "inhertiance tax" / "capital gains tax" /"bank tax"
- Sentences (and topics) occur in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

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

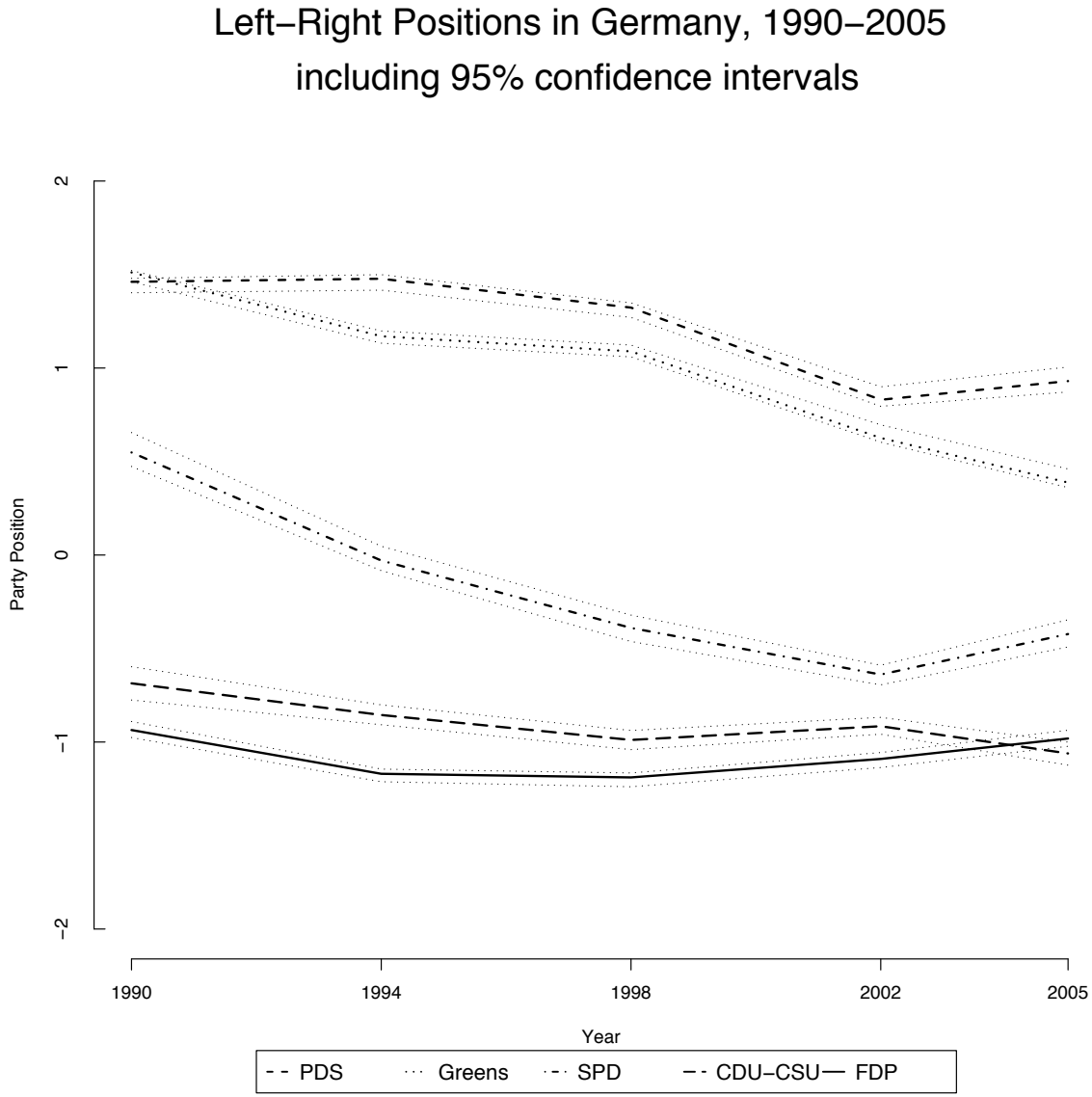
(from Slapin and Proksch 2008)



How to account for uncertainty?

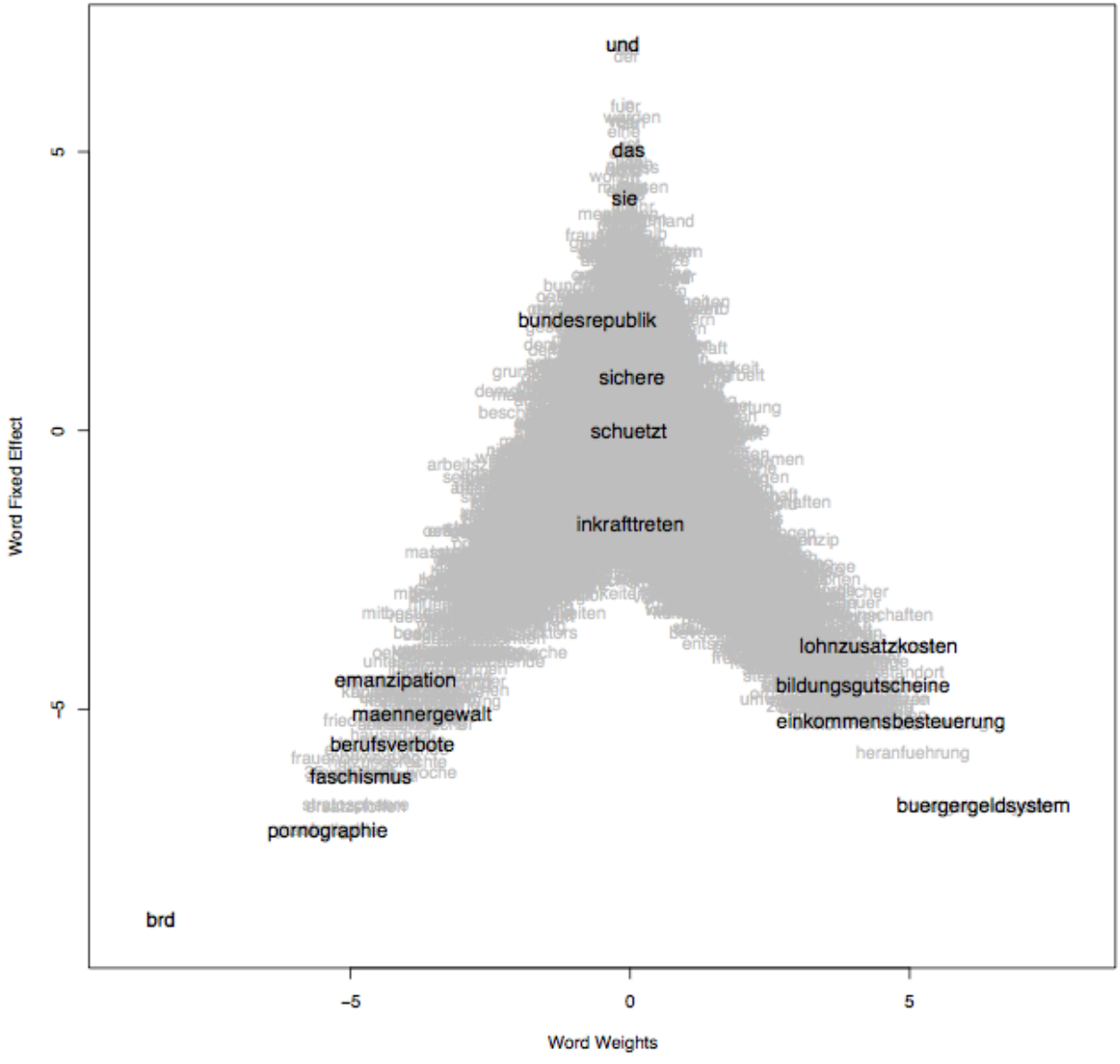
- Don't. (SVD-like methods, e.g. correspondence analysis)
- Analytical derivatives
- Parametric bootstrapping (Slapin and Proksch, Lewis and Poole)
- Non-parametric bootstrapping
- (and yes of course) Posterior sampling from MCMC

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



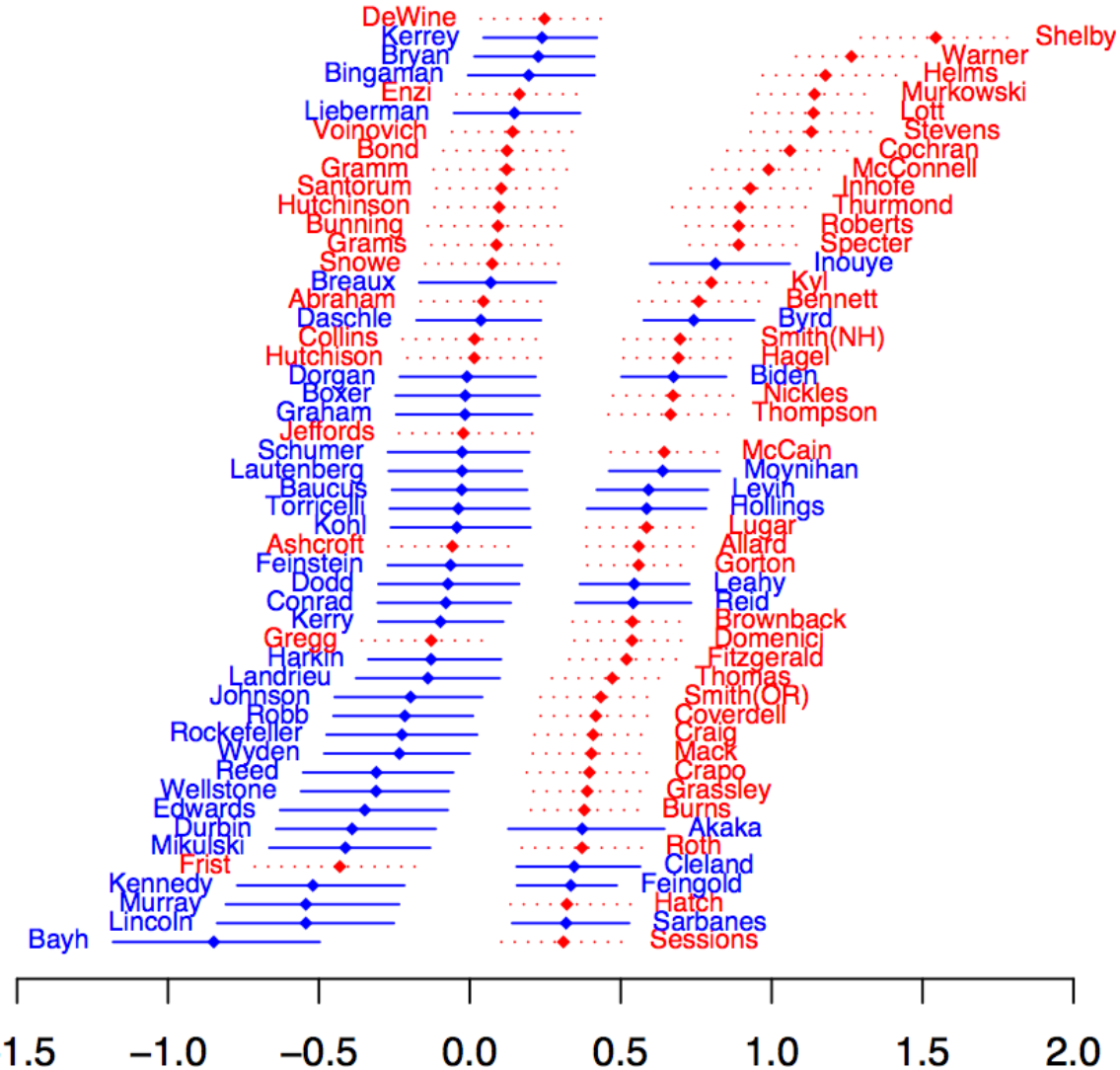
Frequency and informativeness

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:



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?

Interpreting multiple dimensions

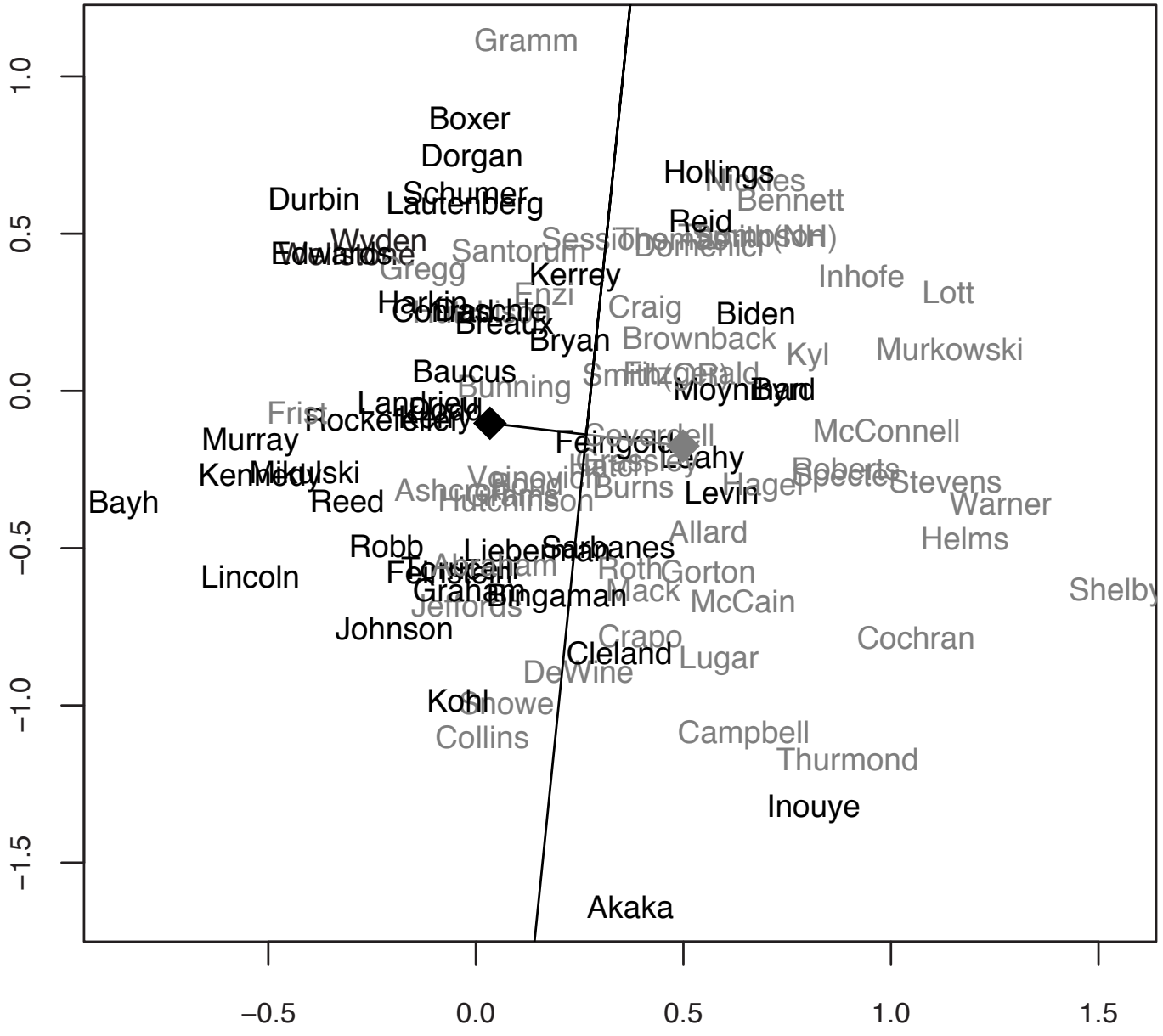
To get one dimension for each policy area, split up the document by hand and use the subparts as documents (the Slapin and Proksch method)
There is currently *no* implementation of Wordscores or Wordfish that extracts two or more dimensions at once

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

Interpreting scaled dimensions

- How to interpret $\hat{\theta}$s substantively? *assert...*
- Billy Joe Jimbob's "Make a Wish Foundation"

The hazards of ex-post interpretation illustrated

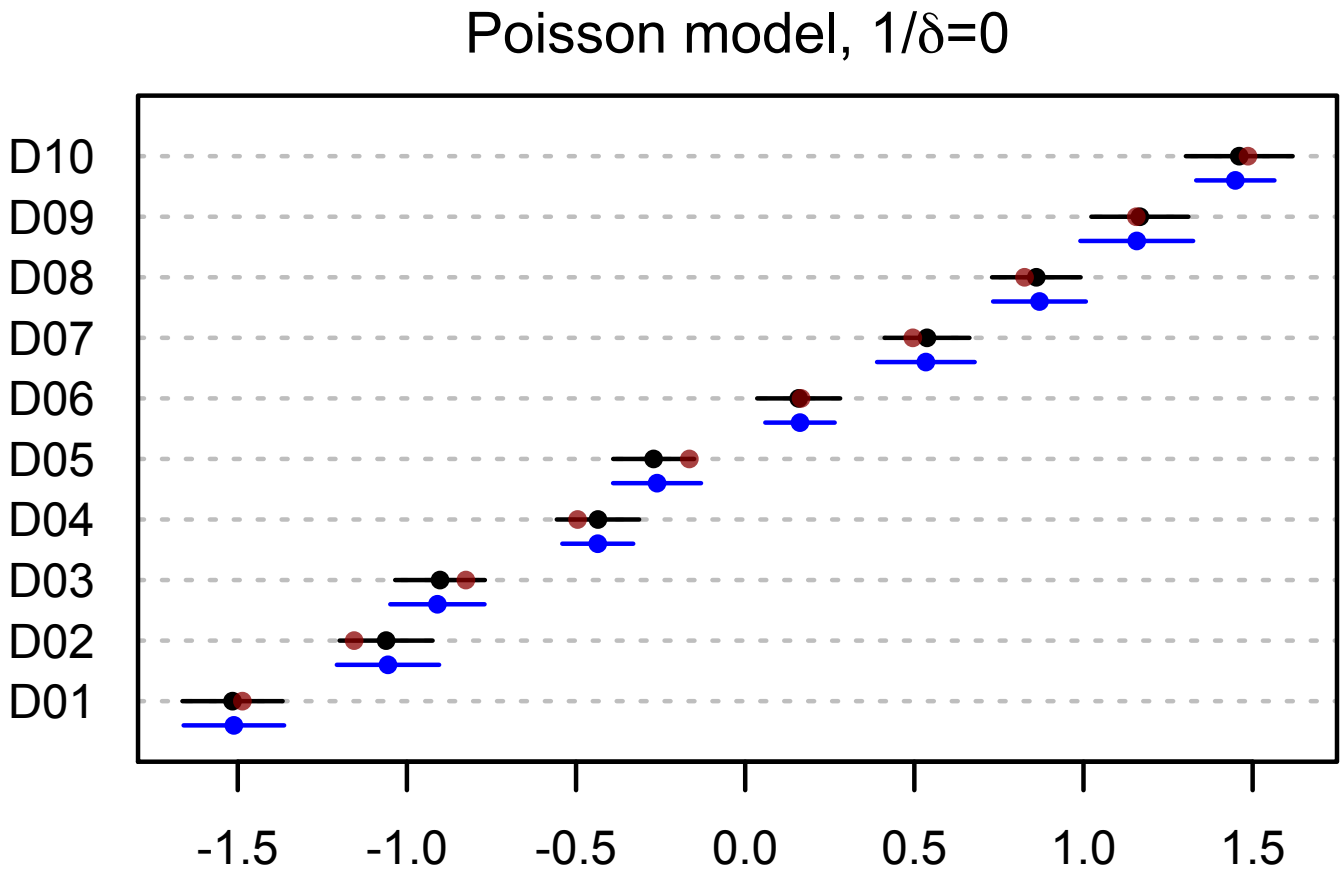


(Billy Joe Jimbob)

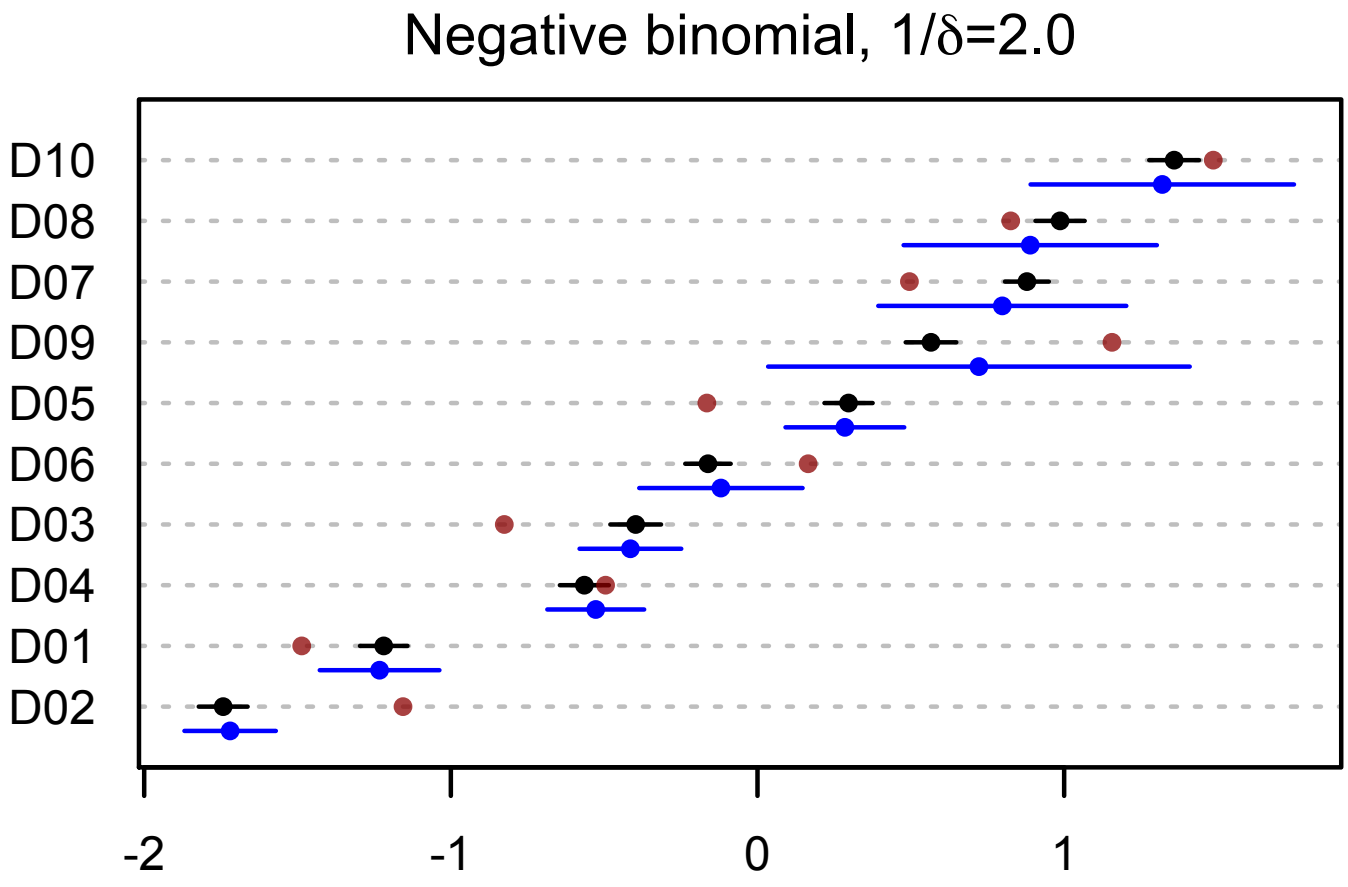
Interpreting scaled dimensions

- 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

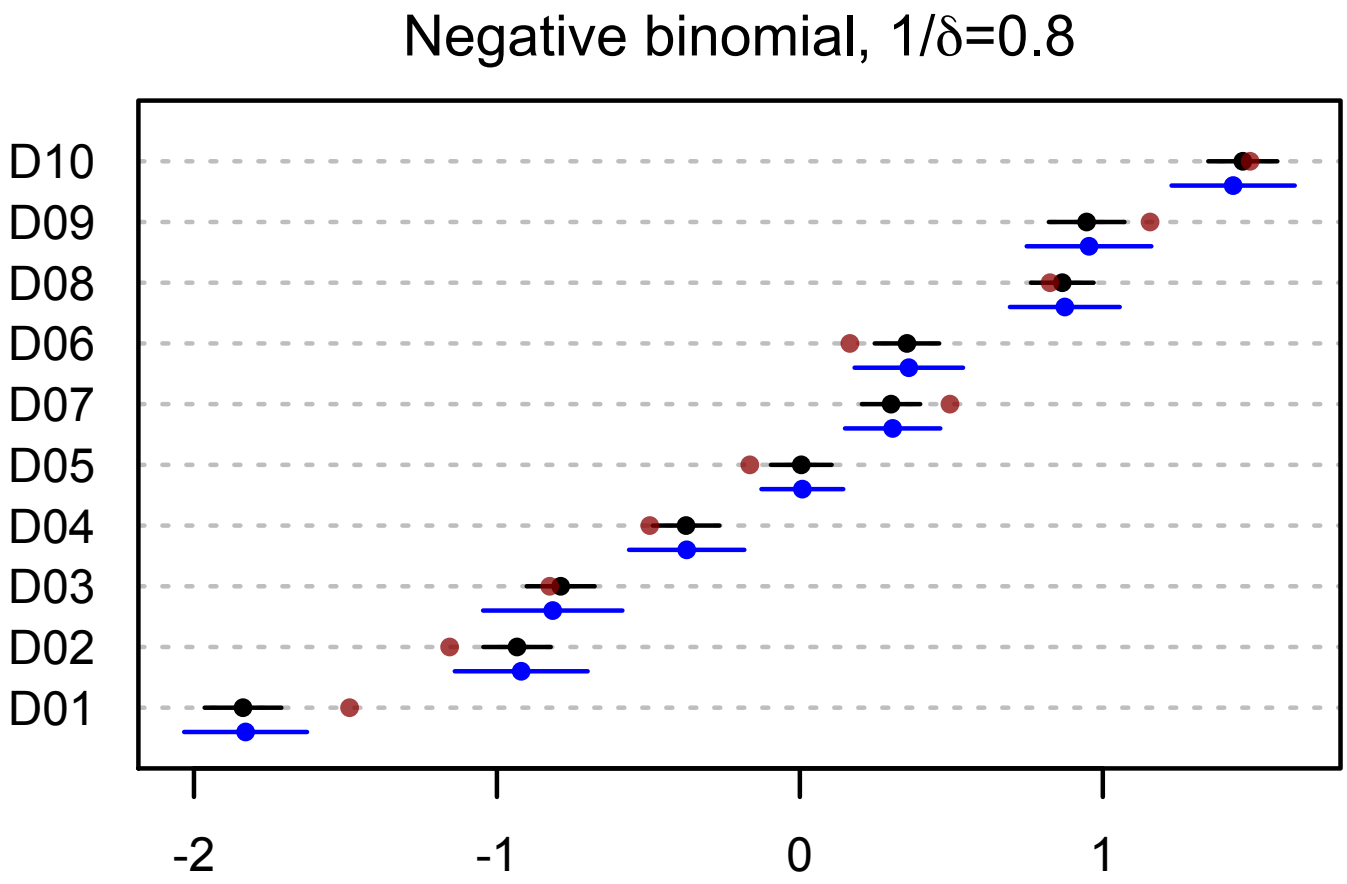
Diagnosis I: Estimations on simulated texts



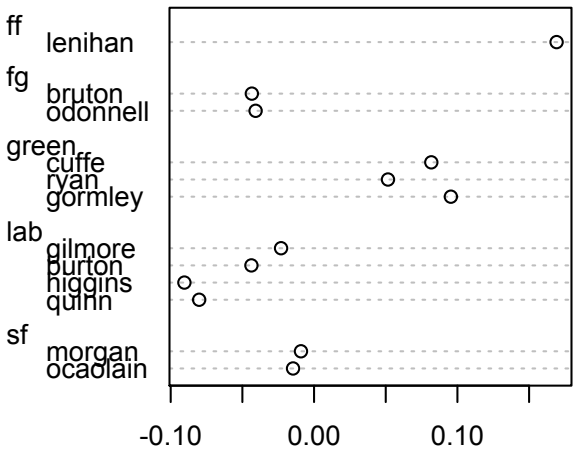
Diagnosis I: Estimations on simulated texts



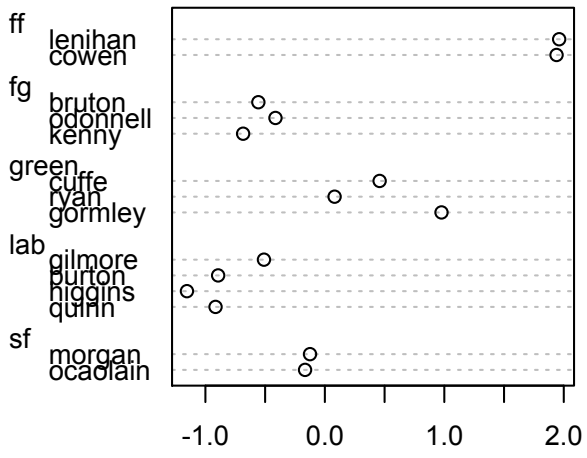
Diagnosis I: Estimations on simulated texts



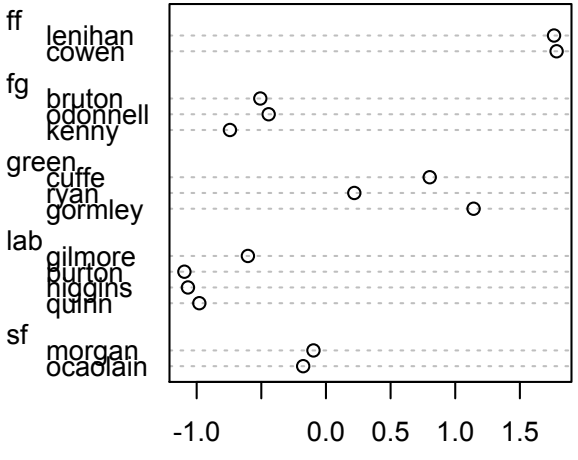
Diagnosis 2: Irish Budget debate of 2009



Wordscores LBG Position on Budget 2009



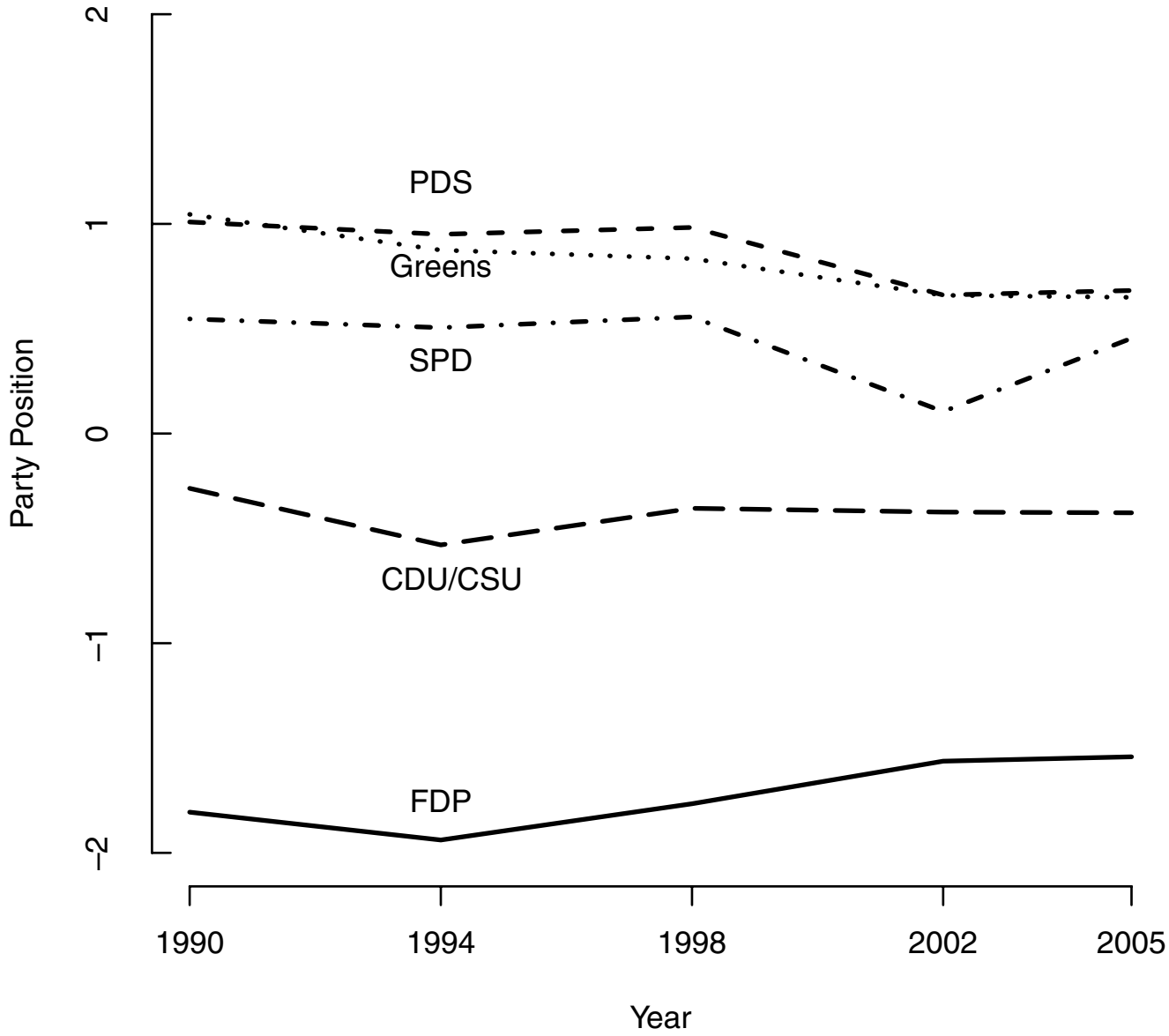
Normalized CA Position on Budget 2009



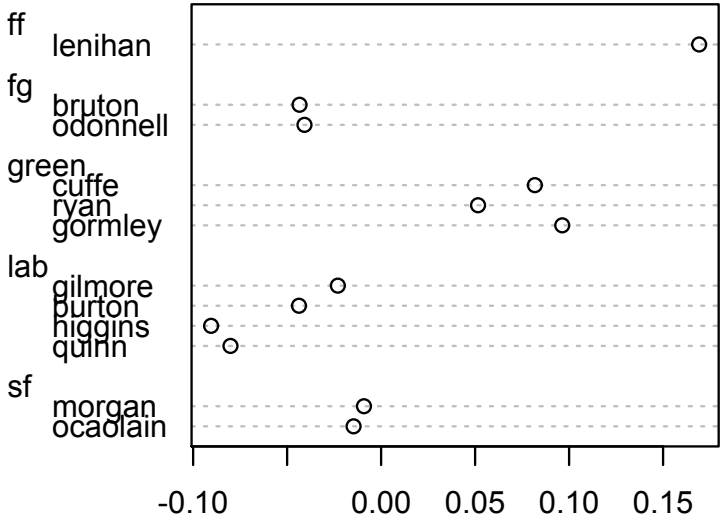
Classic Wordfish Position on Budget 2009

Diagnosis 3: German party manifestos (economic sections)

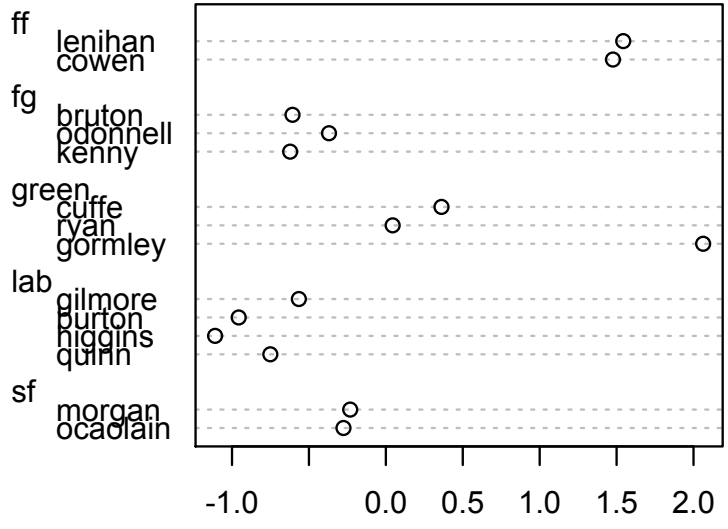
(Slapin and Proksch 2008)



Diagnosis 4: What happens if we include irrelevant text?



Wordscores LBG Position on Budget 2009



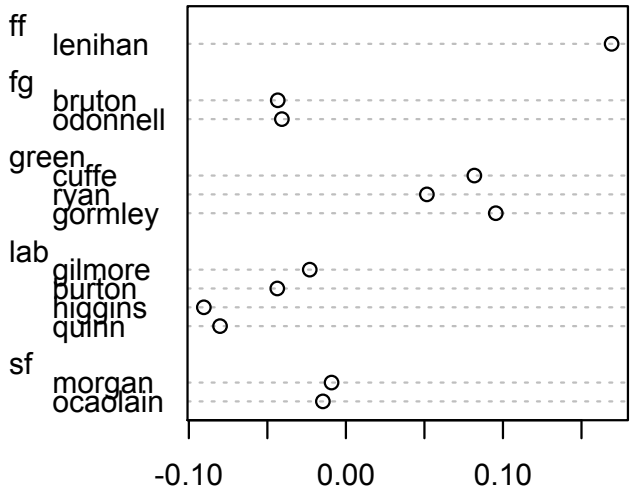
Normalized CA Position on Budget 2009

Diagnosis 4: What happens if we include irrelevant text?

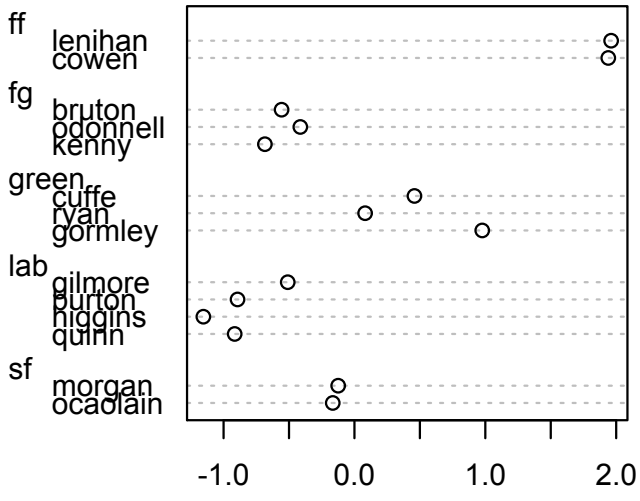
John Gormley: leader of the Green Party and Minister for the Environment, Heritage and Local Government

"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

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:

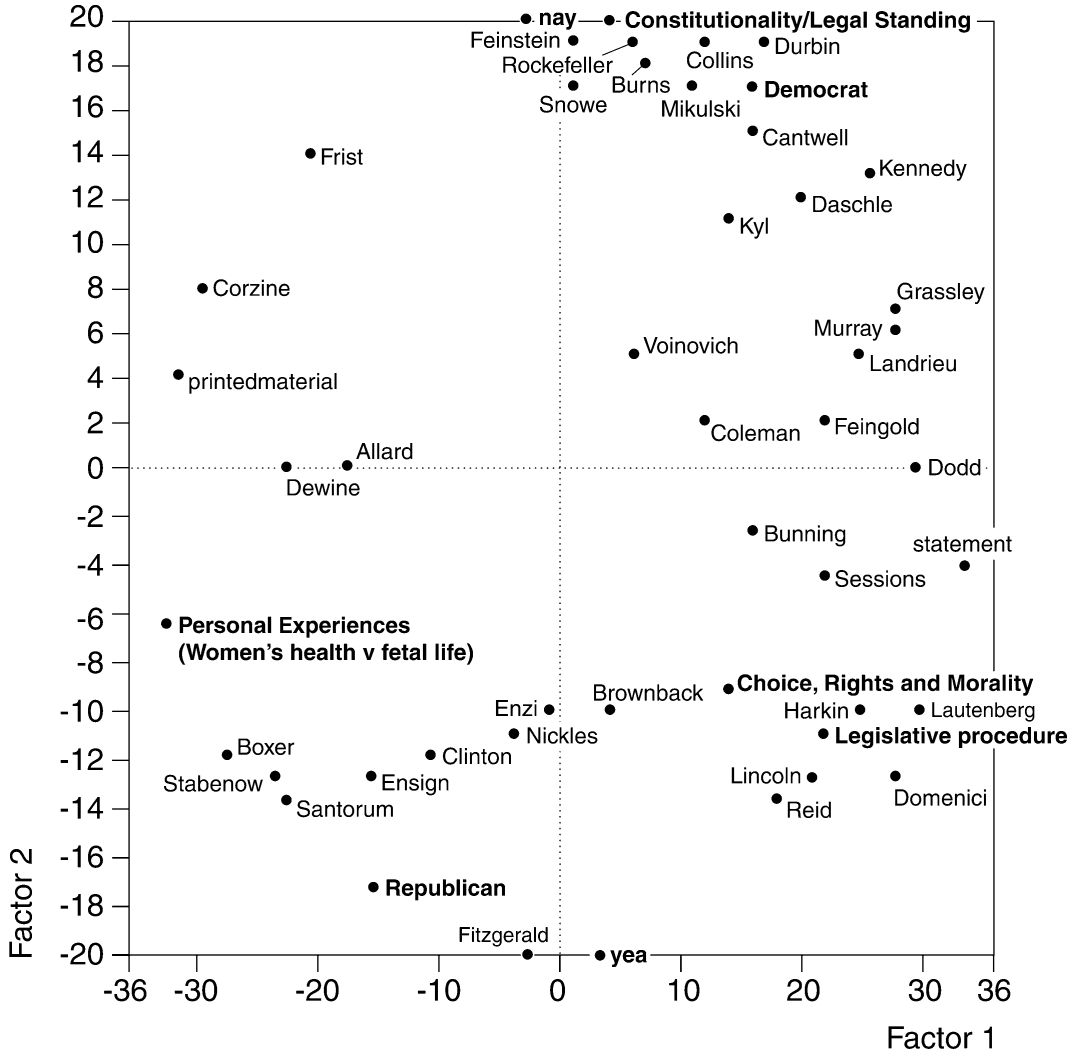
$$\underset{i \times j}{\mathbf{X}} = \underset{i \times k}{\mathbf{U}} \; \underset{k \times k}{\mathbf{d}} \; \underset{j \times k}{\mathbf{V}'} \tag{1}$$

- 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 $V_j \sqrt{f_{..}/f_{j.}}$)

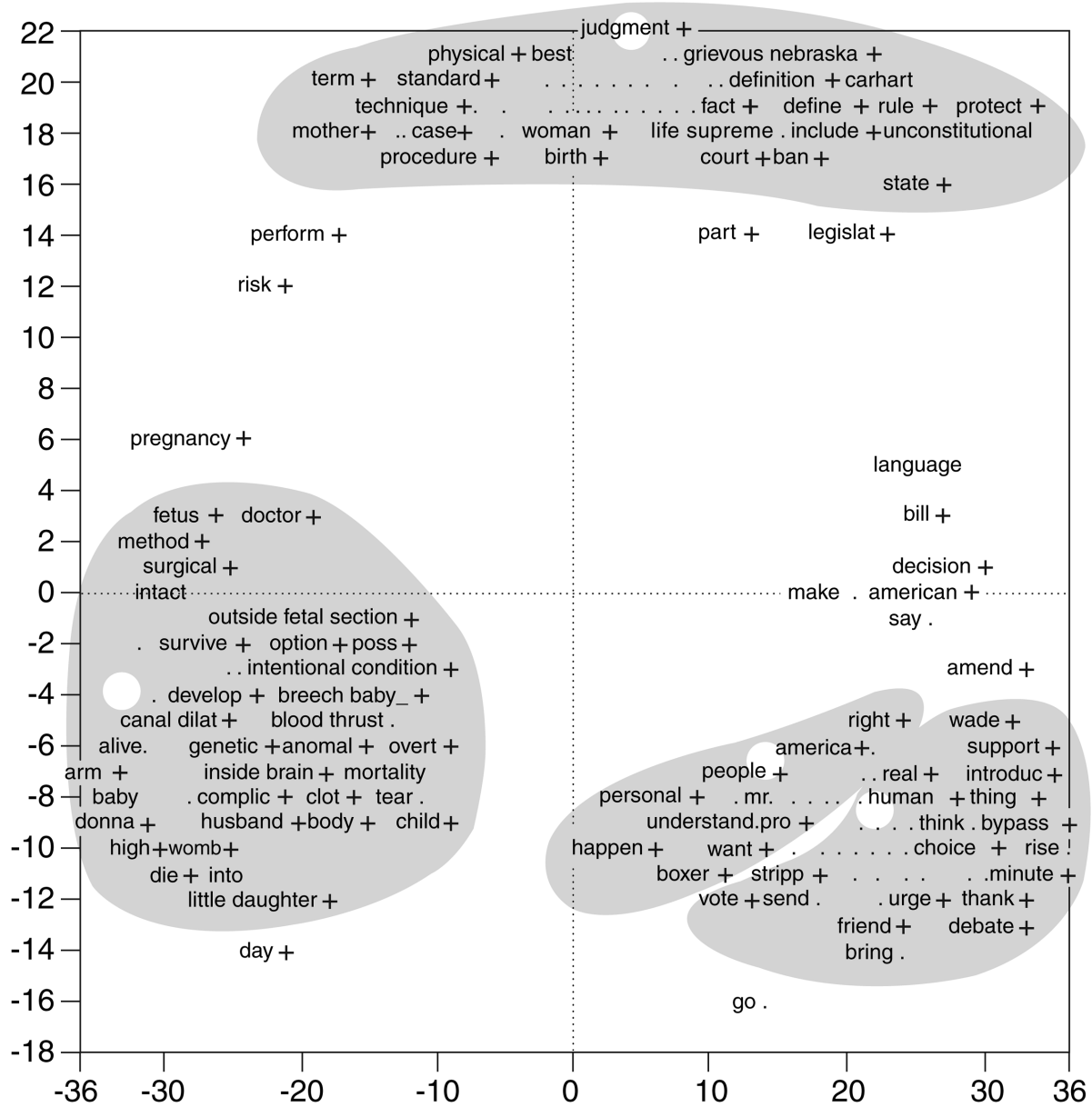
Example: Schonhardt-Bailey (2008) - speakers



	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?

- 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		$\surd$	?	$\surd$