Unsupervised scaling models

Kenneth Benoit & Pablo Barberá

MY 459: Quantitative Text Analysis

February 26th, 2018

Course website: lse-my459.github.io

Outline

- Unsupervised scaling of documents
  - Basics of supervised scaling methods
  - Parametric scaling models: Wordfish and Wordshoal
  - Non-parametric scaling methods: correspondence analysis
  - Practical aspects: computing uncertainty, multiple dimensions, sensitivity to inclusion of irrelevant text
- Unsupervised scaling of features
  - Word embeddings
  - Examples with word2vec

Unsupervised methods scale distance

- Text gets converted into a quantitative matrix of features
  - words, typically
  - could be dictionary entries, or parts of speech
- Documents are scaled based on similarity/distance in feature use
- Fundamental problem: distance on which scale?
  - Ideally, something we care about, e.g. policy positions, ideology, preferences, sentiment
  - But often other dimensions (language, rhetoric style, authorship) are more predictive
- First dimension in unsupervised scaling will capture main source of variation, whatever that is
- Unlike supervised models, validation comes after estimating the model

Unsupervised scaling methods

Two main approaches

- 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
  - e.g. Wordfish (Slapin and Proksch 2008) and Wordshoal (Lauderdale and Herzog 2016)

- Non-parametric methods typically based on the Singular Value Decomposition of a matrix
  - correspondence analysis
  - factor analysis
  - other (multi)dimensional scaling methods

Outline

- Unsupervised scaling of documents
  - Basics of supervised scaling methods
  - Parametric scaling models: Wordfish and Wordshoal
  - Non-parametric scaling methods: correspondence analysis
  - Practical aspects: computing uncertainty, multiple dimensions, sensitivity to inclusion of irrelevant text
- Unsupervised scaling of features
  - Word embeddings
  - Examples with word2vec

Wordfish (Slapin and Proksch 2008)

- Goal: unsupervised scaling of ideological positions
- The frequency with which politician i uses word k is drawn from a Poisson distribution:

$$w_{ik} \sim \mathrm{Poisson}(\lambda_{ik})$$
$$\lambda_{ik} = exp(\alpha_i + \psi_k + \beta_k \times \theta_i)$$

- with latent parameters:

  α_i is "loquaciousness" of politician i
  ψ_k is frequency of word k
  β_k is discrimination parameter of word k
  θ_i is the politician's ideological position

- Key intuition: controlling for document length and word frequency, words with negative β_k will tend to be used more often by politicians with negative θ_i (and vice versa)

Wordfish (Slapin and Proksch 2008)

Why Poisson?

- Poisson-distributed variables are bounded between $(0, \infty)$ and take on only discrete values $0, 1, 2, \ldots, \infty$
- Exponential transformation: word counts are function of log document length and word frequency

$$
\begin{aligned}
w_{ik} &\sim \mathrm{Poisson}(\lambda_{ik}) \\
\lambda_{ik} &= exp(\alpha_i + \psi_k + \beta_k \times \theta_i) \\
log(\lambda_{ik}) &= \alpha_i + \psi_k + \beta_k \times \theta_i
\end{aligned}
$$

How to estimate this model

Conditional maximum 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 `quanteda` package as `textmodel_wordfish`

An alternative is MCMC with a Bayesian formulation or variational inference using an Expectation-Maximization algorithm (Imai et al 2016)

Conditional maximum likelihood for wordfish

Start by guessing the parameters (some guesses are better than others, e.g. SVD)

Algorithm:

1. Assume the current legislator parameters are correct and fit as a Poisson regression model
2. Assume the current word parameters are correct and fit as a Poisson regression model
3. Normalize θs to mean 0 and variance 1

Iterate until convergence (change in values is below a certain threshold)

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

“Features” of the parametric scaling approach

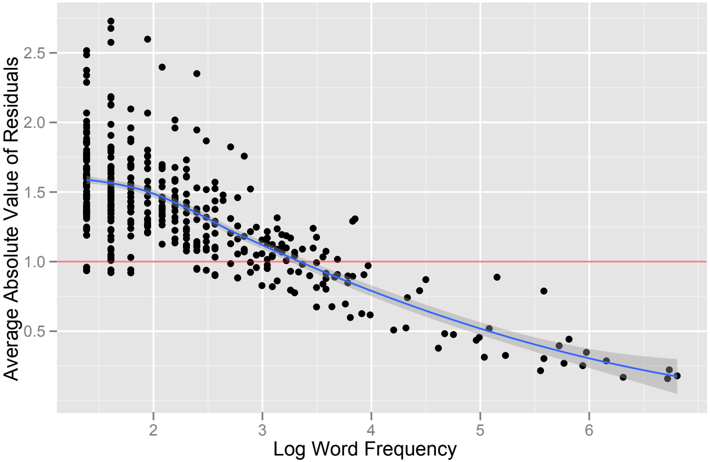
- Standard (statistical) inference about parameters
- Uncertainty accounting for parameters
- Distributional assumptions are made explicit (as part of the data generating process motivating the choice of stochastic distribution)
  - *conditional independence*
  - *stochastic process* (e.g. $\mathrm{E}(Y_{ij}) = \mathrm{Var}(Y_{ij}) = \lambda_{ij}$)
- 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

- Violations of conditional independence:
  - Words occur in sequence (serial correlation)
  - Words occur in combinations (e.g. as collocations) "carbon tax" / "income tax" / "inheritance tax" / "capital gains tax" /"bank tax"
  - Legislative speech uses rhetoric that contains frequent synonyms and repetition for emphasis (e.g. "Yes we can!")
- Heteroskedastic errors (variance not constant and equal to mean):
  - 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)

Overdispersion in German manifesto data

(data taken from Slapin and Proksch 2008)



One solution to model overdispersion

Negative binomial model (Lo, Proksch, and Slapin 2014):

$$w_{ik} \sim \mathrm{NB}\left(r, \frac{\lambda_{ik}}{\lambda_{ik} + r_i}\right)$$
$$\lambda_{ik} = exp(\alpha_i + \psi_k + \beta_k \times \theta_i)$$

where r_i is a variance inflation parameter that varies across documents.

It can have a substantive interpretation (ideological ambiguity), e.g. when a party emphasizes an issue but fails to mention key words associated with it that a party with similar ideology mentions.

Example from Slapin and Proksch 2008

FIGURE 1 Estimated Party Positions in Germany, 1990–2005

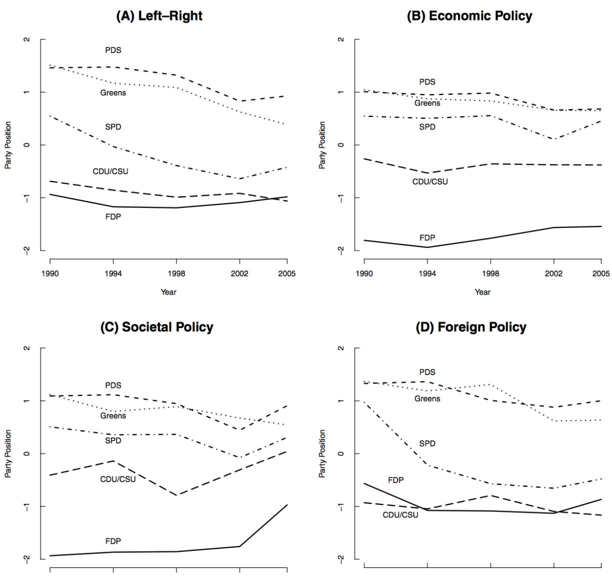
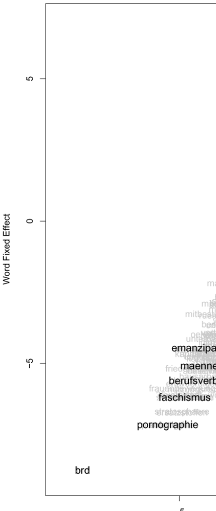


FIGURE 2 Word Weights vs. Wo
1990–2005 (Translati



Wordshoal (Lauderdale and Herzog 2016)

Two key **limitations** of wordfish applied to legislative text:

- Word discrimination parameters assumed to be constant across debates (unrealistic, think e.g. "debt")
- May not capture left-right ideology but topic variation

Slapin and Proksch partially avoid these issues by scaling different types of debates separately.

But resulting estimates are confined to set of speakers who spoke on each topic.

Wordshoal solution: aggregate debate-specific ideal points into a reduced number of scales.

Wordshoal (Lauderdale and Herzog 2016)

- The frequency with which politician i uses word k in debate j is drawn from a Poisson distribution:

$$
\begin{aligned}
w_{ijk} &\sim \mathrm{Poisson}(\lambda_{ijk}) \\
\lambda_{ijk} &= exp(\alpha_{ij} + \psi_{jk} + \beta_{jk} \times \theta_{ij}) \\
\theta_{ij} &\sim \mathcal{N}(\nu_j + \kappa_j \mu_i, \tau_i)
\end{aligned}
$$

- with latent parameters:

  α_{ij} is "loquaciousness" of politician i in debate j
  ψ_{jk} is frequency of word k in debate j
  β_{kj} is discrimination parameter of word k in debate j
  θ_{ij} is the politician's ideological position in debate j
  ν_j is baseline ideological position of debate j
  κ_j is correlation of debate j with common dimension
  μ_i is overall ideological position of politician i

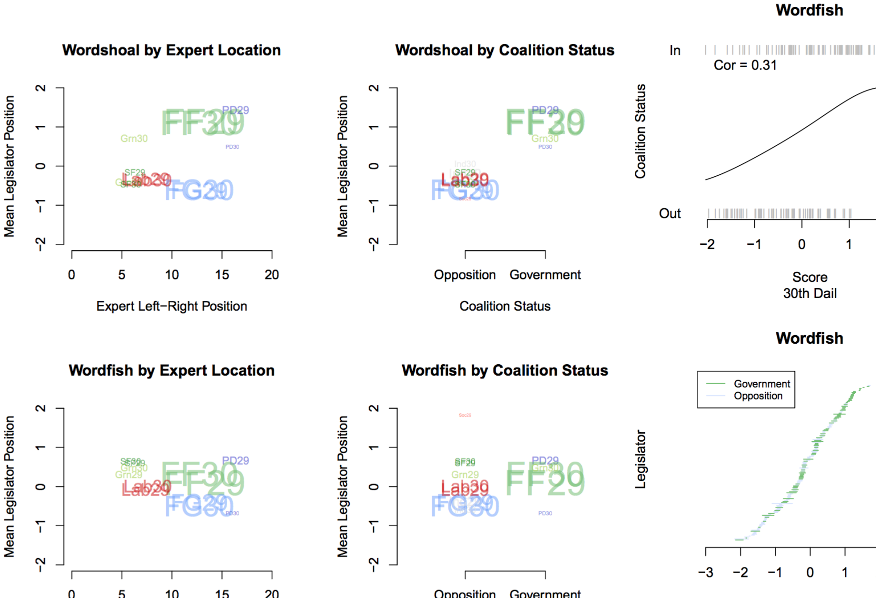
- Intuition: debate-specific estimates are aggregated into a single position using dimensionality reduction

Wordshoal (Lauderdale and Herzog 2016)

New quantities of interest to estimate:

- Politicians' overall position vs debate-specific positions
- Strength of association between debate scales and general ideological scale
- Association of words with general scales, and stability of word discrimination parameters across debates

Example from Lauderdale and Herzog 2016



Outline

- Unsupervised scaling of documents
  - Basics of supervised scaling methods
  - Parametric scaling models: Wordfish and Wordshoal
  - Non-parametric scaling methods: correspondence analysis
  - Practical aspects: computing uncertainty, multiple dimensions, sensitivity to inclusion of irrelevant text
- Unsupervised scaling of features
  - Word embeddings
  - Examples with word2vec

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 distributional 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 document-feature 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 $\underset{n \times k}{\mathbf{X}}$ can be represented in a dimensionality equal to its rank d as:

$$\underset{n \times k}{\mathbf{X}} = \underset{n \times d}{\mathbf{U}} \underset{d \times d}{\mathbf{\Sigma}} \underset{d \times k}{\mathbf{V}'} \tag{1}$$

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

Correspondence analysis

1. Compute matrix of standardized residuals, $\mathbf{S}$:
$$\mathbf{S} = \mathbf{D}_r^{1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^T)\mathbf{D}_c^{1/2}$$
where $\mathbf{P} = \mathbf{Y} / \sum_{ij} y_{ij}$
$\mathbf{r}$, $\mathbf{c}$ are row/column masses: e.g. $r_i = \sum_j p_{ij}$
$\mathbf{D}_r = \mathrm{diag}(\mathbf{r})$, $\mathbf{D}_c = \mathrm{diag}(\mathbf{c})$

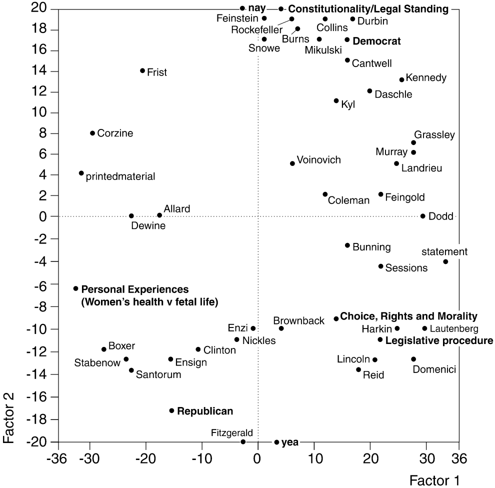
2. Calculate SVD of $\mathbf{S}$

3. Project rows and columns onto low-dimensional space:
$$\theta = \mathbf{D}_r^{1/2}\mathbf{U} \text{ for rows (documents)}$$
$$\phi = \mathbf{D}_c^{1/2}\mathbf{V} \text{ for columns (words)}$$

Mathematically close to log-linear poisson regression model

(Lowe, 2008)

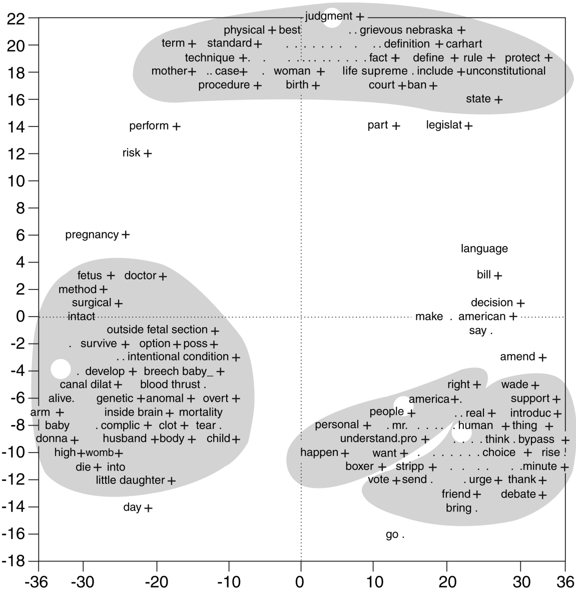
Example: Schonhardt-Bailey (2008) - speakers



	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



Outline

- Unsupervised scaling of documents
  - Basics of supervised scaling methods
  - Parametric scaling models: Wordfish and Wordshoal
  - Non-parametric scaling methods: correspondence analysis
  - Practical aspects: interpretation, computing uncertainty, multiple dimensions, sensitivity to inclusion of irrelevant text
- Unsupervised scaling of features
  - Word embeddings
  - Examples with word2vec

Interpreting scaled dimensions

How can we validate that we are measuring a construct of interest?

1. Semantic validity
   - Most discriminant words correspond to extremes of dimension of interest
2. Convergent/discriminant construct validity
   - Estimated positions match other existing measures where they should match, and depart where they should depart
3. Predictive validity
   - Variation in positions or word usage corresponds with expected events
4. Hypothesis validity
   - Variation in positions or word usage can be used effectively to test substantive hypotheses

How to account for uncertainty in parametric models

- Option 1: Analytical derivatives
  - Reformulating the Poisson model as a multinomial 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 in parametric models

- Option 2: 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
- Option 3: 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
- (and yes of course) Posterior sampling from MCMC

How to account for uncertainty in non-parametric models

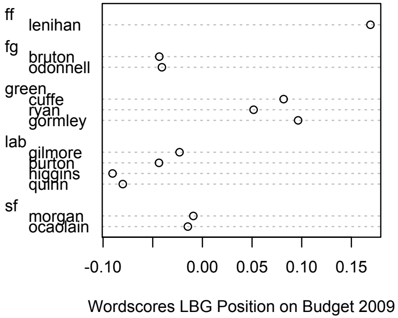
- There are problems with bootstrapping results from correspondence analysis (Milan and Whittaker 2004):
  - rotation of the principal components
  - inversion of singular values
  - reflection in an axis
- Ignore the problem and hope it will go away?
  - SVD-based methods (e.g. correspondence analysis) typically do not present errors
  - and traditionally, point estimates based on other methods have not either

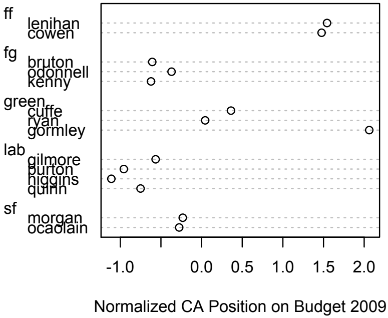
Interpreting multiple dimensions

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
- Correspondence analysis by definition gives you multiple dimensions

What happens if we include irrelevant text?



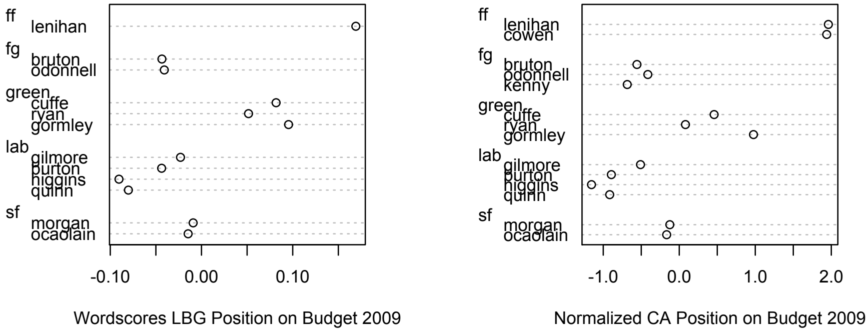


What happens if we include irrelevant text?

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



Outline

- Unsupervised scaling of documents
  - Basics of supervised scaling methods
  - Parametric scaling models: Wordfish and Wordshoal
  - Non-parametric scaling methods: correspondence analysis
  - Practical aspects: interpretation, computing uncertainty, multiple dimensions, sensitivity to inclusion of irrelevant text
- Unsupervised scaling of features
  - Word embeddings
  - Examples with word2vec

Beyond bag-of-words

Most applications of text analysis rely on a bag-of-words representation of documents

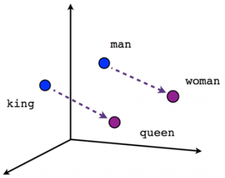
- Only relevant feature: frequency of features
- Ignores context, grammar, word order...
- Wrong but often irrelevant

One alternative: word embeddings

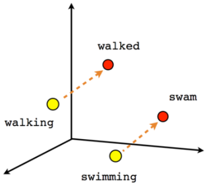
- Represent words as real-valued vector in a multidimensional space (often 100–500 dimensions), common to all words
- Distance in space captures syntactic and semantic regularities, i.e. words that are close in space have similar meaning
  - How? Vectors are learned based on context similarity
  - Distributional hypothesis: words that appear in the same context share semantic meaning
- Operations with vectors are also meaningful

Word embeddings example

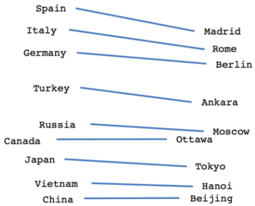
word	D_1	D_2	D_3	...	D_N
man	0.46	0.67	0.05	...	...
woman	0.46	-0.89	-0.08	...	...
king	0.79	0.96	0.02	...	...
queen	0.80	-0.58	-0.14	...	...



Male-Female



Verb tense



Country-Capital

word2vec (Mikolov 2013)

- Statistical method to efficiently learn word embeddings from a corpus, developed by Google engineer
- Most popular, in part because pre-trained vectors are available
- Two models to learn word embeddings: