Scaling

William Lowe

Hertie School

12th November 2020

Scaling

We can think about document living in some kind of *space* with θ as the positione.g.

- affect, a.k.a. *sentiment analysis*
- unidimensional policy preferences
- multidimensional ideological position

How to place documents in space?

- Think of a row in the document term matrix as a vocabulary profile, e.g. by normalize the counts
- This is a point in a (very high-dimensional) space
- Which has distances to every other document in that space

But we can also collapse them down into a smaller space, e.g. one or two dimensions

- Often we think they really live there
- Sometimes it's just visualization

Plan

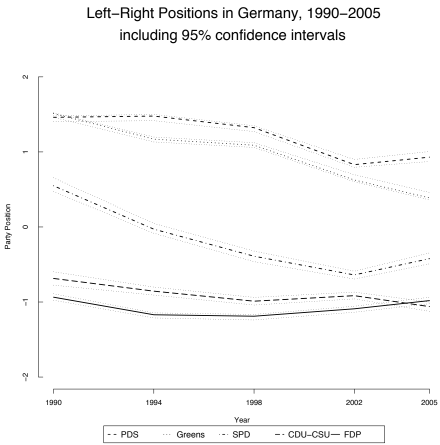
- → Where does information about position live?
- → The model
- → Spatial talking
- → Special cases
- → Validation
- → Comparing positions

Reminder

A matrix of document by word/topic counts is a *contingency table*

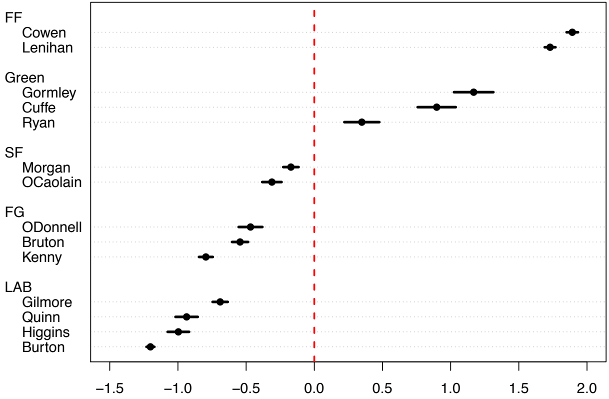
	neue	vor	Menschen	wie	nur	Arbeitsplätze	...
...							
FDP-2005	11	20	6	22	31	17	...
FDP-2002	17	17	27	30	35	9	...
PDS-2005	5	10	17	10	9	12	...
PDS-2002	15	19	8	9	3	9	...
GREENS-2005	42	21	47	46	19	17	...
GREENS-2002	27	18	27	28	22	21	...
SPD-2005	8	15	26	11	13	10	...
SPD-2002	16	18	16	16	9	7	...
CDU-2005	21	12	10	13	19	22	...
CDU-2002	20	20	14	15	18	7	...
...							

Scaling: Party position dynamics



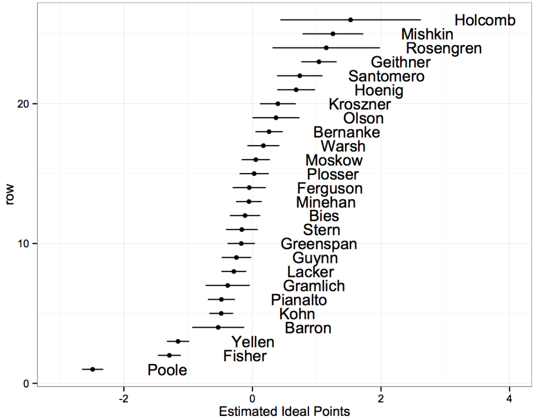
German party position on the economy (Slapin & Proksch, 2008)

Scaling: Irish budget debates



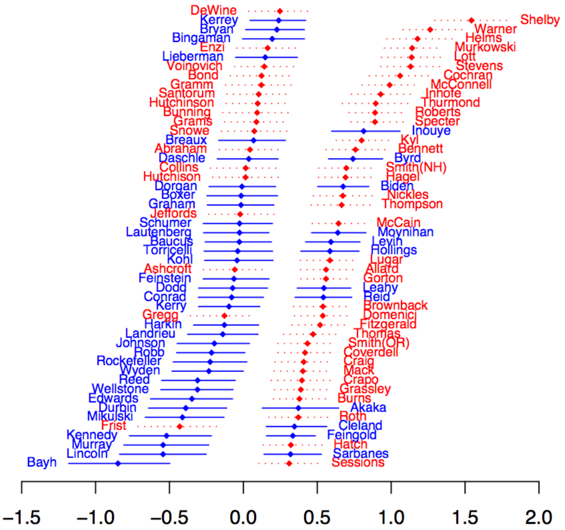
Estimated FOMC member ideal points from meeting transcripts (Lowe & Benoit, 2013)

Scaling: FOMC transcripts



Estimated FOMC member ideal points from meeting transcripts (Baerg & Lowe, 2020)

Scaling: Senators (Monroe & Maeda)



Where positional information lives

		Word			
	Party	Wirtschaft	soziale	Förderung	...
2002	FDP	14	4	15	
	CDU	11	8	20	
	SPD	15	9	10	
	PDS	7	16	9	
	Grüne	2	41	12	
	...				

Assumptions:

→ Position does not depend on *document length*

→ Position does not depend on *word frequency*

Where positional information lives

		Word			
	Party	Wirtschaft	soziale	Förderung	...
2002	FDP	14	4	15	
	CDU	11	8	20	
	SPD	15	9	10	
	PDS	7	16	9	
	Grüne	2	41	12	
	...				

Assumptions:

→ Position does not depend on *document length*

→ Position does not depend on *word frequency*

Implication

→ table margins are uninformative

Where positional information lives

	Word				
	Party	Wirtschaft	soziale	Förderung	…
2002	FDP	14	4	15	
	CDU	**11**	**8**	20	
	SPD	15	9	10	
	PDS	7	16	9	
	Grüne	2	41	12	
	…				

That leaves only *association structure*.

Where positional information lives

		Word			
	Party	Wirtschaft	soziale	Förderung	...
2002	FDP	14	4	15	
	CDU	**11**	**8**	20	
	SPD	15	9	10	
	PDS	7	16	9	
	Grüne	2	41	12	
	...				

That leaves only *association structure*.

The CDU uses 'Wirtschaft' (business) $11/8 = 1.38$ times more than 'soziale' (social).

Where positional information lives

		Word			
	Party	Wirtschaft	soziale	Förderung	...
2002	FDP	**14**	**4**	15	
	CDU	11	8	20	
	SPD	15	9	10	
	PDS	7	16	9	
	Grüne	2	41	12	
	...				

The FDP uses 'Wirtschaft' (business) 14/4 = 3.5 times more than 'soziale' (social).

Where positional information lives

Many $(N-1)(V-1)$ small but relevant facts about relative proportional emphasis

1. FDP's emphasis on Wirtschaft over soziale is 3.5/1.375 = 2.55 times larger than that of the CDU.
2. CDU's emphasis on Wirtschaft over soziale is 0.82...
3. …

You might recognize 2.55 and 0.82 and so on as *odds ratios*

$$\frac{P(\text{Wirtschaft} \mid \text{FDP})}{P(\text{soziale} \mid \text{FDP})} \bigg/ \frac{P(\text{Wirtschaft} \mid \text{CDU})}{P(\text{soziale} \mid \text{CDU})} = \frac{14}{4} \bigg/ \frac{11}{8}$$

which are delightfully indifferent to document lengths and word frequencies.[1]

[1] Add k the frequency of Wirtschaft, keeping the odds ratio the same, and notice that it just adds (some function of) k to both numerator and denominator, which cancel.

Where positional information lives

Actually this is where *all* substantively interesting information in document term matrices lives

→ where else is there?

Any kind of text model, e.g. a topic model

→ implies constraints on how these odds ratios can vary
→ reduces the dimensionality of word distributions to a lower than V space

So let's think about building a model of them from first principles

Modeling the associations

First we'll assume that each C_{ij} is a Poisson distributed with some expected rate

$$C_{ij} \sim \text{Poisson}(\mu_{ij})$$

Modeling the associations

First we'll assume that each C_{ij} is a Poisson distributed with some expected rate

$$C_{ij} \sim \text{Poisson}(\mu_{ij})$$

There are two *log-linear models* of any contingency table

$$\begin{aligned} \log \mu_{ij} &= \alpha_i + \psi_j && \text{(boring)} \\ &= \alpha_i + \psi_j + \lambda_{ij} && \text{(pointless)} \end{aligned}$$

Modeling the associations

First we'll assume that each C_{ij} is a Poisson distributed with some expected rate

$$C_{ij} \sim \text{Poisson}(\mu_{ij})$$

There are two *log-linear models* of any contingency table

$$\begin{aligned} \log \mu_{ij} &= \alpha_i + \psi_j && \text{(independence)} \\ &= \alpha_i + \psi_j + \lambda_{ij} && \text{(saturated)} \end{aligned}$$

Modeling the associations

First we'll assume that each C_{ij} is a Poisson distributed with some expected rate

$$C_{ij} \sim \text{Poisson}(\mu_{ij})$$

There are two *log-linear models* of any contingency table

$$\begin{aligned} \log \mu_{ij} &= \alpha_i + \psi_j && \text{(independence)} \\ &= \alpha_i + \psi_j + \lambda_{ij} && \text{(saturated)} \end{aligned}$$

All the *relative emphasis*, all the odds ratio information, and all the *position-taking* is in λ

Reminder:

- → In log linear model land, the matrix of λ values is just the same size as C
- → but the influence of the row and column margins has been *removed* by the α and ψ parameters

Infer dimensional structure

Intuition: λ has an orthogonal decomposition

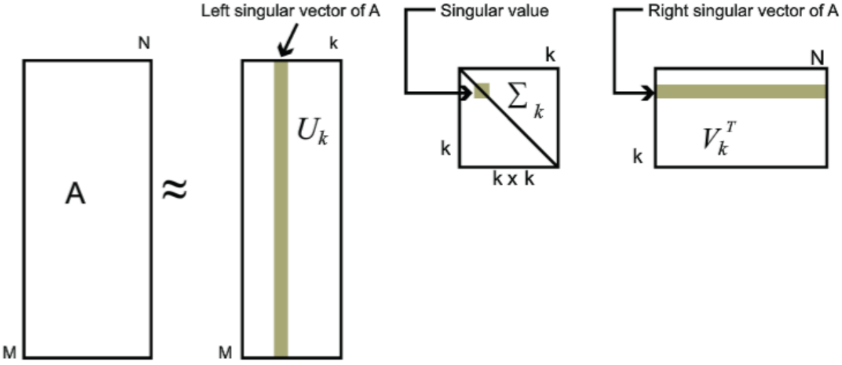
$$
\begin{aligned}
\lambda &= \Theta\Sigma B^T && \text{(SVD)} \\
&= \sum_m^M \theta_{(m)} \sigma_{(m)} \beta_{(m)}^T \\
&\approx \theta\, \sigma\, \beta^T && \text{(Rank 1 approx.)}
\end{aligned}
$$

θ are *document positions*

β are *word positions*

σ says *how much relative emphasizing* is happening in this dimension

SINGULAR VALUE DECOMPOSITION



where A is our λ, U is our θ and V is our β

Modeling proportional relative emphasis

That small fact from earlier

→ FDP's emphasis on 'Wirtschaft' over 'soziale' is 3.5/1.375 = 2.55 times larger than that of the CDU.

According to the model:

$$\log\left(\frac{3.5}{1.375}\right) \approx (\theta_{\text{FDP}} - \theta_{\text{CDU}})\,\sigma\,(\beta_{\text{Wirtschaft}} - \beta_{\text{soziale}})$$

This is a very good idea

Everybody has it...

- → Ecology, archaeology, psychology, political science

and has been having it since Hirschfeld (1935), as

- → the RC Association model (Goodman, 1981)
- → Wordfish (Slapin & Proksch, 2008)
- → Rhetorical Ideal Points (Monroe & Maeda, 2004)

This is a very good idea

Everybody has it...

→ Ecology, archaeology, psychology, political science

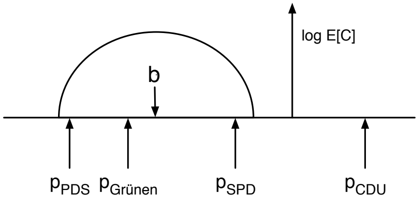
and has been having it since Hirschfeld (1935), as

→ the RC Association model (Goodman, 1981)
→ Wordfish (Slapin & Proksch, 2008)
→ Rhetorical Ideal Points (Monroe & Maeda, 2004)

That was just algebra – *why* is this a very good idea?

Spatial talking

How much will each party use word b?



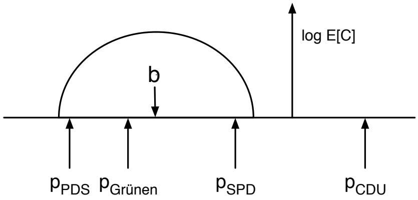
As a model

$$\log \mu_{ij} = r_i + c_j + \frac{(p_i - b_j)^2}{\nu}$$

where ν describes how fast the tendency to say b declines with distance

Spatial talking

How much will each party use word b?



As a model

$$\log \mu_{ij} = r_i + c_j + \frac{(p_i - b_j)^2}{v}$$

where v describes how fast the tendency to say b declines with distance

That we've seen before

$$r_i + c_j + \frac{(p_i - b_j)^2}{k}$$

$$r_i + c_j + (p_i^2 - 2p_i b_j + b_j^2)/v$$

$$\underbrace{(r_i + p_i^2/v)}_{\alpha_i} + \underbrace{(c_j + b_j^2/v)}_{\psi_j} + \underbrace{p_i}_{\theta_i}\ \underbrace{(1/v)}_{\sigma}\ \underbrace{(-2b_j)}_{\beta_j}$$

For the microeconomists, think

→ stochastic utility decision model with V choices

→ and very simple underlying preference structure

→ i.e. a huge structured IIA violation…

For the political scientists

otherwise. Legislators are assumed to have quadratic utility functions over the policy space, $U_i(\zeta_j) = -\|\boldsymbol{x}_i - \zeta_j\|^2 + \eta_{ij}$, and $U_i(\psi_j) = -\|\boldsymbol{x}_i - \psi_j\|^2 + \nu_{ij}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is the *ideal point* of legislator i, η_{ij} and ν_{ij} are the errors or stochastic elements of utility, and $\|\cdot\|$ is the Euclidean norm. Utility maximization implies that $y_{ij} = 1$ if $U_i(\zeta_j) > U_i(\psi_j)$ and $y_{ij} = 0$ otherwise. The specification is completed by assigning a distribution to the errors. We assume that the errors η_{ij} and ν_{ij} have a joint normal distribution with $E(\eta_{ij}) = E(\nu_{ij})$, $\mathrm{var}(\eta_{ij} - \nu_{ij}) = \sigma_j^2$ and the errors are independent across both legislators and roll calls. It follows that

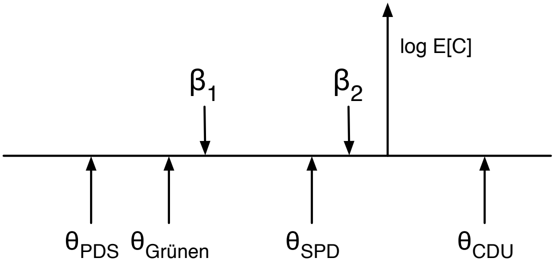
$$\begin{aligned} P(y_{ij} = 1) &= P(U_i(\zeta_j) > U_i(\psi_j)) \\ &= P(\nu_{ij} - \eta_{ij} < \|\mathbf{x}_i - \psi_j\|^2 - \|\mathbf{x}_i - \zeta_j\|^2), \\ &= P(\nu_{ij} - \eta_{ij} < 2(\zeta_j - \psi_j)'\mathbf{x}_i \\ &\quad + \psi_j'\psi_j - \zeta_j'\zeta_j) \\ &= \Phi(\beta_j'\mathbf{x}_i - \alpha_j), \end{aligned} \tag{1}$$

where $\beta_j = 2(\zeta_j - \psi_j)/\sigma_j$, $\alpha_j = (\zeta_j'\zeta_j - \psi_j'\psi_j)/\sigma_j$, and

⟵ from Clinton et al. (2004)

Condition on document length to get a spatial 'voting' model (via the 'Multinomial-Poisson' transform', Baker, 1994; Lang, 2004)

For the political scientists



Decision time: Should I say word 1 or word 2?

→ Depends on my distance to each of them
→ If I can say the word exactly once before being presented with another pair then this is the roll-call voting context and this is a logistic regression (embedded in an IRT model)

Special cases: rile

According to the folk at the WZB (Budge et al., 1987; Volkens et al., 2020), parties signal their ideology by making just such a sequence of choices, over topics

- → Manually identify 'Right' topics R and 'Left' topics L, from a 56 topic codebook
- → For each party manifesto, sum up the Right topic proportions and subtract the sum of the Left topic proportions
- → That's Right-Left position, a.k.a. *RILE*

$$\hat{\theta}_i = \sum_{j \in R} \frac{C_{ij}}{C_{i.}} - \sum_{k \in L} \frac{C_{ik}}{C_{j.}} \qquad \text{RILE}$$

where

$$C_{i.} = \sum_j C_{ij}$$

is the document length

Consequences: Validation

Open questions:

- → Are these the correct category choices?
- → How could we get policy-specific scales? (Benoit et al., 2012; Lowe et al., 2011)
- → What about new categories – where do they fall, ideologically speaking?

Some intermediate answers (Lowe et al., 2011)

- → Probably, but the functional form is not a difference of proportions

$$\hat{\theta}_i = \log \frac{\sum_{j \in R} C_{ij}}{\sum_{k \in L} C_{ik}} \qquad \text{logit scores}$$

- → By careful manual choice of categories
- → ??

Consequences: Validation

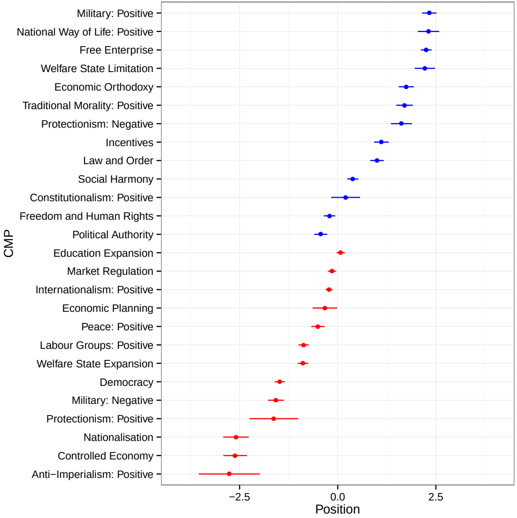
Better answer: Let's check

- → We know already that logit scores are a special case of the model with two 'words'

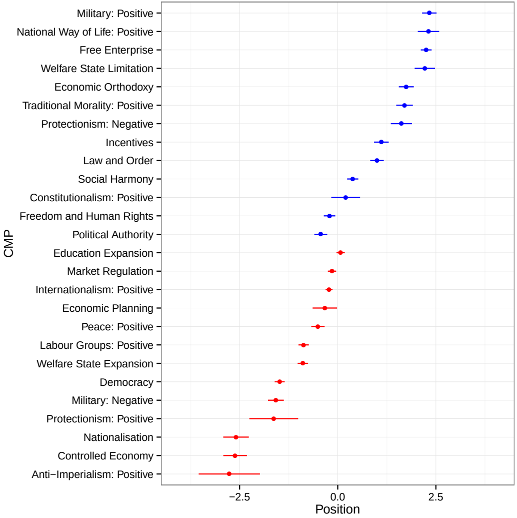
Plan:

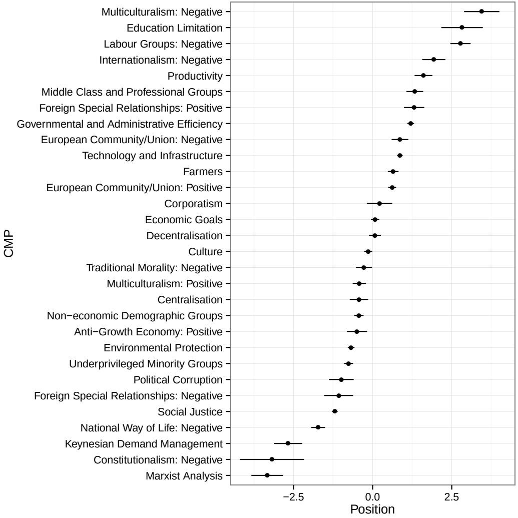
- → Fit the scaling model – here to post-1989 Germany
- → See if the θs agree with *RILE*
- → See if the βs fall into two homogenously position groups
- → Get scores for topics not in L and R

RILE AND OTHER CATEGORIES

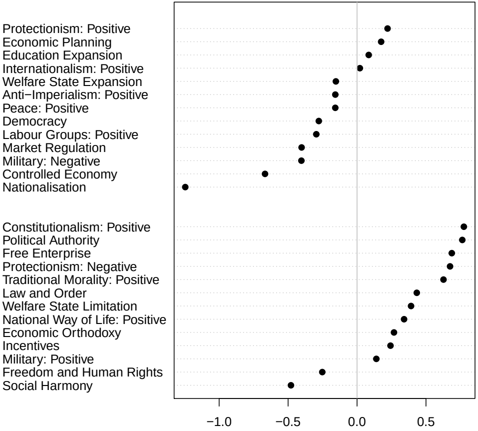


RILE AND OTHER CATEGORIES



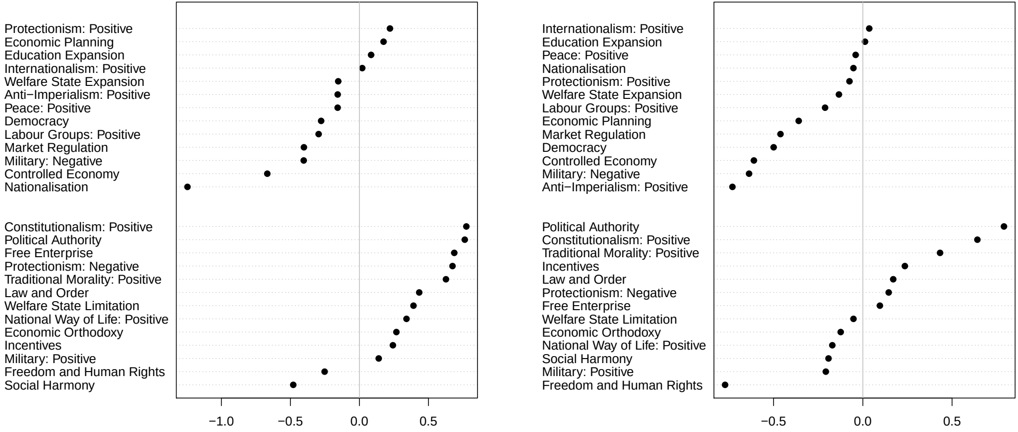


Brits...



Protectionism: Positive
Economic Planning
Education Expansion
Internationalism: Positive
Welfare State Expansion
Anti−Imperialism: Positive
Peace: Positive
Democracy
Labour Groups: Positive
Market Regulation
Military: Negative
Controlled Economy
Nationalisation
Constitutionalism: Positive
Political Authority
Free Enterprise
Protectionism: Negative
Traditional Morality: Positive
Law and Order
Welfare State Limitation
National Way of Life: Positive
Economic Orthodoxy
Incentives
Military: Positive
Freedom and Human Rights
Social Harmony
−1.0
−0.5
0.0
0.5

Brits...


Protectionism: Positive
Economic Planning
Education Expansion
Internationalism: Positive
Welfare State Expansion
Anti−Imperialism: Positive
Peace: Positive
Democracy
Labour Groups: Positive
Market Regulation
Military: Negative
Controlled Economy
Nationalisation
Constitutionalism: Positive
Political Authority
Free Enterprise
Protectionism: Negative
Traditional Morality: Positive
Law and Order
Welfare State Limitation
National Way of Life: Positive
Economic Orthodoxy
Incentives
Military: Positive
Freedom and Human Rights
Social Harmony
−1.0
−0.5
0.0
0.5
Internationalism: Positive
Education Expansion
Peace: Positive
Nationalisation
Protectionism: Positive
Welfare State Expansion
Labour Groups: Positive
Economic Planning
Market Regulation
Democracy
Controlled Economy
Military: Negative
Anti−Imperialism: Positive
Political Authority
Constitutionalism: Positive
Traditional Morality: Positive
Incentives
Law and Order
Protectionism: Negative
Free Enterprise
Welfare State Limitation
Economic Orthodoxy
National Way of Life: Positive
Social Harmony
Military: Positive
Freedom and Human Rights
−0.5
0.0
0.5

Theoretical questions

→ How can we know what θ represents?
→ How can we get policy-specific scores?
→ How comparable are these document and word position estimates?
→ (When) does it make sense to project different documents into a space
→ What would a multi-dimensional version of this model look like

Implementations

Association model

→ What quanteda calls 'wordfish'
→ limited to scaling in one dimension
→ Provides uncertainty estimates for document positions (too small, Lowe and Benoit, 2013)
→ Fit by alternating maximum likelihood, so rather slow

Correspondence Analysis

→ The least squares version of the association model, so tends to agree with it
→ Very fast to fit – just one SVD
→ Multiple dimensions possible at no extra cost
→ θ called 'row coordinates' and β called 'column coordinates'
→ Uncertainty estimates harder to get (bootstrap is possible)
→ Very useful general purpose contingency table visualization tool

Next week

→ Interpretation
→ Multidimensional models
→ Comparisons
→ Connections to other methods

References

Baerg, N. & Lowe, W. (2020). 'A textual taylor rule: Estimating central bank preferences combining topic and scaling methods'. *Political Science Research and Methods*, *8*(1), 106–122.

Baker, S. G. (1994). 'The multinomial-poisson transformation'. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *43*(4), 495–504.

Benoit, K. R., Conway, D., Laver, M. & Mikhaylov, S. (2012). *Crowd sourced data coding for the social sciences*.

Budge, I., Robertson, D. & Hearl, D. (Eds.). (1987). 'Ideology, strategy and party change: Spatial analyses of post-war election programmes in 19 democracies'. Cambridge University Press.

Clinton, J., Jackman, S. & Rivers, D. (2004). 'The statistical analysis of roll call data'. *American Political Science Review*, *98*(02), 1–16.

References

Goodman, L. A. (1981). 'Association models and canonical correlation in the analysis of cross-classifications having ordered categories'. *Journal of the American Statistical Association*, *76*(374), 320–334.

Hirschfeld, H. O. (1935). 'A connection between correlation and contingency'. *Mathematical Proceedings of the Cambridge Philosophical Society*, *31*(4), 520–524.

Lang, J. B. (2004). 'Multinomial-poisson homogeneous models for contingency tables'. *The Annals of Statistics*, *32*(1), 340–383.

Lowe, W. & Benoit, K. R. (2013). 'Validating estimates of latent traits from textual data using human judgment as a benchmark'. *Political Analysis*, *21*(3), 298–313.

Lowe, W., Benoit, K. R., Mikhaylov, S. & Laver, M. (2011). 'Scaling policy preferences from coded political texts'. *Legislative Studies Quarterly*, *36*(1), 123–155.

Monroe, B. L. & Maeda, K. (2004). *Talk's cheap: Text-based estimation of rhetorical ideal-points*.

References

Slapin, J. B. & Proksch, S.-O. (2008). 'A scaling model for estimating time-series party positions from texts'. *American Journal of Political Science*, *52*(3), 705–722.

Volkens, A., Burst, T., Krause, W., Lehmann, P., Matthieß, T., Merz, N., Regel, S., Weßels, B., Zehnter, L. & Wissenschaftszentrum Berlin Für Sozialforschung (WZB). (2020). Manifesto project dataset.