Quantitative Text Analysis

Bruno Castanho Silva

Day 3 - Scaling

Cologne Center for
Comparative Politics

Any questions from yesterday?

Scaling

Identifying positions

- The Politics:
  - Actors *generate, infer, change* and *frame* positions on continuous policy dimensions
- The Methods:
  - Scaling models explain the *generation* and the *inference* parts
  - **Vote** scaling models explain actors' positions with votes; **text** scaling models explain it with text
- The Data:
  - Speeches, election manifestos, social media posts, press releases,...

Position as a latent variable

- What does that mean? Preferences are fundamentally unobservable.
    - Politicians reveal ideology indirectly through their actions, i.e. through voting or talking
    - No matter what measurement instrument we use, there is no directly observable position
    - Available data are manifestations of the latent quantity
- It's all about relative emphasis. . .

Assumptions

- Typically scaling models assume that
    - Relative word usage is reflective of position (θ)
    - Positions are unidimensional
    - Positions drive word counts according to a particular form for $P(W_j \mid \theta)$
    - Bag of words: counts of W_j are conditionally independent given θ

Wordscores

Laver, Benoit, and Garry (2003): supervised scaling

- Each word *j* has a policy position (word score) α_j. This means some words are more extreme (used by one of extreme outliers on the scale), while others are moderate (used by everyone equally).
- The supervision part: some reference document positions are known
- Document positions are average of its words' positions in relation to these reference texts.

Wordscores

Consider two reference texts A and B

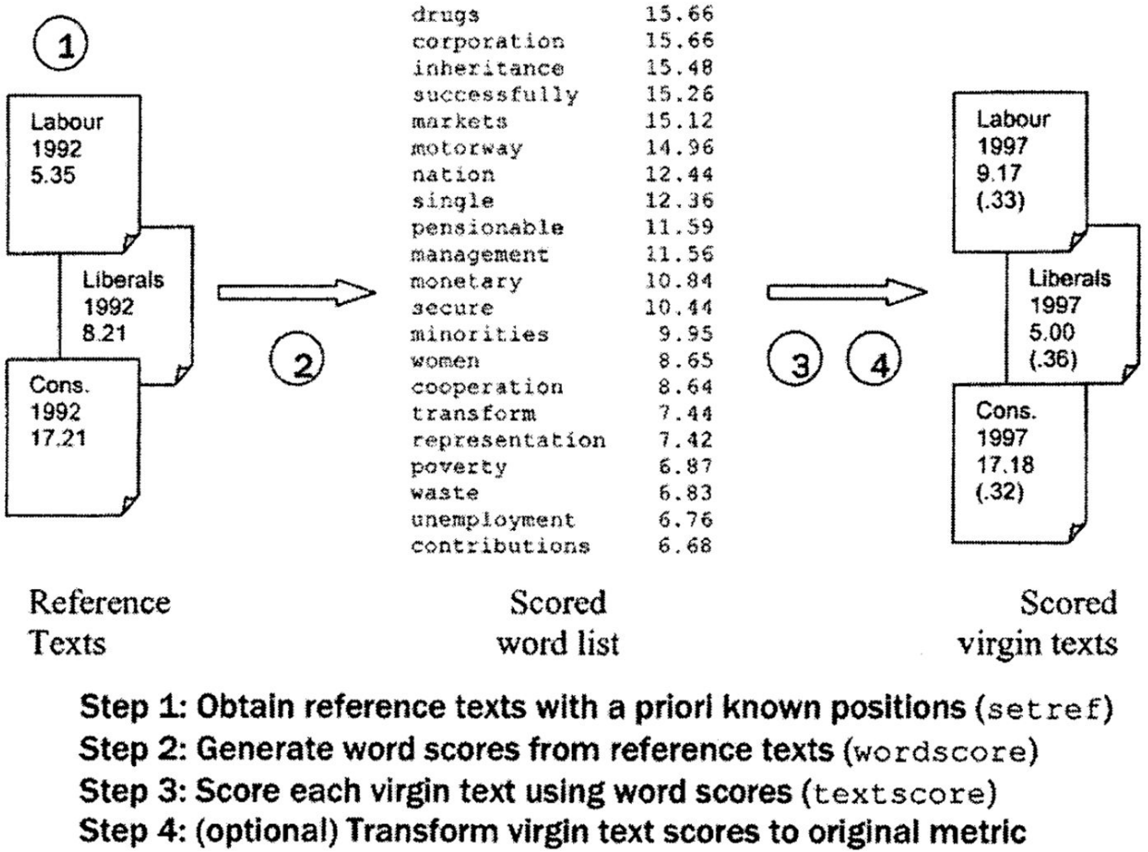
- The word ”healthcare” is used 10 times per 10,000 words in text A and 30 times per 10,000 words in text B
- Conditional on observing the word “healthcare”, we are reading text A with probability 0.25 and text B with probability 0.75
- We can then compute a ”word score” once we assign reference values to the reference texts
- Suppose reference text A has position -1, and text B position +1
- Then the score of word “healthcare” is:

$$0.25(-1.0) + 0.75(1.0) = -0.25 + 0.75 = +0.5$$

Wordscores: how to

1. Define reference texts: Which are the known outliers/most extreme texts?
2. Generate Word scores from reference texts: Which words are used more by one of the actors than the others?
3. Score remaining texts according to their word usage

FIGURE 1. The Wordscore procedure, using the British 1992–1997 manifesto scoring as an illustration



Note: Scores for 1997 virgin texts are transformed estimated scores; parenthetical values are standard errors. The scored word list is a sample of the 5,299 total words scored from the three reference texts.

Wordscores

- Final document scores are not directly comparable to reference documents - the variance in reference texts is much higher
- LBG propose a rescaling method, and others have also proposed alternatives
  - The default today is Martin-Vanberg 2007

Wordscores: limitations

- Which reference texts when more than one election/debate?
- Comparison of reference texts and your documents?
- Influence of the researcher by setting the references
- What kind of policy dimensionality? Completely defined by the reference texts!
    - Might not be what you thought at first

Wordfish

Wordfish

- Unsupervised scaling. No reference texts. More similarity to item response models (e.g. NOMINATE for roll call voting).
- Assumption: There is ONE underlying dimension that is expressed in a collection of texts. We look at words that are predominantly used by some of the actors but not others to maximize differentiation.

Wordfish

The position-word relationship is:

$$W_{ij} \quad \sim \quad \mathrm{Poisson}(\mu_{ij})$$

Where y_{ij} is the count of word j in speaker i's text. Determined by word and document parameters with the form of:

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

Breaking it down

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

- ψ_{ij} are word fixed-effects (frequency of a word overall, irrespective of position)
- β_j is the word weight, capturing the importance of the word in differentiating positions. How fast does the word count increase/decrease with changes in position?
- Θ_i is the position of the document (what we're actually interested in)

Estimation

Wordfish models are fit using Conditional Maximum Likelihood (regression without independent variables)

Iterate:

- Fix document parameters (α and Θ) and maximize word parameters (β and ψ)
- Fix new word parameters (β and ψ) and maximize document parameters (α and Θ) This can be quite slow depending on the size of your dataset, but generally runs in seconds

Model identification

- Much like other scaling or latent variable models, some parameters must be fixed for identification
- Otherwise, there are infinite combinations of Θ and β, which could provide the same likelihood (we would not arrive at a unique solution).
- Solution: fix mean of document positions Θ to 0 and SD to 1. Set one document's α to 0. Set directionality of scale.
- This means that you cannot directly compare estimates ACROSS different estimations.

Dimension issues

- What the heck have we estimated? What is Θ?
- How do we know that positions on only one dimension are being expressed in the text?

Dimension issues

What the heck is Θ?

- Whatever maximizes the Likelihood
- Approximately the first principal component of log W
- Like all scaling techniques (e.g. NOMINATE), Wordfish is effectively exploratory - you have to figure out what the dimension really is. This is the reason why you need to think about your data before applying the method.

Wordfish is about differences

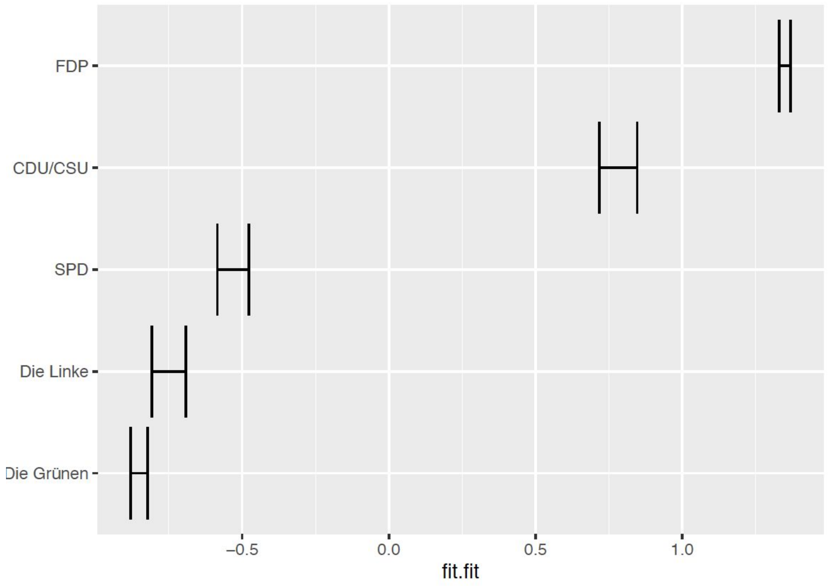
- It will pick up on what differentiates the texts the most

Wordfish is about differences

- It will pick up on what differentiates the texts the most

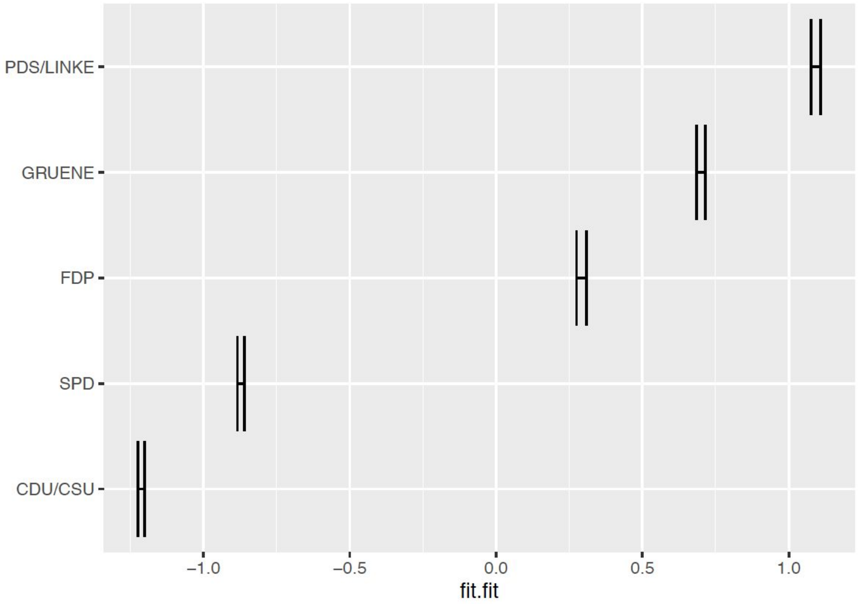
In politics

Wordfish of 2005 German party manifestos



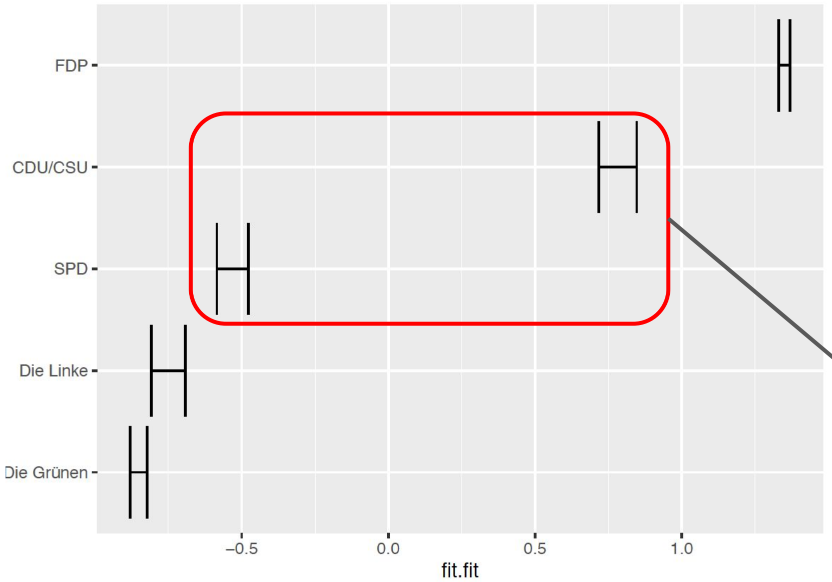
In politics

Wordfish of 2005 German parliamentary speeches

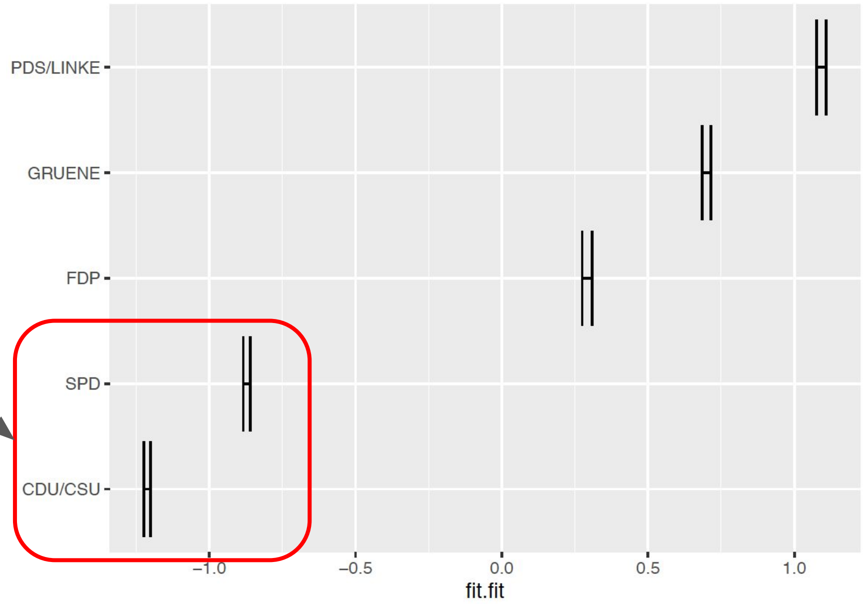


Same actors, different dimensions

- SPD and CDU/CSU were in a governing coalition



Manifestos



Parliamentary speeches

(In)Stability of the Political Lexicon

- What if the political lexicon is unstable over time? New issues appear, old issues disappear
- Scaling algorithms will pick up shifts in the policy agenda rather than shifts in positions.
  - In fact, this is one assumption: that word usage reflects ideology.
  - For example, it becomes seriously problematic when all parties start talking about the ”issue” of the day. Then we can distinguish between elections, but not very well between parties
  - This gets oven more problematic once we start dealing with challenger parties
  - We can (try to) get around this by focusing on those words that remain in the political vocabulary across time.
- There are models such as Wordshoal that implement debate level Wordfish scaling and can deal with different policy contexts in each debate. It estimates a general latent position.

Warnings

Scaling works only...

- if all documents deal with a similar topic and use similar languages (e.g. we can't directly compare newspapers with speeches)
- all speak to the same underlying dimension