Scaling Methods

SICSS Bologna 2024

Paride Carrara

University of Bologna

September 24, 2025

Outline

Outline

Text Analysis

Social science is typically interested in locating things on latent traits. This is known as **scaling**.

- Policy positions
- Inter- and intra-party differences
- ... and any other continuous scale

The intuition behind text scaling

Texts represent an observable implication of some underlying trait of interest (i.e. Ideology).

Each actor has a position θ on a latent dimension d.

Inference about the unobservable latent trait θ can be made by looking at observed features in texts.

Modelling Word Counts to Infer Positions

Getting θ from W words

Assumptions:

- Actors express their positions through their statements.
- Positions drive word counts according to a particular stochastic process.
- Wordscores and Wordfish make different assumptions about the stochastic process that drives word count and about how to deal with dimensionality.

Outline

Wordscores Method

Wordscores (Laver, Benoit, and Garry 2003): supervided model

Wordscores estimates policy positions by comparing two sets of political texts:

Reference texts: texts about which we know something (i.e., a scalar dimensional score)

- Texts whose policy positions are known
- Discriminating as possible (i.e extremes on the dimension) and long texts
- Position of reference text can be taken from external sources(i.e. expert surveys) or assumed

Wordscores Method

Virgin texts: texts about which we know nothing, but whose dimensional score we would like to know

- A set of texts whose policy positions we do not know but want to find out.
- Should be from the same lexical universe as reference texts.
- All we do know about the virgin texts is the words we find in them.
- Wordscores compares these words to the words we have observed in reference texts with "known" policy positions.
- It is an extension of the logic behind dictionary methods.

Wordscores: calculate scores of words

- Begin with a set of R reference documents, each with an associated score A_r on a dimension d.
- Represent reference documents with a document-feature matrix (dfm) C_{rw}, where r indexes the document and w the feature.
- Convert the dfm C_{rw} into a relative dfm by computing the relative frequency of feature counts (i.e., the proportion of the feature counts tf_{rw} of total feature counts $\sum_w tf_{rw}$):

$$F_{rw} = \frac{tf_{rw}}{\sum_w tf_{rw}}$$

Wordscores: calculate scores of words

Once we have observed F_{wr} for each of the reference texts, we have a matrix of relative word frequencies that allows us to calculate a matrix of **conditional probabilities**.

Each element in this matrix tells us the **probability** that we are reading reference text r, **given** that we are reading word w.

In other words, given a **set of reference texts**, the probability that an occurrence of word w implies that we are reading text r is:

$$P(r \mid w) = \frac{F_{wr}}{\sum_R F_{wr}}$$

This quantity **is the key** to the Wordscores a-priori approach.

Wordscores: calculate scores of words

As an **example** consider two reference texts, A and B.

We observe that the word *"Homeland"* is used 10 times per 100 words in Text A and 30 times per 100 words in Text B.

If we know simply that we are reading the word *"Homeland"* in one of the two reference texts, then which is the probability of reading Text A (and Text B)?

The probability that we are reading Text A is $(0.1/(0.1+0.3)) = \mathbf{0.25}$

The probability that we are reading Text B is $(0.3/(0.1+0.3)) = \mathbf{0.75}$

Wordscores: calculate scores of words

We can then use this matrix $P_{r|w}$ to produce a **score** for each word w on dimension d.

This is the expected position on dimension d of any text we are reading, given **only** that we are reading word w, and is defined as:

$$S_{d|w} = \sum_r \left(P_{r|w} \times A_{rd} \right)$$

To continue with this example, imagine that Reference Text A **is assumed** to have a position of 0 on dimension d, and Reference Text B is assumed to have a position of 10 on the same dimension d.

The **score** of the word *"Homeland"* is then

$$S_{d|\text{Homeland}} = 0.75 \times (0) + 0.25 \times (10) = 0 + 2.5 = 2.5$$

Wordscores: calculate scores of words

What happens if a word appears only in one Reference Text and not in the other?

Wordscores: calculate scores of words

What happens if a word appears only in one Reference Text and not in the other? Example: The word "**House**" is contained 0 times in Text A and 1 time in Text B

Wordscores: calculate scores of words

What happens if a word appears only in one Reference Text and not in the other? Example: The word "**House**" is contained 0 times in Text A and 1 time in Text B

$$S_{d|\text{House}} = 0.00 \times (0) + 1 \times (10) = 0 + 10 = 10$$

Wordscores: calculate scores of words

What happens if a word appears only in one Reference Text and not in the other? Example: The word "**House**" is contained 0 times in Text A and 1 time in Text B

$$S_{d|\text{House}} = 0.00 \times (0) + 1 \times (10) = 0 + 10 = 10$$

The reference score is equal to the Reference value for Text B. Given the score, we can consider "House" a **discriminating** word

Wordscores: calculate scores of words

What happens if a word appears only in one Reference Text and not in the other? Example: The word "**House**" is contained 0 times in Text A and 1 time in Text B

$$S_{d|\text{House}} = 0.00 \times (0) + 1 \times (10) = 0 + 10 = 10$$

The reference score is equal to the Reference value for Text B. Given the score, we can consider "House" a **discriminating** word

What happens if a word equally appears in both Reference Texts?

Wordscores: calculate scores of words

What happens if a word appears only in one Reference Text and not in the other? Example: The word "**House**" is contained 0 times in Text A and 1 time in Text B

$$S_{d|\text{House}} = 0.00 \times (0) + 1 \times (10) = 0 + 10 = 10$$

The reference score is equal to the Reference value for Text B. Given the score, we can consider "House" a **discriminating** word

What happens if a word equally appears in both Reference Texts? The word "**Government**" is contained 1 time in Text A and 1 time in Text B

Wordscores: calculate scores of words

What happens if a word appears only in one Reference Text and not in the other? Example: The word "**House**" is contained 0 times in Text A and 1 time in Text B

$$S_{d|\text{House}} = 0.00 \times (0) + 1 \times (10) = 0 + 10 = 10$$

The reference score is equal to the Reference value for Text B. Given the score, we can consider "House" a **discriminating** word

What happens if a word equally appears in both Reference Texts? The word "**Government**" is contained 1 time in Text A and 1 time in Text B

$$S_{d|\text{Government}} = 0.5 \times (0) + 0.5 \times (10) = 0 + 5 = 5$$

Wordscores: calculate scores of words

What happens if a word appears only in one Reference Text and not in the other? Example: The word "**House**" is contained 0 times in Text A and 1 time in Text B

$$S_{d|\text{House}} = 0.00 \times (0) + 1 \times (10) = 0 + 10 = 10$$

The reference score is equal to the Reference value for Text B. Given the score, we can consider "House" a **discriminating** word

What happens if a word equally appears in both Reference Texts? The word "**Government**" is contained 1 time in Text A and 1 time in Text B

$$S_{d|\text{Government}} = 0.5 \times (0) + 0.5 \times (10) = 0 + 5 = 5$$

The reference score is equal to the mean of Reference Text A and Reference Text B. Given the score, we can consider "Government" a **non-discriminating** word

Wordscores: Scoring virgin texts

Finnaly, we can use the estimated scores for virgin text

- Compute first the **relative frequency** F_{wv} of each virgin text word, as a proportion of the total number of words (that received a score) in the virgin text, as:

$$F_{wv} = \frac{tf_{wv}}{\sum_w tf_{wv}}$$

- The **estimated score** of any virgin text v on dimension d, S_{vd}, is then the **mean score** of all of the scored words that it contains, **weighted** by the relative frequency F_{wv} of the scored words:

$$S_{vd} = \sum_w \left(F_{wv} \times S_{d|w}\right)$$

Wordscores: Example

Virgin Text Word Distribution

To calculate the score of a virgin text that contains words with the following distribution:

Word	**Count**
Homeland	6
House	3
Government	1

Word Scores

And these scores calculated from Reference Text A and B:

Word	**Score**
Homeland	2.5
House	10
Government	5

Step-by-step calculation:

$$F_{\text{Homeland}} = \frac{6}{10} = 0.60$$

$$F_{\text{House}} = \frac{3}{10} = 0.30$$

$$F_{\text{Government}} = \frac{1}{10} = 0.10$$

$$F_{\text{Homeland}} \times S_{\text{Homeland}} = 0.60 \times 2.5 = 1.5$$

$$F_{\text{House}} \times S_{\text{House}} = 0.30 \times 10 = 3.0$$

$$F_{\text{Government}} \times S_{\text{Government}} = 0.10 \times 10 = 1.0$$

$$S_{vd} = 1.5 + 3.0 + 1.0 = 5.5$$

So, the score of the virgin text is **5.5**.

Pros and Cons of Wordscores

Pros

- Language-blind: all we need to know are reference scores.
- Estimates unknown positions on a priori scales.
- No inductive scaling with a posteriori interpretation of unknown policy space.
- We "control" the meaning of dimensions through the selection of the reference texts.

Cons

- Very dependent on correct identification of:
  - Appropriate reference texts
  - Appropriate reference scores

Exercise in **Rstudio**

Outline

Wordfish

Wordfish Slapin and Proksch 2008: unsupervised model

- Unsupervised scaling of ideological positions
- Documents are scaled based on similarity/difference in feature use
- WF assume that documents' relative word usage conveys information about their positions in the latent dimension
- Which dimension? Being unsupervised, the interpretation comes after

Wordfish: which dimension?

The first dimension in unsupervised scaling will capture the main source of variation, whatever that is

- Ideally: policy positions, ideology, preferences etc
- But it can also be other dimensions (language, rhetoric style, authorship, topics, etc.)

The validation of Wordfish is essential to understand on which dimension the documents are scaled

Wordfish

Wordfish assumes that the frequency of word m in a document is drawn from a Poisson distribution.

Poisson-distributed variables are bounded between $(0, \infty)$ and take on only discrete values $0, 1, 2, \ldots, \infty$ (like counting words).

Wordfish model

- The frequency W with which politician i uses word m is assumed to be drawn from a Poisson distribution:

$$W_{im} \sim \text{Poisson}(\lambda_{im})$$

$$\lambda_{im} = \exp(\alpha_i + \psi_m + \beta_m \times \theta_i)$$

$$\log(\lambda_{im}) = \alpha_i + \psi_m + \beta_m \times \theta_i$$

- The λ_{im}, the expected count of the word m in document i, is driven by a set of word and document parameters parameter:
  - α_i is the length of the document i (i.e., fixed effect for the length of the document).
  - ψ_m is the frequency of word m (i.e., fixed effect for the frequency of word use).
  - β_m is the discrimination parameter of word m.
  - θ_i is i's position on the latent trait.
- Controlling for document length and word frequencies, words with negative β_m will tend to be used more often by documents with negative θ_i (and vice versa).

Wordfish model

Wordfish uses an expectation maximization (EM) algorithm to retrieve maximum likelihood estimates for all parameters through an iterative process

- Document parameters are held fixed at certain values, while word parameters are estimated
- Word parameters are held fixed at their new values while the document positions are estimated
- These steps are repeated until convergence is reached: The goal is to maximize the likelihood of observing the actual word counts given the model. The process of maximizing the likelihood involves adjusting the model parameters to make the predicted counts as close as possible to the actual counts.

Wordfish: pros and cons

Pros

- Language-blind and fully automated
- No need for reference texts or human supervision

Cons

- Inductive scaling with a posteriori interpretation of unknown policy space
- The identification of the policy dimension is strictly dependent on vocabulary choice Divide the text into single-topic sections can help

Exercise in **Rstudio**

Outline

Alternative Models

The choice of the method will depend on your data (size, topic and temporal variation) and on your specific goal:

- Wordshoal (Lauderdale and Herzog 2016/ed)
- Class Affinity Model (Perry and Benoit 2017)
- Latent Semantic Scaling (Watanabe 2021)
- PartyEmbeddings (Rheault and Cochrane 2020)
- Text Based Ideal Points (Vafa, Naidu, and Blei 2020)
- Manifestoberta (Burst et al. 2023)

Burst, Tobias et al. (2023). *Manifestoberta*. Version 56topics.sentence.2023.1.1. URL: https://manifesto-project.wzb.eu/doi/manifesto.manifestoberta.56topics.sentence.2023.1.1 (visited on 06/11/2024). preprint.

Lauderdale, Benjamin E. and Alexander Herzog (2016/ed). "Measuring Political Positions from Legislative Speech". In: *Political Analysis* 24.3, pp. 374–394. URL: http://www.cambridge.org/core/journals/political-analysis/article/measuring-political-positions-from-legislative-speech/35D8B53C4B7367185325C25BBE5F42B4 (visited on 10/17/2021).

Laver, Michael, Kenneth Benoit, and John Garry (2003). "Extracting Policy Positions from Political Texts Using Words as Data". In: *The American Political Science Review* 97.2, pp. 311–331. JSTOR: 3118211. URL: http://www.jstor.org/stable/3118211.

Perry, Patrick O. and Kenneth Benoit (Oct. 24, 2017). *Scaling Text with the Class Affinity Model*. arXiv: 1710.08963 [cs, stat]. URL:

http://arxiv.org/abs/1710.08963 (visited on 06/11/2024). preprint.


Rheault, Ludovic and Christopher Cochrane (Jan. 2020). "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora". In: *Political Analysis* 28.1, pp. 112–133. URL: https://www.cambridge.org/core/journals/political-analysis/article/word-embeddings-for-the-analysis-of-ideological-placement-in-parliamentary-corpora/017F0CEA9B3DB6E1B94AC36A509A8A7B (visited on 05/31/2021).


Slapin, Jonathan B. and Sven-Oliver Proksch (July 2008). "A Scaling Model for Estimating Time-Series Party Positions from Texts". In: *American Journal of Political Science* 52.3, pp. 705–722. URL: http://doi.wiley.com/10.1111/j.1540-5907.2008.00338.x (visited on 03/13/2021).


Vafa, Keyon, Suresh Naidu, and David Blei (July 2020). "Text-Based Ideal Points". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL 2020. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics,

pp. 5345–5357. URL: https://aclanthology.org/2020.acl-main.475 (visited on 11/07/2023).

Watanabe, Kohei (Apr. 3, 2021). "Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages". In: *Communication Methods and Measures* 15.2, pp. 81–102. URL: https://doi.org/10.1080/19312458.2020.1832976 (visited on 02/05/2022).