# Scaling Political Positions from Text

## – Assumptions, Methods and Pitfalls

Benjamin C.K. Egerod[1]

Robert Klemmensen

**Abstract**

In this chapter we review the different automated techniques for scaling text most commonly used in political science. We start by relating text scaling to the broader field of measurement models aimed at estimating latent positions. Through this comparison, we outline the assumptions underlying the scaling of political text. We proceed to show how the most commonly used scaling techniques build statistical models of text through these assumptions. In doing so, we also show the utility of the various techniques as well as their vulnerable spots. We then focus on two assumptions that are common across all techniques: that the texts are sufficiently a) long and b) similar in the way meaning is ascribed to words. Through simulations, we investigate how sensitive various techniques are to violations of these assumptions. We conclude by discussing how the need for awareness about model assumptions illustrates how automated scaling should not replace human judgement. We also discuss how a) techniques for estimating word and text embeddedness may improve our scaling techniques by incorporating context, and b) the need for conceptualizing measurement error arising from wrong models.

This draft: July 1, 2019

[1]We thank Lucas Leemann as well as the editors Luigi Curini and Robert Franzese, who provided very valuable feedback on earlier versions of this chapter.

# 1 Why Do We Want to Scale Texts?

Virtually all instances of political conflict can be thought of in spatial terms. In everyday language as well in academic discourse we use metaphors relating to space when describing politics. Indeed, it is difficult to even talk about politics without "using the notions of position, distance, and movement" (Benoit and Laver 2006, p. 12). In politics, the left-right distinction—by definition a spatial notion—may be the most enduring organizing principle (Bobbio 1996), and the underlying conceptualization of political preferences distributed along different latent dimensions is closely linked to the spatial models of politics often associated with Downs (1957), Smithies (1941) and Hotelling (1990). While the prevalence of the left-right distinction has made it natural to focus on political ideology, most instances concerning differences between preferences can be thought of in spatial terms. For instance, interest group scholars often conceptualize the degree to which special interests attain their preferences in terms of how policy proposals move relative to the stated positions of groups (Bernhagen et al. 2014; Dür 2008; Klüver 2009). Indeed, there is good reason to believe that spatial models are not just good ways of representing multidimensional data, but good approximations of how humans think about preferences (Armstrong et al. 2014). Therefore, considerable amounts of energy has been devoted to developing methods, which reliably and validly can place actors in political spaces. Scaling methods are devoted to precisely that, and have a long history of successfully placing legislators, political parties, and judges in ideological spaces (e.g. Martin and Quinn (2002) and Poole and Rosenthal (1985, 2000)). More recently, the surge in computational power and availability of new forms of data has provided the possibility of scaling political preferences of extremely diverse sets of actors (Barberá 2015; Bond and Messing 2015; Bonica 2013, 2014; Crosson et al. 2018).

It is in this landscape the scaling of political positions from text fits in. The techniques that are used to scale texts are in large part parallel to the ones used for estimating positions from other data sources, and the use of text scaling has evolved in a similar fasion. While scaling positions from hand-coded databases has a long and successful history (Klingemann et al. 2006), we have seen a surge in the application of computationally intensive methods for scaling texts without first manually coding them. This is both due to the explosion of texts available, and to an increase in the techniques and the computational power that allows us to use them (Laver et al. 2003a; Lo et al. 2016; Martin and Yurukoglu 2017; Monroe and Maeda 2004; Slapin and Proksch 2008). Computational text scaling offers an extremely wide range of potential applications, and while its use is in constant growth, new estimators are continuously invented and applied to innovative data sources, there is no doubt that the field will continue its development for many years to come.

This chapter is dedicated to introducing computational techniques for scaling policy positions from political texts. Because text scaling is similar in theory to other forms of

scaling, but also presents its own challenges, we do not only introduce the reader to how common models can be applied, but also the particular assumptions they make—and how they can be broken.

We start by discussing how text scaling relates to the broader field of measurement theory as it has evolved in political science, and the core assumptions that are needed to scale a set of texts. This structures our review of specific methods for text scaling. We discuss how techniques vary in their assumptions, and illustrate their use with a diverse set of political texts. We discuss the work done by Laver et al. (2003a), which introduced the use of automatic scaling of text to political science. We proceed to review the Poisson scaling model (Monroe and Maeda 2004; Slapin and Proksch 2008), which scales policy position with practically no input from the researcher. The techniques we discuss in this chapter represent but a small subset af the universe of potential scaling techniques. Therefore, we also discuss how each estimator has been extended. While it is impossible to cover all possible estimators in a single chapter, we hope that this chapter can serve as a starting point for the reader. Finally, to illustrate some of the potential pitfals when scaling texts, we include two simulation studies investigating a) how short texts can be, and b) how differently they can use their words, before common methods no longer will be able to meaningfully place them in space. The final section concludes with a discussion of how our existing toolbox can be extended by incorporating new methods for taking word context into account, and how we can think about measurement error more productively than by simply discarding models.

# 2 Text Scaling As Inference About Latent Positions

The goal of methods for scaling positions is to use some observed set of outcomes to draw inferences about an actor's (in the widest sense of the word) unobservable position on a latent dimension relative to other actors. Position is here to be understood as the political preference on some dimension. To get at such a position, the observed outcomes must reveal some kind of preference on the part of the actor. While this holds regardless of the nature of the observed data and the context in which it was produced, different types of data obviously require varying models of the data generating process. Without a good theoretical model of how the observed outcome can discriminate between different latent positions, it simply becomes unclear what exactly it is that is being scaled. The spatial model of politics is probably the widest used model, to relate behaviors—including textual behaviors—to positions. While it obviously is not the only possible model of any single data generating process, it is highly appealing because it is well-tested through years of refinement. In scaling techniques that rely on variations of the spatial model, actors are assumed to choose the outcomes that are most closely aligned with their ideal point (their

political preference).[2] For example, donors contribute to campaigns of candidates, legislators cast their vote for policies, and Facebook users follow political pages—and all choose the ones that are most closely aligned with their preferences. Actors recieve monotonically decreasing utility from choosing outcomes (candidates, policies, Facebook follows) as they increase in distance from their ideal point. The process of choosing a candidate, a roll call vote or a page follow, however, is inherently random in nature, which is typically modeled with some distributional assumption (for a more thorough review, see Armstrong et al. (2014)).

When scaling the political positions of a corpus of texts, similar assumptions are needed, and the spatial model generalizes well to this setting. Here, we can view the choice of words as the outcome. Whenever certain statements are associated with particular political positions, we can use them to discriminate between positions in a certain political space. In other words, the use of a particular (set of) word(s) provides us with a revealed preference for a specific (kind of) policy. Whenever we can think of the data generating process in these terms, the spatial model of politics is likely to provide a good approximation, and scaling a set of documents might be feasible.

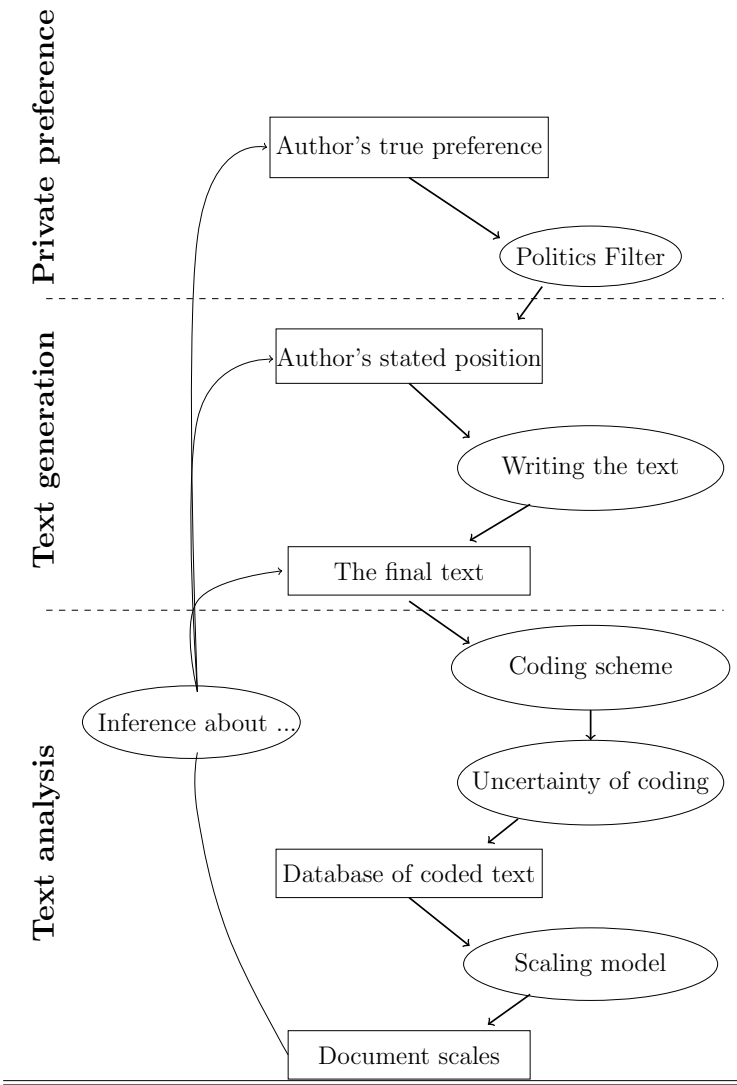
## 2.1 Which Assumptions are Needed to Scale a Text?

To infer from the text in a document to a political position, we need assumptions about three stages in the data generating process, which together operationlize the spatial model of politics in the setting of text scaling. First, we make assumptions about the author's private preference and intent with the analyzed document. The second set of assumptions is about how those prefereces are expressed in any particular document, and how that document relates to the others in the corpus, which in combination make up the relative political space, we want to estimate.

While the first two sets of assumptions are about modelling the causal process that relates text generation to political position, we also need assumptions to translate text into data and from data to scales. Specifically, we need a statistical model that operationalizes the considerations in the underlying causal model about how preferences are communicated in text, and how this relates to the latent position of the document. Table 1 presents the stylized version of the full text and data generating process proposed in Benoit et al. (2009).

[2] This is the so-called proximity based model of space. While a *directional* model obviously is possible as well none of the currently implemented models use it (Armstrong et al. 2014)

Table 1: A stylized model of political text generation & model assumptions in each step




*Note: The stylized model of data generating process is based on Benoit et al (2009).*

**True Preferences & the Politics Filter:** Even though it is called text scaling, what we most commonly want to draw inferences about is not the political position articulated in the text, but the preference held by the author. But if the cost to articulating a position is low, authors' might engage in cheap talk. Conversely, if costs are high, they might choose not to articulate the position for strategic reasons. All of the techniques, we review here, assume that authors do not censor their statements for political reasons. Therefore, we sideline this discussion, even though it obviously can cause significant measurement error.

**Model of the Political Space:** Second, we need to make assumptions about how any given author translates her position into text, and how that relates to the other authors in the corpus. Specifically, the language used in the texts must discriminate between the intended messages of different authors. In other words, the authors should receive varying levels of utility from their choice of words, and this variation should be related to the political space, we want to measure. If authors of different preferences receive the same utiliy from similar choices of words, we cannot use the texts to discriminate between their positions. The documents should be informative about the political differences, we seek to estimate (Slapin and Proksch 2014). Particularly in contexts where there are strong common norms about how to phrase a document (as with highly technical legislative or legal documents) or the texts do not communicate any preference at all, it can be difficult to scale documents. An interesting special case of incomparability is when authors simply use different languages. Importantly, De Vries et al. (2018) have shown that—at least when it comes to topic models—automated translation software (e.g. Google Translate) provides a good way of making languages comparable.

Second, and regarding the relation between documents, a set of texts is only be scalable, if they can be placed in the same euclidian space. This is an often undiscussed assumption, and it can be broken in three ways. One way is if political preferences are discrete views, not matters of continuous differences. Another violation would be if the language used in the documents is incomparable in the way meaning is ascribed to words. Analyzing text that is produced under very different conditions or in varying contexts; that are from different time periods or actors; or have very different audiences in mind would make it difficult to place them relative to each other—let alone in the same space (Slapin and Proksch 2014). Finally, we need assumptions about the dimensionality of the space.

**Stochastic Writing Process:** Whereas the particular spatial model constitutes the assumptions about the systematic component of the data generating process, scaling also requires assumptions regarding the stochastic part. An important assumption in most scaling techniques is that the analyzed units are conditionally independent. When the unit of analysis is word frequencies, this is labeled 'the bag of words' assumption (Grimmer and

Stewart 2013). Conditional on the statistical model, any relation between the use of words is purely noise, and their ordering is inconsequential to the positions that we obtain through our scaling techniques. This assumption allows us to determine differences across texts based on the relative frequencies with which words occur in a text. The main difference in the scaling techniques we review in this chapter is how these frequencies are converted into positions. While this assumption is certainly wrong, it provides an effective simplification, and often scaling techniques work well despite the obviously wrong assumption (Grimmer and Stewart 2013).

Finally, the writing process is inherently stochastic: the same writer would not write the same text, if she sat down to communicate the same message repeatedly. Even if a perfect model of political text could be constructed, and no assumptions were violated at all, the randomness of the writing process would still produce uncertainty in the resulting position estimates. We cannot observe all possible texts the actors could have written, but we can estimate the variance in their positions. While this is clearly not the same as uncertainty, it can help us identify a range within which the author's intended message is likely to be.

**The Model Specification:** Finally, each assumption about the data generating process has to be incorporated into a combined scaling model. This requires assumptions about how the use of words is related to the latent policy position through a functional form and a statistical distribution. This is the step which operationalizes the theoretical considerations in the text generation process. If we use frequencies of individual words, we need to specify our expectations about how a change in frequency helps us discriminate the underlying position. For example, is the relationship log-linear? How does utility decrease as words are further removed from the author's ideal point? We should also consider which other parameters (such as controls, prior information etc.) to include in the model. Together, the assumptions about functional form, distribution and relations between words represent principal statistical assumptions, which implement the theoretical model of text generation.

Most of the estimators, we will discuss here, are highly greedy in their data requirements, however. Thus, an additional and crucial assumption is that the texts, we apply scaling techniques to, contain sufficient information for the technique to pick up differences across texts. The limiting factor is of course how words are distributed in the document-term matrices which are the analytic underpinnings of the methods we discuss. There has to be different distributions of words, and these distributions have to be meaningfully linked to the latent dimensions that we are trying to estimate.

## 2.2 Summing Up

In this section, we discussed how *text* scaling in particular relates to the broader discipline of estimating latent political preferences. We saw how this requires assumptions about the political context, and how it shapes the manner in which authors state their preferences through text. We discussed which assumptions are needed to implement a theoretical model of the political space in the context of text data, and how this forms the basis of scaling the positions of a set of documents. It is worth reiterating that while all of these assumptions are wrong, the model might still be useful in the sense that the estimate of the latent variable may correlate well with the true political preference of an author.

In the next sections, we review the scaling techniques that have been used by political scientists. We relate each technique to the spatial model of politics, and discuss its assumptions about how observed text is related to the author's latent position. This continuously reminds us that while most of the *visible* discretion when scaling a text is contained in the choice of unit of analysis and preprocessing, the choice of even off-the-shelf scaling techniques involves making a number of model assumptions.

# 3 Using Machine Learning to Scale Document Positions

Before analysis we need to make the choice of scaling technique, each of which embodies a set of assumptions about the data generating process. While most techniques resemble statistical ideal point models, we can further distinguish between supervised and unsupervised models. Supervised models use human input, typically in the form of a set of training texts. These estimates can then be used to predict the positions of texts the model has not encountered previously. The training set also serves to define the policy space that the researcher seeks to estimate. Unsupervised techniques simultaneously learn about the latent space and estimate document positions in it, without input from the researcher.

When using any kind of computational technique, preprocessing of the documents is the first step in the modelling process, as it involves making the decision about where in the text we expect the signal about policy positions to be located. A first problem that any scaling needs to deal with is how to translate words to numeric values. That is, what is our unit of analysis, and how do we quantify the prevalence of certain phrases? The techniques we review in the following sections, have traditionally used counts of individual words (unigrams) as their units of analysis. Pairs of consecutive words (bigrams), and words in all possible three contiguous sequences (trigrams) have also been used in political science. But one could use any length of word string (n-grams), which are more common in the broader field of natural language processing. Additionally, noise is often sifted out

by removing extremely common words (so-called 'stopwords'), numbers, punctuation and symbols as well as by reducing words to their stem. Especially for unsupervised models it can be a good idea to remove highly infrequent words. It is extremely important to note that all of these choices regarding preprocessing have consequences, and should not reflect some standardized procedure, but be seen as a step in building the model (Denny and Spirling 2018). The output of this is typically a document-term matrix, where documents are identified along the in rows, terms (words, bi-grams, n-grams etc.) are in columns, and the frequencies are in the cells. This is the input on which we estimate our statistical scaling models.[3] We use the quanteda package (Benoit and Nulty 2016), which is available in R, and comes with an excellent online tutorial that walks the reader through each step in a computational scaling model.

An important caveat when using ideal point models for scaling is that even if all assumptions held, and we could think of a perfect statistical model of text, the incidental parameters problem—that there are too many parameters to estimate relative to observations—would still mean that our position estimates would be wrong. In this regard, it is important to note that the use of computational techniques is no substitute for careful reading of the text and understanding of the subject matter. Automated scaling serves to amplify human ability—not replace it—and its use should be subject to careful subject-specific validation (Grimmer and Stewart 2013).

# 4 Supervised Techniques: Wordscores

The Wordscores algorithm is a one-step approximation of a reciprocal averaging estimator for correspondance analysis on words (Lowe 2008). Originally developed by Laver et al. (2003b) (LBG), it was pioneering because it was one of the first attempts to introduce computational scaling techniques to a wider political science audience. And the model has been hugely successful for a number of reasons. First of all the model is easy to implement because the authors made their software available to the wider public. Secondly, it relies on prior information in the form of reference texts with known positions. This makes it very stable compared to, e.g., unsupervised techniques. This also partially defines the the latent political space before estimation, which makes it extremely flexible. Third, the algorithm is very clear and simple, which further broadens the group of potential users.

## 4.1 The Wordscores Model & Assumptions

Wordscores begins from the premise that we have access to a set of texts $R$ with known positions on the dimension we are interested in. Hence the precondition for a Wordscores

[3] While this is typically what we model, it is not the only concievable form.

model is that we have reliable and valid measures of the positions in a set of reference texts. Wordscores works through the core assumption that each word $w$ has a specific political position, and that the position of a document can be found by averaging over these word scores. The simple idea is that if we first ascribe positions to each word $w$ by observing their frequencies in our reference texts $r$, where document positions are known, then we can use those word scores to predict the positions of out-sample texts by simply observing frequencies of words that also occurred in the reference texts. This is done by developing a measure of the probability of observing a given word $w$ in our reference texts $r$, and using this to infer the positions of a set of out-of-sample texts from their word frequencies.

Specifically, LBG propose to calculate a score $S$ for each individual word in the text using the following equation.

$$S_{wd} = \sum_{r} P_{wr} \cdot A_{rd}, \tag{1}$$

where $P_{wr}$ is the probability $(P)$ of word $(w)$ ocurring in text $(r)$, and $A_{rd}$ is the position given to reference text $r$ on the dimension $d$. Now we have values for each word in our text and we can therefore use the in-sample word scores to infer the position of the out-of-sample texts by using the frequencies of the word, whose positions we know:

$$S_{vd} = \sum_{w} F_{wv} \cdot S_{wd}, \tag{2}$$

where $F_{wv}$ is the word frequency in the out-sample texts $v$. Lowe (2008) outlines the conditions under which bias in policy positions estimated in this fashion is minimized:

1. Positions of reference texts are equally spaced and extend over the range of the positions of individual words.[4]

2. Positions of individual words in the reference texts are equally spaced and extend *past* document positions in both direction.

3. All words are equally informative.

While it is obvious that the first two conditions cannot hold simultaneously in any real-world setting, they provide guidance, when choosing reference texts in a way that minimizes bias. Specifically, the conditions suggests that there should be sufficient overlap between distributions of words in the reference texts, and that they should include a sufficient range of potential word positions in the out-of-sample texts.

[4]Additionally, in the statistical model for Wordscores proposed by (Lowe 2008), where words differ in informativeness, text and word positions should be closely spaced relative to each word's discriminatory power (informativeness).

As mentioned previously, a strong assumption when scaling in general is that the vocabulary does not change radically over texts. When using Wordscores alongside a good choice of reference texts (defined by the above conditions) estimates are generally less sensitive to differences in the meanings and uses of words. We illustrate this point later.

## 4.2 Using Wordscores

To illustrate the use of Wordscores, we draw on data from Baturo and Mikhaylov (2013), who use speeches by Russian governors to estimate how aligned they are with Putin and Medvedev, respectively. By leveraging the fact that the main policy dimension in a Wordscores estimation is defined a priori through the use of reference texts, they are able to estimate where each governor's address to the local parliament falls on a scale from Medvedev to Putin. This use of prior information is what makes supervised techniques like Wordscores extremely flexible. In terms of the spatial model of politics, we can think of the underlying policy space as one in which two leaders compete for control, and state slightly different policy preferences. The assumed utility function of the authors is one, where the governors prefer to converge on the policy position of the most powerful national leader. Thus, we have defined a coherent policy space, and have a clear idea about how written words reveal a preference. In combination, this provides us with a foundation for maping words onto a latent position in this particular space. This is an interesting case, in part because it shows how broadly we can construe the spatial model of politics. An additional interesting feature is that it does not necessarily assume that authors communicate sincere preferences.

We use the texts made available by Baturo and Mikhaylov (2013), which excludes segments on foreign policy. Otherwise, the only preprocessing we do is to remove punctuation. We set the reference scores of Putin and Medvedev to be 1 and -1, respectively. Thus, we fully replicate the original study. Figure 1 shows how alignments estimated through Wordscores changed over time.

To validate the Wordscore estimates, we follow Baturo and Mikhaylov (2013) and use monthly expert evaluations of how powerful Putin and Medvedev are, respectively. To get a direct estimate of partisan alignment, we compute the difference between the Wordscores estimate and Putin's reference position. We use the monthly averages of this difference to facilitate comparison with the expert survey. Figure 2 shows the results. It is clear that when the average governor's speech is more aligned with Putin, expert perception of Medvedev's influence is lower. The correlation is strong and precisely estimated. The correlation between expert perception and the alignment of governor speeches with Putin's position is somewhat lower. This is likely, because there is relatively little variation in both estimates of Putin's influence – he remains continuously powerful by both estimates.

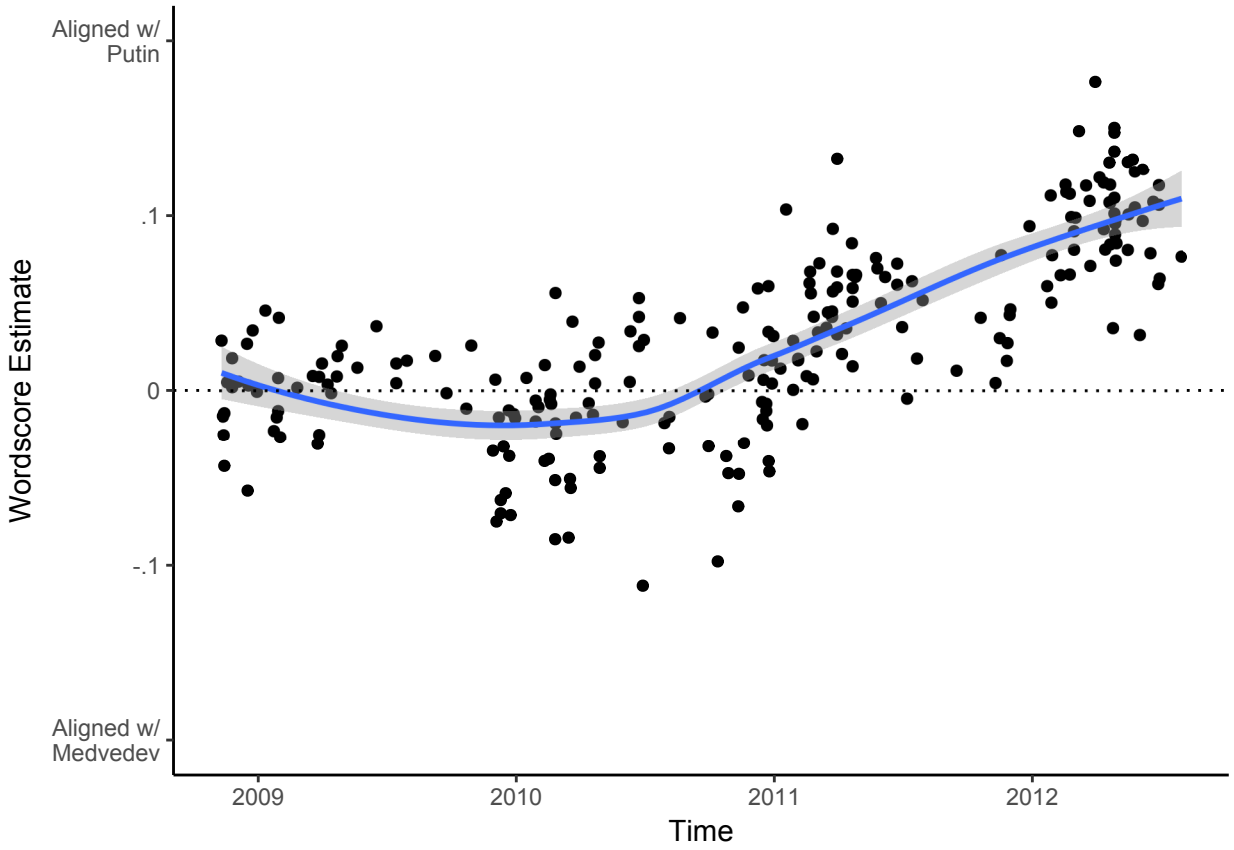



**Figure 1: On A Scale from Medvedev to Putin.** *Note: Each point represents the Wordscores estimate of a governor's speech at a given point in time. The dimension is identified using the most recent parliamentary address by Medvedev (reference score = -1) and Putin (reference score = 1), respectively. The solid blue line is a loess smoother, and the shaded region is a 95 pct. pointwise confidence interval.*

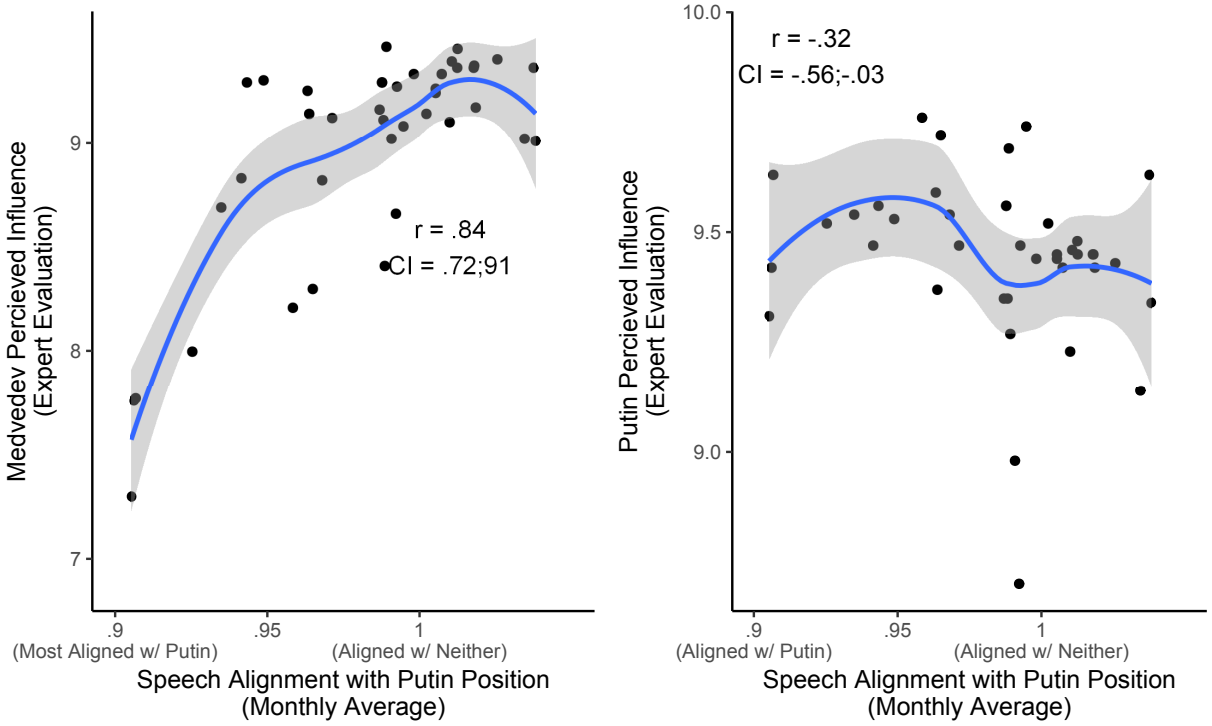



**Figure 2: Perception and Governor Speeches.** *Note: The figure shows the correlations between expert perception of Medvedev's and Putin's power, respectively, and the Wordscores estimate of the average governor's alignment with Putin. The solid blue lines are loess smoothers, and the shaded regions are 95 percent. pointwise confidence intervals.*

## 4.3 Extensions to Wordscores

The original incarnation of Wordscores was strong in its simplicity, but made many of its assumptions implicitly. Lowe (2008) clarifies the underlying assumptions and provides a statistical model for Wordscores, which also serves to relate it to the broader family of statistical ideal point estimators and correspondance analysis. Perry and Benoit (2017) implements a scaling technique using an affinity class model, which is highly similar to the original Wordscores model, and solves some of the problems identified in Lowe (2008).

A second issues that has drawn some attention is how to transform the raw scores obtained by the procedure described above to the same scale used to score the reference texts. In their original article LBG assume that the raw scores for the reference texts have the correct mean, but that the variance is incorrect. Lowe (2008) argues that this assumption might lead to biased scores, because of the shrinkage discussed above. Martin and Vanberg (2008) criticize the original transformation arguing that there the original transformation is dependent on the choice of reference text leading to the uncomfortable position that any desired result could be obtain provided that the right combination of references text are chosen. Consequently, they propose a transformation of the raw scores which build on the relative distance ratios using two anchoring texts which serve as the unit in which all other positions are expressed in relation to. Lowe (2008) argues that researchers are then confronted with a choice when choosing which transformation to use. Either we use the original LBG transformation is dependent on the reference texts and is indifferent to the virgin texts while the opposite is true of the Martin-Vanberg transformation.

# 5 Unsupervised Techniques: Wordfish

Wordfish (introduced in Slapin and Proksch (2008)) is an unsupervised machine learning algorithm, which is based on a Poisson item response theory (IRT) model (Lowe 2015). Being unsupervised, it simultaneously estimates policy positions and learns the policy space using only the texts provided and no external information in the form of virgin texts or anchoring (Grimmer and Stewart 2013). While this is a strength in many aspects, it requires strong modelling assumptions and presents challenges, particularly in regards to validation of the particular policy scales and the dimension as such (Grimmer and Stewart 2013; Lowe and Benoit 2013). In this section, we briefly introduce the statistical model underlying Wordfish, its assumptions as well as how they can be broken, and how a Wordfish model can be estimated and interpreted.

## 5.1 The Wordfish Model & Assumptions

The Wordfish estimator assumes the data generating process to be as follows:

$$y_{ij} \sim Poisson(\lambda_{ij})$$
$$\lambda_{ij} = exp(\alpha_i + \psi_j + \beta_j * \omega_i)$$

Where $y$ is the count of word $j$ in the position document of actor $i$. $y$ is assumed to be drawn at from a Poisson distribution and connected through its mean, $\lambda$, to the systematic component, where $\omega$ is the the estimated position of document $i$. $\psi$ is a word fixed effect, which signifies the frequency of word $j$, when a document expresses a center position on the estimated scale (the difficulty parameter in IRT language). $\beta$, a word's weight in estimating positions, is an estimate of how sensitive the use of a word is to the political position. In IRT, it is called the discrimination parameter, because it measures how the latent position parameter changes in response to word frequencies. It is parallel to a variable's loading in factor analysis (Jackman 2001). $\omega$ is the position of actor $i$ as estimated through its position document. Finally, this leaves $\alpha$, a set of document fixed effects. Estimation is done by iterating over conditional maximum likelihoods.[5]

The Wordfish requires a number of the previously introduced assumptions about the underlying spatial model. However, especially dimensionality is key to the Wordfish model, and it uses a different statistical operationalization than Wordscores.

**The Statistical Model**

Wordfish operates under the assumption that the generation of words in a text—conditional on the model—follows a Poisson process. This has consequences for estimation of both the latent position parameter and the uncertainty of all quantities in the model. First, regarding parameter estimation, it translates into an assumption about the functional form of the relation between word frequency and the latent parameter being log-linear. The model specification introduced above implies assuming monotonicity and that the weight of each word must be the same in any subgroup with the same latent parameter. The former would be violated in situations where word weights do not always increase or decrease with the latent parameter. The latter violation occurs, if two groups with the same policy position use the same word differently.

Second, assuming a Poisson process implies that the variance of the rate of the word count is equal to its expected rate. This assumption will be broken in the presence of both under- and overdispersion as well as structural zeros (when there is zero probability of a word occurring in a text). This induces well-known problems of underestimation of the

[5]This is where the informativeness of the text corpus is particularly important: for estimation to be done, there has to be enough data for the curvature of the log likelihood to be approximately quadratic (Lowe and Benoit 2013).

uncertainty in Poisson models for count data (King 1998), which translate directly into the Wordfish setting (Lowe and Benoit 2013). In practice, violations of the distributional assumption will also lead to poor coverage (Lowe and Benoit 2011).

### (Uni)dimensionality of the policy space

While all scaling techniques require assumptions about dimensionality, the unsupervised ones—like Wordfish—are particularly vulnerable, because they learn about the policy space without input from the researcher.[6] If a researcher misspecifies dimensionality, she risks estimating a policy dimension, which is either meaningless or not the one, she is interested in. There is a number of reasons, why this is a risk. When the generative model specifies a unidimensional policy space, when it really is multidimensional, or the word weights are misspecified, we risk misspecifying the policy dimension. But even if all modeling assumptions hold, the dimension identified by Wordfish might simply be wrong. First, word use in texts adressing the same concerns are likely to be highly correlated. Wordfish will recognize differences in word use between two texts as indicative of their different political positions, but in reality these differences could be due to the topics addressed by the authors (Lowe and Benoit 2013). A special case of this, is in situations where texts use completely incomarable language or do not address similar topics at all. In these situations they cannot be scaled together, and if they are, it will often result in the main policy dimension being misspecified. Finally, the Wordfish estimator's likelihood function is prone to have many local minima, and estimation can easily get stuck in an uninteresting one. This problem could be compounded if there is not enough data for the curvature of the likelihood to be estimated correctly. In this situation, the algorithm might capture noise and a meaningless policy dimension.

## 5.2 Using Wordfish

To illustrate the use of Wordfish, we investigate its performance in estimating the policy positions of European interest groups on three specific issues. Here, we draw on data from Egerod (2016). We use texts from the European Commission's online consultations regarding *Reinforcing sanctioning regimes in the financial services sector*, *A New European Regime for Venture Capital* and *Review of the Investor Compensation Scheme Directive*. We will refer to them, simply, as Sanctions, Venture Capital and ICSD, respectively. For the present purposes, we include only a subset of the interest group responses. Below, when we examine the consequences of violated assumptions, we include all groups. Alongside the

[6] When using supervised techniques like Wordscores, the researcher to some extent defines the policy space herself through her choice of reference texts. This still entails an assumption about unidimensionality, which is obviously likely to be wrong, but if the reference texts are chosen well enough, the estimator is unlikely to estimate a policy space that is very different from the one the researcher is interested in.

interest group position papers, we include the Commission's original Green Paper, which outlines the issues within each consultation, and the final policy proposal.

The main fracture between the interested parties in all three consultations was whether the EU should impose *more* or *less* rules. Therefore, we can think of the underlying spatial model as one, where actors are placed along a continuum ranging from wanting more to less supranational rules. This is the underlying political space in which we wish to place actors. To gauge Wordfish performance, we compare its estimates to hand-coded positions, which aim directly at capturing this space. See Egerod (2016) for more information on the hand-coding.

To prepare documents for Wordfish scaling, we reduce words to their stem, remove stopwords, numbers and punctuation. Figure 3 shows the positions estimated through Wordfish, and how they correlate with the hand-coded positions for documents in each of the three online consultations, we investigate here. The two correlate highly in all three cases, although by far most strongly in the case of National Sanctions, and clearly the least in the Venture Capital case. To save space, we do not discuss the reasons for discrepances between automated and human scaling, which are present in all three cases.

We can use the word weights, or $\beta$ parameters, (i.e. the word discrimination parameter) to analyse the substantive content of the dimension recovered by Wordfish. This can potentially be used to explain, why the two sets of scales diverge for some documents. Figure 4 shows the 21 words with the, respectively, highest and lowest weights for each consultation.

We can take the National Sanctions consultation as an example of how to diagnose divergence between human and machine based scales. There, we can observe that 'labour', 'claus' and 'employ' all have very negative weights—far out in the tail of the full distribution of word weights. This can help us explain why the Belgian union of employees in the financial sector are estimated to be more advocating fewer EU rules. A thorough reading of their position document reveals that, while they are relatively positive overall, they spend many words strongly arguing against employees of financial institutions being liable to prosecution when laws are broken. This seems to be the aspect Wordfish has caught.

## 5.3 Extensions to Wordfish

While Wordfish builds on item response theory, it is closely related to correspondance analysis (like Wordscores) (Lowe 2008). Therefore, correspondance analysis will often provide similar position scales at a lower computational cost. In an early implementation of the Poisson scaling model, Monroe and Maeda (2004) use a Bayesian setup to estimate a two-dimensional model. This is one way of dealing with some dimensionality issues. Slapin and Proksch (2008) have done so by manually separating out the parts of a text that are most closely aligned with predefined dimensions.

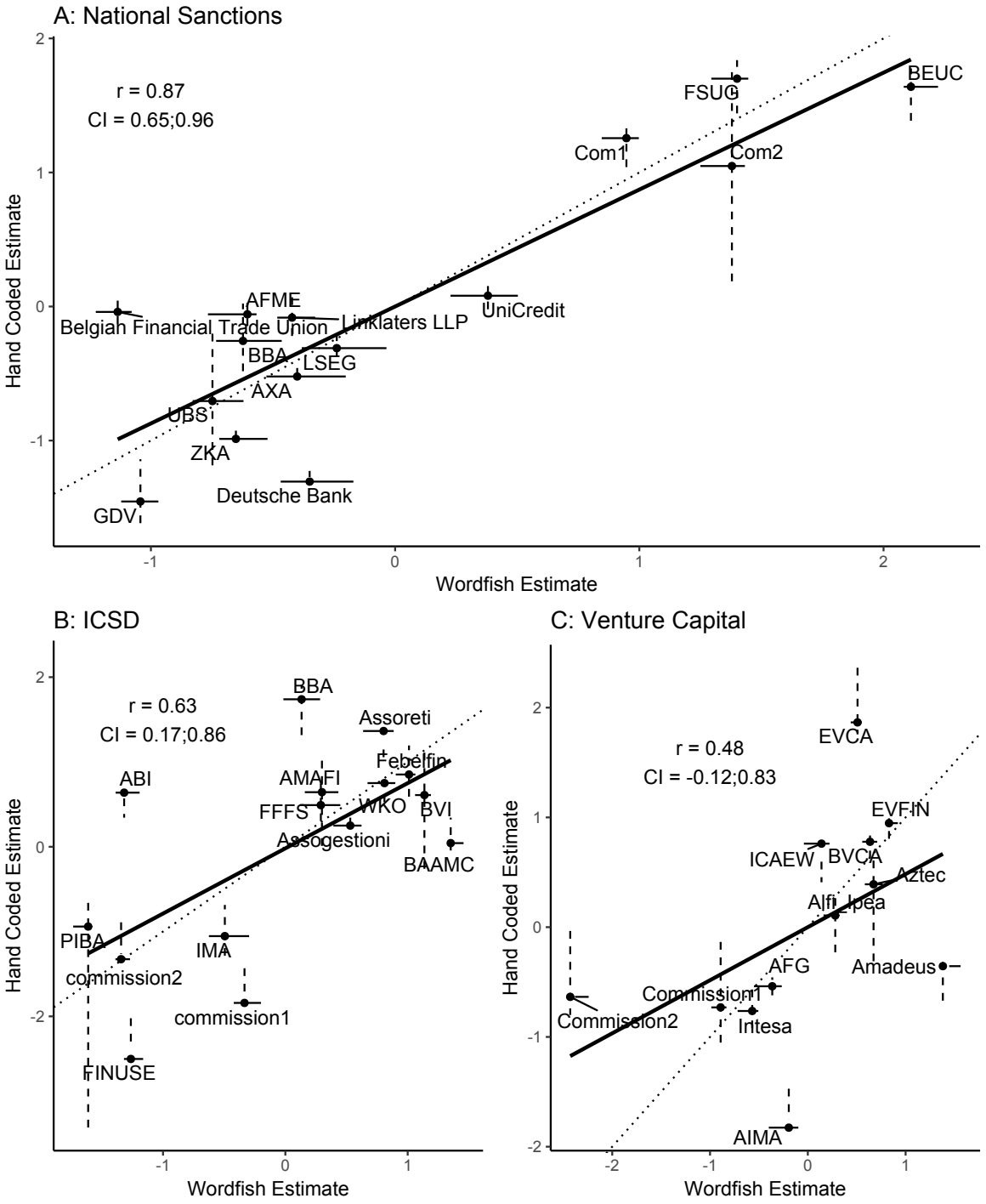



**Figure 3: Validating Wordfish Estimates Against a Human Benchmark.** *Note: Horisontal lines around points are 95 percent parametric Poisson bootstrapped confidence intervals (CIs) around the Wordfish estimate. Vertical, dashed lines are 95 percent CIs from non-parametric bootstraps of the hand-coded scales. 500 resamples used. Solid line is the best linear fit, dotted diagonal line shows what a perfect fit would look like.*

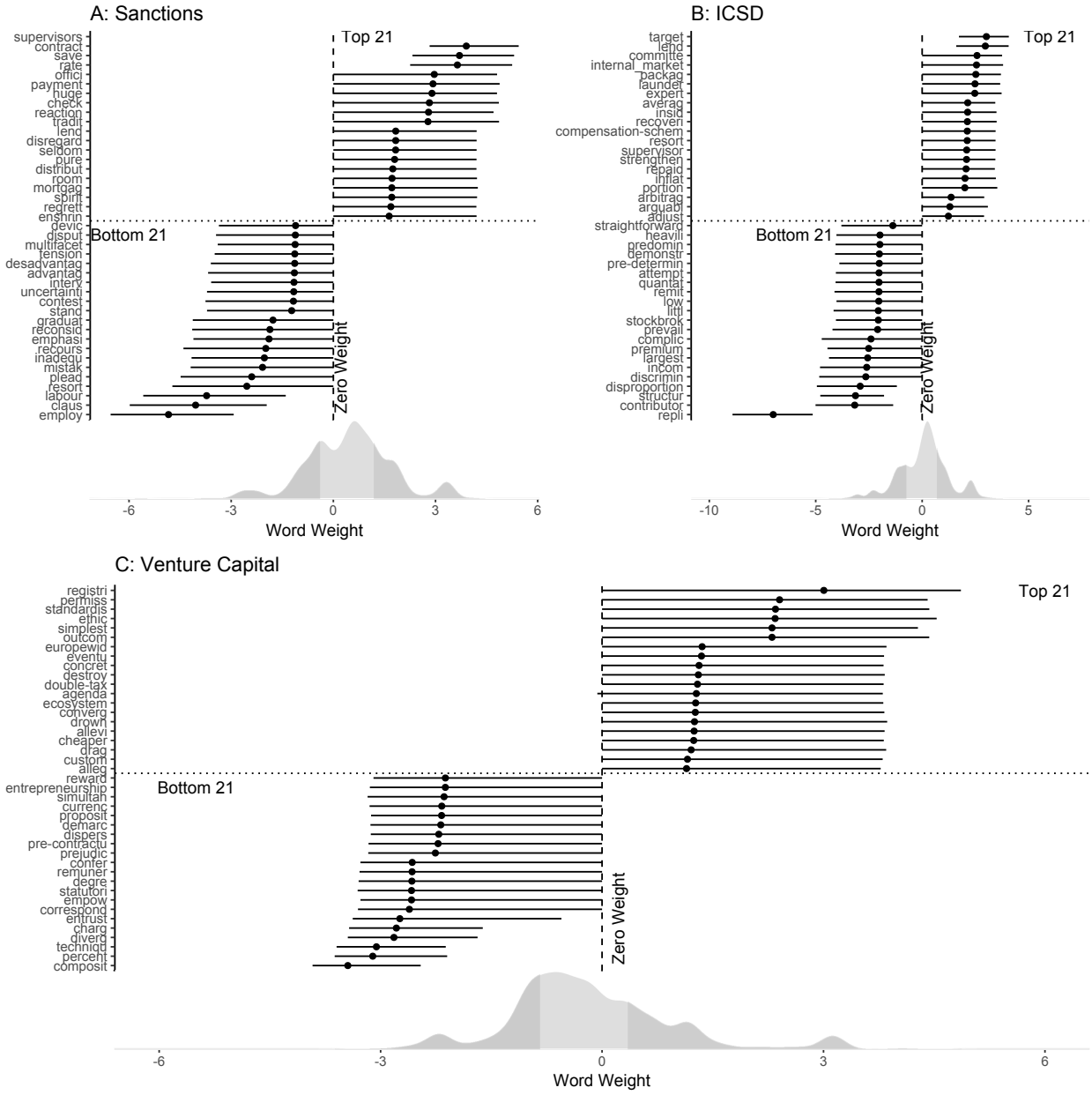



**Figure 4:** Which Words Define the Policy Space? *Note: Points show the point estimate for each word weight. Lines are 95 pct. CIs based on 500 resamples from a conditional Poisson distribution. Density plots show the marginal distribution of word weights in the full corpora. Dark shaded areas are below the 25th and 75th percentiles, respectively.*

Besides these issues of dimensionality, there are number of relevant extensions to the Wordfish model itself. Lowe and Benoit (2013) introduced the use of asymptotic standard errors, instead of the very computationally intensive Poisson bootrstrapped standard errors used in Slapin and Proksch (2008). The Lowe and Benoit (2013) implementation also allows for varying levels of dispersion. Both the analytical and the original technique, however, rely heavily on the model being correctly specified. As a way of obtaining uncertainty estimates with weaker assumptions, Lowe and Benoit (2013) also introduced a non-parametric bootstrap procedure. The quanteda package in R supplies functionality for random sampling of words, which can be used to implement the bootstrap with relative ease. Lauderdale and Herzog (2016) deal with problems of comparability between corpora by using Wordfish to estimate issue specific positions, which they aggregate using Bayesian linear factor analysis to get estimates of overall ideology from text data.

Finally, Lo et al. (2016) exploit the fact that as the rate of failure converges to the limit, a Poisson distribution can be reparameterized as a negative binomial one. This allows them to incorporate a document level dispersion parameter, which can be interpreted as the clarity associated with a document's stated position.

# 6 When Assumptions Are Broken

Assumptions about the generative model and the use of prior information vary between models. Thus, when one model is obviously misspecified and performs poorly a researcher can choose another, more suitable one. However, it is not well understood, how to handle violations of assumptions that are common across models, or how to proceed, when no algorithm performs satisfactorily. In this section, we will investigate the consequences of violations of two basic assumptions that all common scaling techniques rely on: the comparability of language use and the length of the texts.[7]

We use real-world texts and simulate changes to the corpora to quantify the effect of marginal changes to document length and word use. This allows us to inspect what happens, when core assumptions are broken in realistic but controlled settings.

## 6.1 Language Differences

When we previously illustrated the use of Wordfish by using position documents from interest groups in EU consultations, we only relied on a subset of the actual position papers, consisting of 14, 13 and 15 documents, respectively, in the National Sanctions, Venture Capital and ICSD consultations. In reality, however, each corpus consists of 42, 44 and 57 documents from a very wide range of different actors ranging from individuals over govern-

[7] E.g. (Lowe and Benoit 2011) investigate what happens, when distributional assumptions are broken.

ment branches and NGOs through different kinds of corporations and capital funds. These three corpora present an extremely hard test for both the Wordscores and Wordfish algorithms, in particular regarding dimensionality and the comparability of the authors use of words.

To gauge the impact of language differences, we ran both scaling techniques several times on different subsets of consultation documents. We began by running it on all documents in each consultation, then we randomly removed five to eight documents. In each consultation, we chose not to remove the type of actor which was most active in the consultation, which ensures that the included documents become more comparable with each iteration. In the case of Sanctions, we only removed non-corporations. In Venture Capital, we removed documents from actors that were not venture capital funds. In the case of the ICSD, we removed all other documents than those from national employer associations. After numerous iterations, this left only one or few types of organizations and the Commission. The strength of this framework lies in its approximation of counterfactual scenarios—as documents are removed in a semi-random way, and the consultations otherwise remain the same, we hope to estimate the causal impact of altering the composition of the different corpora. Additionally, the dimensionality of the policy space is close to predefined by the Commission in its original policy paper, since it directs interested parties to comment on specific topics.

To save space, we do not show the performance of scalers within each iteration. However, we do find that both Wordfish and Wordscores are highly sensitive to the subset of documents being used, and that both perform best in the smallest, most homogeneous sets of texts. To quantify the degree to which these improvements are driven by decreased differences in word use, we use the correlation between the recovered scales and hand coded positions as the dependent variable in two linear regressions—one for each algorithm. We measure differences in language with two proxies: the average correlation in word use and the number of unique words in each iteration. Because other quantities change besides similarity of the documents, we include as controls the number of positions to be estimated, the average document length, an inverted Hirschmann-Herfindahl index capturing how many different types of interest groups that were included in the estimation and fixed effects for consultations. The results for our variables of interest are presented in Figure 5.

As we can see, the improvements in algorithm performance follow predictable patterns. For both algorithms, the correlation between the human benchmark and the computer-based scales decreases by almost .1 for each percent the number of unique words increases. Note that because we control for average document length, the increase in the number of unique words captured here, is for an unchanged document length. For Wordfish, performance improves by more than .1 every time the average correlation of word use in the corpus increases by .1. This effect is smaller for Wordscores, where the improvement in performance

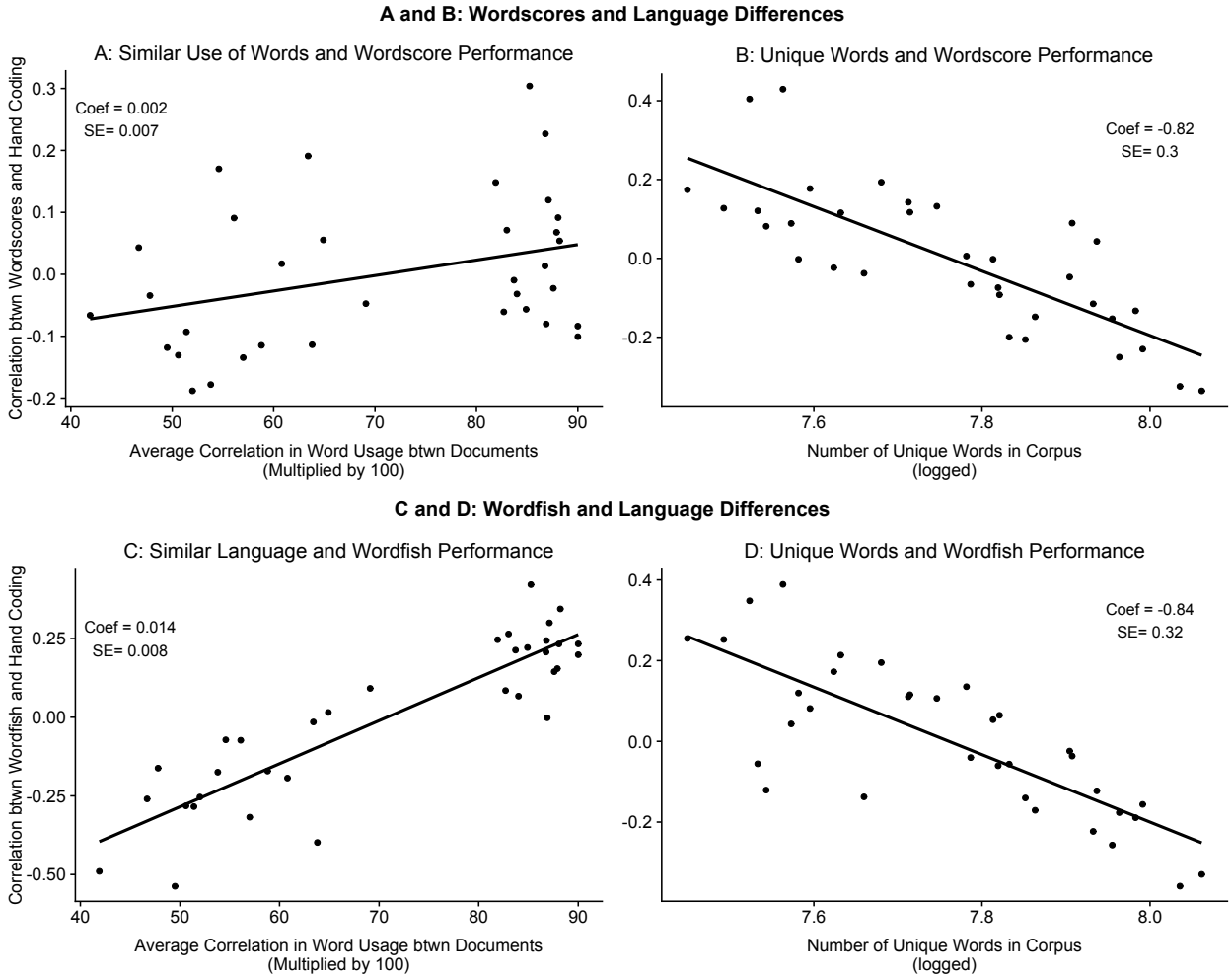



**Figure 5: How the Performance of Scaling Algorithms Vary with the Comparability of Texts.**

is .02. While the latter estimate is not statistically significant, it is still a strong correlation.

This illustrates that as word use in documents becomes more dissimilar, automatic scaling becomes less feasible. The fact that Wordscores is less vulnerable to differences in correlations in word use illustrates a key difference between the two algorithms. Wordfish relies heavily on documents addressing the same concerns using the same words. If they do not, the algorithm is likely to pick up differences in the topics the authors address, not in their political positions. For Wordscores, performance relies more on the reference texts being representative of the broader universe of texts in the corpus. As long as that is the case, differences in word frequencies matter less (although they are not irrelevant), but as they become less representative (e.g. because the number of unique words increase), performance of Wordscores decreases markedly.

## 6.2 Document Length and Informativeness

To get at the effect of document length on the performance of scalers, we use a text corpus, where we know that Wordscores and Wordfish provide good estimates of policy positions—the Lowe and Benoit (2013) data on parliamentary speeches during the debate on the 2010 budget in the Irish Dail. This data also includes estimates of the position of each speech based on human judgement. We simulate changes in document length by randomly reducing the number of times an actor articulates each word in the corpus by between 0 and 3. We reduce the word frequencies in the corpus in this way 100 times, estimating both Wordscores and Wordfish models in each iteration. To measure the performance of the algorithms, we predict the expert coded benchmark using using scales recovered from each algorithm and compute the root mean squared error (RMSE). Because reductions in word frequencies are random, we can get estimates of uncertainy through randomization inference repeating the entire process a 100 times. The results are presented in Figure 6.

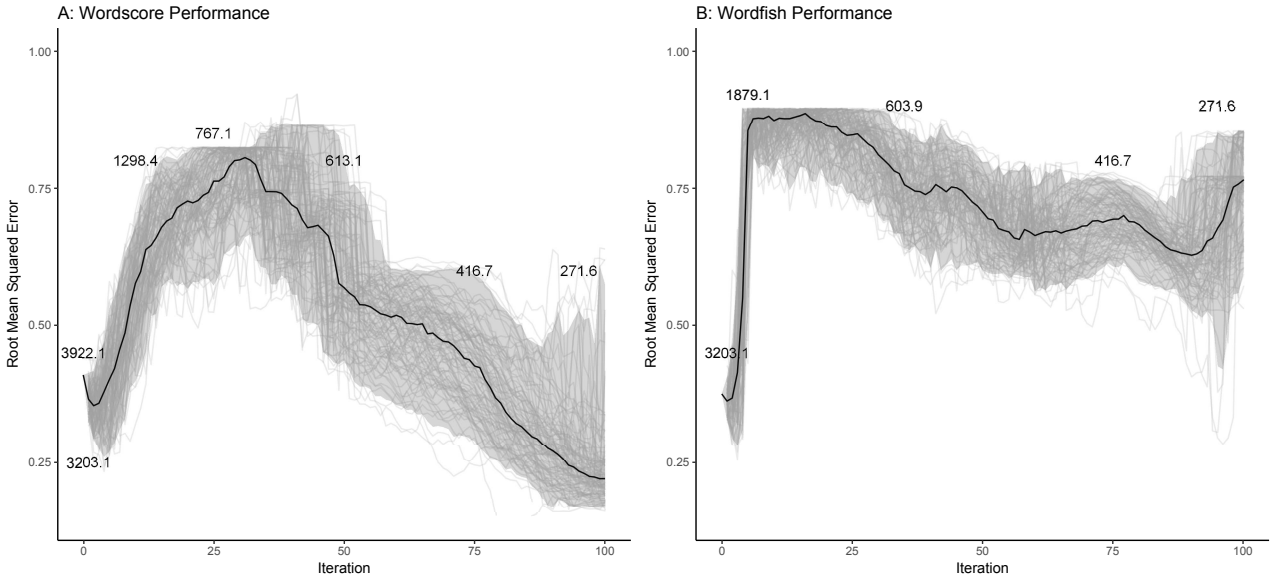



**Figure 6: Document Length and Performance of Scaling Models.** *Note: In each iteration, we reduce the frequency with wich each actor uses a word by a random integer between 0 and three. We run 100 iterations, and rerun the algorithm 100 times to get uncertainty. Shaded lines represents the root mean squared error in each iteration, the solid line shows the average across different random removals. Shaded area is the 95 percent confidence interval.*

The results show that the performance of Wordfish and Wordscores decreases dramatically, as the average document length decreases from the baseline of approximately 3,900. To begin with, scales from both models predict the positions of speeches with a relatively small RMSE of between .3 and .4—corresponding to approximately one-third of the standard deviation. The error increases quickly and stabilizes at about 80 percent of a standard deviation for Wordfish, when it hits an average document length of 1,800 words. For Word-

scores, the RMSE stabilizes at approximately 75 percent of a standard deviation, when the average document length is around 700.[8] With an error of that size, the recovered scales are close to useless.

As in the previous simulation, the use of reference texts to guide the estimates of the Wordscores model proves an important feature. As the average document length decreases below 600 words, Wordscores performance actually starts to improve again. This happens because the word usage in the reference texts becomes more representative, as the length of the left-out texts decreases. While this obviously hinges on a good choice of reference texts, it suggests that the performance of the Wordscores model is a non-monotonic function of document length. With the right reference texts, the algorithm may perform equally well in small and large corpora. The obvious caveat is, of course, that because the texts are very short in the final 20 iterations of each chain, the uncertainty around the average RMSE is relatively high.

## 6.3 Guidelines for Constructing Your Corpus

With the results from these two simulations studies, we can provide some tentative guidelines for researchers. While it is difficult to give precise advise beyond the particular cases, we have investigated here, the results show that the performance of the two scaling techniques is strongly influenced by observable characteristics of the documents being scaled. While we probably cannot infer precise thresholds from these cases to the universe of potential texts to be scaled, it does tells us, what we should be aware of, when we construct those corpora.

In our cases, scaling documents, when the average correlation between their word frequencies is below .6-.7, resulted in poor performance. Additionally, increasing the number of unique words beyond 2,000 without also increasing document length resulted in the recovery of biased positions. Regarding the length of the included documents we found that scaling corpora with fewer than approximately 1,800 words in the average text is infeasible using both algorithms. For Wordscores, however, corpora consisting of very short texts (below 400 words on average) can be scaled, if the reference documents provide good coverage of the virgin texts.

The analyses also provide some insight into how these problems can potentially be handled. First, when there are many unique words relative to the average document length, it will often make sense to trim away very infrequent words, as this removes noise in the Wordfish estimation and improves the representativeness of the reference texts for the Wordscores algorithm.

Second, if texts simply are incomparable, it can make sense to redefine the population of interest to a more coherent group of texts, if it is possible given the particular research

[8]It should be noted that the speed with which the error increases in part is driven by the fact that more word are deleted within each iteration in the beginning, because there simply is more content to delete.

question. If estimation is done over time, it would make sense to split the sample up and run the algorithm within more narrow time periods. If the positions of many different types of groups are to be recovered, one can focus on the most relevant one and only include that one in the estimation.

Finally, if the average document is very short, and there is no way to increase the amount of data available for estimation, Wordscores seems to be able to recover good, but noisy, positions—under the important condition that the reference texts are representative of the remaining corpus.

# 7 What is the Road Ahead?

In this chapter, we have laid out the conceptual foundations for scaling policy positions from hand coded texts as well as automated content analysis. We argued that a spatial model of how policy positions relate to language use is a necessary condition for scaling texts, and that the choice of statistical model should follow from this conceptualization. Based on this framework for understanding text scaling, we outlined the necessary and sufficient assumptions for estimating positions from political texts. This highlights how important conceptualization and the accompanying statistical assumptions are—which, yet again, shows that automated scaling can never replace human judgement, only augment it (Grimmer and Stewart 2013).

We then introduced commonly used techniques for scaling text through supervised as well as unsupervised machine learning (the Wordscores and Wordfish estimators, respectively). Based on the conceptual framework outlined in the beginning of the chapter, we discussed the assumptions embedded in these techniques, and how they could be broken. We illustrated the use of these three scaling models with diverse corpora of political text.

We then proceeded to investigate what happens, when two important assumptions are broken. By simulating random changes in two corpora of real-world text, we illustrated the consequences of estimating positions from texts that a) use their words too differently, and b) are too short. The results suggested that the performance of Wordscores is hurt less by changing these factors, as long as the reference texts are representative, and yielded some tentative guidelines about how to construct a corpus. The key take-away is that performance varies systematically and according to observable features of a corpus. While the thresholds uncovered in the specific corpora investigated here, may not hold in any given other setting, the findings can still inform researchers, when they construct a corpus. The results illustrate how we can use observable features of a corpus to judge its suitability for scaling—we do not always have to rely on abstract argumentation.

In conclusion, we will briefly discuss two potential venues for future research into text scaling, which we think may be fruitful: a) potential ways of improving our scaling models

and b) how we can think about measurement error, when we incorporate scaled positions in econometric models. One goal of this diversity was to illustrate how broadly the spatial model of politics can be construed.

## 7.1 Improving Scaling Models: Dealing with Comparability and Conditional Independence

Throughout this chapter, we have repeatedly discussed two assumptions about the textual context: 1) conditional independence of word (or n-gram) frequencies, and 2) similarity in the way authors ascribe meaning to words (i.e. stability of the vocabularium). We believe that important new work on word and text embeddedness (Mikolov et al. 2013; Ng 2017) may offer ways to deal with this within existing scaling frameworks. We will briefly discuss two such possibilities.

### Conditional Independence of Words

While most scaling techniques peform relatively well, in spite of relying on the simplifying bag-of-words assumption, there is little doubt that it can hurt algorithm performance – and sometimes severely. Throughout the text, we have emphasized that using the frequencies of single word—while the political science standard—is not the only feasible level of analysis. Any n-gram could conceivably be used. This, however, could induce more noise than looking at single words. One way to think about the utility of word embeddedings for scaling techniques is that they offer a statistical model for learning about word context—and, among other things, ways to construct n-grams in a principled manner. The iconical word2vec (Mikolov et al. 2013; Ng 2017) technique, for instance, either uses the context of any given word to predict it, or that word to predict its own context. Either provides a set of probabilities that words co-occur, which, in turn, gives a foundation for finding the optimal n-gram, which could make the conditional indepence assumption more valid.

### Grouping Most Similar Documents

A well-know fact—which we substantiated through simulations—is that dissimilarity in word use and the way in which meaning is ascribed to individual words can severely harm the performance of scalers. Embedding texts in their contexts migh help in this. If we think about dissimilarity of documents as a confounding factor, learning the context of documents may offer a way of modeling it directly. Seeing as position scales will be severely confounded with the most prevalent topics, when documents are dissimilar enough, one way of doing this would be by using topic modeling to learn about the topics in documents before scaling, and then conditioning on them during the estimation of document positions. By

only comparing documents that concern similar topics during estimation, this could be one way of estimating the policy positions of highly dissimilar actors on the same latent scale.

## 7.2 Improving the Use of Scales: Systematic Measurement Error in Political Positions

No matter how much we improve our models or how complex, they become, they will always be erroneous approximations of real-world text generation. While it is relatively straighforward to deal with the special cases, when our models perform either dismally or in a superb manner, we need more research on how to handle the intermediate cases, where a model provides estimates that correlate with true positions, but far from perfectly. This is especially pertinent, when we use those position scales in econometric models seeking to estimate their correlates or use them to explain other political phenomena. The default solutions seem to be to either disregard the measurement error (or assume that it is random), or to disregard any results based on these imperfect scales—neither of which are satisfactory. Because position scales in any given case will be flawed, they will correlate imperfectly with the true positions of a set of actors. Thus, by definition, measurement error in the estimated scales cannot be random. But on the other hand, it seems foolish to discard estimates that, while wrong, provide some useful information.

While a complete treatment of problems with measurement error is beyond the scope of this chapter, we will briefly illustrate how we can think about problems arising, when including erroneous scaling estimates in econometric models. To do so, we conduct a number of Monte Carlo simulations. We use variations of a simple setup: we include a variable measured with some systematic error as either dependent, independent or both variables in a linear regression. We vary the measurement error in increments of 0.2, so the observed scales correlate with the true position estimates with r $\in \{1; .8; .6; .4; .2\}$. To keep things simple, we assume that all remaining Gauss-Markov assumptions are met. We run the three models in 10,000 random samples each including 1,000 observations. Figure 7 shows the distributions of bias arising from each scenario.

Because the setup does not necessarilly generalize, the most important thing to note is the direction of bias: Systematic error in the dependent variable biases the regression estimate downward, while the opposite holds, when the independent variable is measured with error. There is reason to believe that the errors to some extent cancel out, when both variables are measured erroneously. This conclusion, however, rests on the assumption that errors in variables on either side of the equation induce an equal amount of bias in the estimation. In this particular setup, that only seems to be the case, when the independent variable correlates with more than .6 with the true concept. For lower correlations, bias induced by systematic measurement error on the right hand side is much larger in this

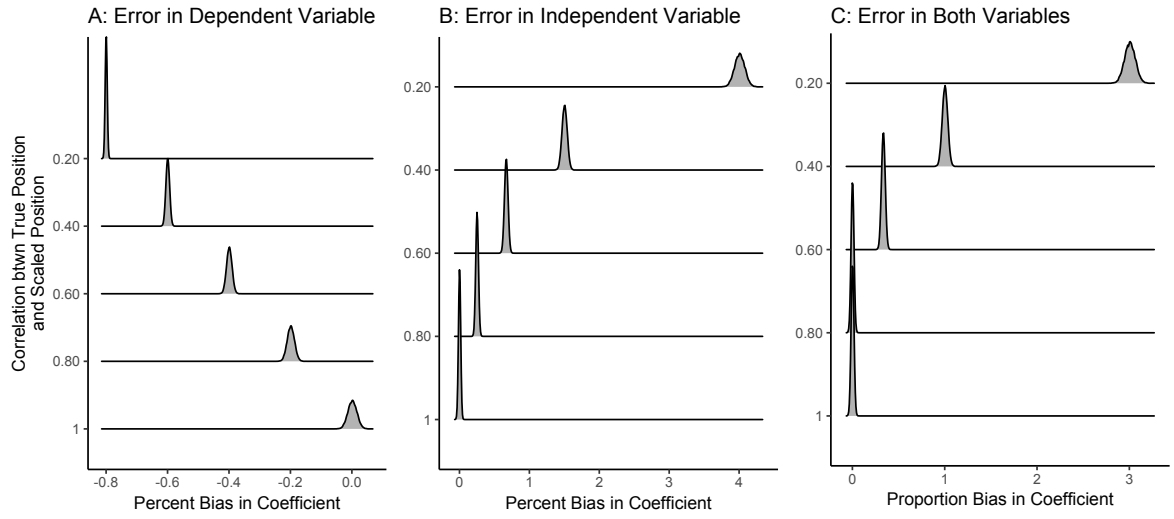



**Figure 7:** Error in Scales and Bias in Econometric Models. *Note: Each distributions shows the percentage difference between estimated coefficient and the true effect as measurement error varies. Each distribution is made up of 10,000 iterations of a OLS regression each with a sample size of 1,000 observations.*

scenario.

This provides an initial illustration of how measurement error in position scales may affect results downstream, when they are included in econometric models. While the scenarios are from general, the exercise shows that researchers should think hard about how measurement error in their scales may impact later stages of their research—not just ignore or discard them—and that more research into the topic might be needed.

# References

Armstrong, David, Ryan Bakker, Royce Carroll, Christopher Hare, Keith Poole, and Howard Rosenthal (2014). *Analyzing spatial models of choice and judgment with R*. CRC Press.

Barberá, Pablo (2015). "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data". In: *Political Analysis* 23.1, pp. 76–91.

Baturo, Alexander and Slava Mikhaylov (2013). "Life of Brian revisited: Assessing informational and non-informational leadership tools". In: *Political Science Research and Methods* 1.1, pp. 139–157.

Benoit, Kenneth and Michael Laver (2006). *Party Policy in Modern Democracies*. Routledge.

Benoit, Kenneth and PP Nulty (2016). *Quanteda: Quantitative Analysis of Textual Data. R package version 0.9. 6-9*.

Benoit, Kenneth, Michael Laver, and Slava Mikhaylov (2009). "Treating words as data with error: Uncertainty in text statements of policy positions". In: *American Journal of Political Science* 53.2, pp. 495–513.

Bernhagen, Patrick, Andreas Dür, and David Marshall (2014). "Measuring lobbying success spatially". In: *Interest Groups & Advocacy* 3.2, pp. 202–218.

Bobbio, Norberto (1996). *Left and right: The significance of a political distinction*. University of Chicago Press.

Bond, Robert and Solomon Messing (2015). "Quantifying social media's political space: Estimating ideology from publicly revealed preferences on Facebook". In: *American Political Science Review* 109.1, pp. 62–78.

Bonica, Adam (2013). "Ideology and Interests in the Political Marketplace". In: *American Journal of Political Science* 57.2, pp. 294–311.

— (2014). "Mapping the ideological marketplace". In: *American Journal of Political Science* 58.2, pp. 367–386.

Crosson, Jesse, Alexander Furnas, and Geoffrey Lorenz (2018). "Estimating Interest Group Ideal Points with Public Position-Taking on Bills in Congress". In:

De Vries, Erik, Martijn Schoonvelde, and Gijs Schumacher (2018). "No longer lost in translation: Evidence that Google Translate works for comparative bag-of-words text applications". In: *Political Analysis* 26.4, pp. 417–430.

Denny, Matthew and Arthur Spirling (2018). "Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it". In: *Political Analysis* 26.2, pp. 168–189.

Downs, Anthony (1957). "An economic theory of political action in a democracy". In: *Journal of political economy* 65.2, pp. 135–150.

Dür, Andreas (2008). "Measuring interest group influence in the EU: A note on methodology". In: *European Union Politics* 9.4, pp. 559–576.

Egerod, Benjamin Carl Krag (2016). "Poisson Scaling of Interest Group Positions from Text in EU Consultations". In: *Amsterdam Text Analysis Conference*.

Grimmer, Justin and Brandon Stewart (2013). "Text as data: The promise and pitfalls of automatic content analysis methods for political texts". In: *Political analysis* 21.3, pp. 267–297.

Hotelling, Harold (1990). "Stability in competition". In: *The Collected Economics Articles of Harold Hotelling*. Springer, pp. 50–63.

Jackman, Simon (2001). "Multidimensional analysis of roll call data via Bayesian simulation: Identification, estimation, inference, and model checking". In: *Political Analysis* 9.3, pp. 227–241.

King, Gary (1998). *Unifying political methodology: The likelihood theory of statistical inference*. University of Michigan Press.

Klingemann, Hans-Dieter, Andrea Volkens, Michael D McDonald, Ian Budge, and Judith Bara (2006). *Mapping policy preferences II: estimates for parties, electors, and gov-*

*ernments in Eastern Europe, European Union, and OECD 1990-2003*. Vol. 2. Oxford University Press on Demand.

Klüver, Heike (2009). "Measuring interest group influence using quantitative text analysis". In: *European Union Politics* 10.4, pp. 535–549.

Lauderdale, Benjamin and Alexander Herzog (2016). "Measuring political positions from legislative speech". In: *Political Analysis* 24.3, pp. 374–394.

Laver, Michael, Kenneth Benoit, and John Garry (2003a). "Extracting policy positions from political texts using words as data". In: *American Political Science Review* 97.2, pp. 311–331.

— (2003b). "Extracting policy positions from political texts using words as data". In: *American Political Science Review* 97.2, pp. 311–331.

Lo, James, Sven-Oliver Proksch, and Jonathan Slapin (2016). "Ideological clarity in multiparty competition: A new measure and test using election manifestos". In: *British Journal of Political Science* 46.3, pp. 591–610.

Lowe, Will (2015). "Austin Vignette". In: *Austin: Do things with words.*

— (2008). "Understanding wordscores". In: *Political Analysis* 16.4, pp. 356–371.

Lowe, Will and Kenneth Benoit (2011). "Estimating uncertainty in quantitative text analysis". In: *Annual Conference of the Midwest Political Science Association*. Vol. 31.

— (2013). "Validating estimates of latent traits from textual data using human judgment as a benchmark". In: *Political Analysis* 21.3, pp. 298–313.

Martin, Andrew and Kevin Quinn (2002). "Dynamic ideal point estimation via Markov chain Monte Carlo for the US Supreme Court, 1953–1999". In: *Political Analysis* 10.2, pp. 134–153.

Martin, Gregory and Ali Yurukoglu (2017). "Bias in cable news: Persuasion and polarization". In: *American Economic Review* 107.9, pp. 2565–99.

Martin, Lanny and Georg Vanberg (2008). "A robust transformation procedure for interpreting political text". In: *Political Analysis* 16.1, pp. 93–100.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean (2013). "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*, pp. 3111–3119.

Monroe, Burt and Ko Maeda (2004). "Talk's cheap: Text-based estimation of rhetorical ideal-points". In: *annual meeting of the Society for Political Methodology*, pp. 29–31.

Ng, Patrick (2017). "dna2vec: Consistent vector representations of variable-length k-mers". In: *arXiv preprint arXiv:1701.06279*.

Perry, Patrick and Kenneth Benoit (2017). "Scaling Text with the Class Affinity Model". In: *arXiv preprint arXiv:1710.08963*.

Poole, Keith and Howard Rosenthal (1985). "A spatial model for legislative roll call analysis". In: *American Journal of Political Science*, pp. 357–384.

Poole, Keith and Howard Rosenthal (2000). *Congress: A political-economic history of roll call voting*. Oxford University Press on Demand.

Slapin, Jonathan and Sven-Oliver Proksch (2008). "A scaling model for estimating time-series party positions from texts". In: *American Journal of Political Science* 52.3, pp. 705–722.

— (2014). "Words as data: Content analysis in legislative studies". In: *The Oxford Handbook of Legislative Studies*, pp. 126–144.

Smithies, Arthur (1941). "Optimum location in spatial competition". In: *Journal of Political Economy* 49.3, pp. 423–439.