UNIVERSITY OF COPENHAGEN

Poisson Scaling of Interest Group Positions from Text in EU Consultations

Egerod, Benjamin Carl Krag

Publication date:
2016

Document version
Peer-review version

Document license:
Ikke-specificeret

Citation for published version (APA):
Egerod, B. C. K. (2016). *Poisson Scaling of Interest Group Positions from Text in EU Consultations*. Paper præsenteret ved Amsterdam Text Analysis Conference, .

Download date: 20. sep.. 2019

When to Wordfish - and when not to

Poisson scaling of interest group positions in EU consultations

Benjamin Carl Egerod[1]

[1]PhD fellow at the Department of Political Science, University of Copenhagen.

1 Introduction

The production of text is an inherent part of virtually all aspects of politics (as noted by Grimmer & Stewart 2013), and in all phases of the policy cycle actors make their preferences known in the form of political text. As such, text represents one of the most abundant and unspoilt sources of political data available (Lowe et al 2011). Even though the information inherent in political texts has never been underestimated, until recently political scholars have only seldom been able to quantify it for the use in statistical modeling – the cost of manually coding this vast amount of knowledge has simply been too great. However, with the recent series of breakthroughs in the computer sciences, we may have come a large step closer in realizing the potential of textual analysis in political science.

For the study of special interest group (SIG) influence the brave new world of automated textual analysis may be of significant interest. Professor James March once noted that *"there is lacking not only an immediately obvious unit of measuring influence, but even a generally feasible means of providing simple rankings"* (cited in Dür 2008). One evident way of solving this operationalizational problem is by using textual analysis to estimate the preferences implied in the position papers of the interested actors and the decision-maker's draft proposals. This allows for tracing how policy-makers respond to the preferences of single SIGs and, hence, attributing influence to these actors.

The amount of text surrounding any single piece of legislation, however, is vast, and the job of coding – let alone making sense of – these hundreds of highly technical pages quickly tends towards being insurmountable as the number of cases increase. Automated textual analysis in this regard seems heaven sent - but there are a range of pitfalls. In this paper, I examine how the unsupervised textual analysis algorithm Wordfish[2] (Slapin & Proksch 2008) performs in estimating preferences of SIGs active in European Union (EU) consultations. Specifically, I will investigate under which circumstances we can expect the use of Wordfish to be feasible and how to design research to make its use optimal. Wordfish performance is, indeed, a vital question to shed light on, since its use to attribute influence to SIGs has become quite controversial.

The use of Wordfish to estimate positions in EU consultations was first introduced in an innovative study by Klüver (2009). The findings were extremely encouraging and held the promise of extending the study of interest group politics significantly. The validity of using the Poisson scaling model in EU consultations has, however, been forcefully called into question in recent work by Bunea and Ibenskas (2015). They find evidence suggesting that EU consultation documents may violate core assumptions necessary for the use of automated textual analysis – and therefore also Wordfish – to be feasible.

[2]Through the text, I will refer to the algorithm either as the poisson scaling model or simply Wordfish.

Throughout this paper, I will try to find a middle ground. Instead of asking whether it is at all appropriate to apply Wordfish to the study of interest groups, I will endeavour to shed light on *when* the poisson scaling model can be expected to provide feasible estimates of SIGs positions – a necessary step towards estimating their influence on public policy. To investigate this, I have hand coded the documents of three online consultations from the Directorate-General of Internal Market (DG MARKT) under the EU Commission. These cases all represent hard tests of the algorithm, which makes generalizations more feasible. I suggest a framework that allows for approximating the counterfactual effect of altering the composition of the corpus of texts. In this way, I show that when a) the documents are results of similar data generating processes, b) the documents contain sufficient information and c) there is a large number of documents included, the Poisson scaling model can be expected to perform well.

The pitfalls associated with use of automated textual analysis to proxy political influence is conceptually similar to the well-known problems associated with measurement error. The consequences of this specific type of measurement error is not well investigated, however. Therefore – after analysing when Wordfish can be expected to perform well – I use Monte Carlo simulations to explore the properties of the logit estimator when using an error laden proxy of influence as dependent variable. The results show that the estimator is biased and inconsistent with the estimated expected value attenuated towards zero.

The remainder of the paper is structured as follows: in the next section I provide some theoretical predictions about, when I expect Wordfish to perform well. Section three presents the hand coding scheme and outlines my estimation strategy. Among other things, I will provide argumentation for the least likely nature of the consultations under scrutiny. In section four I present evidence on how Wordfish estimates correspond to human hand-coding and how the correspondence changes as different subsets of documents are analyzed. In section five, I present my statistical evidence on the determinants of Wordfish performance. In section six I present the results from my Monte Carlo simulations. The final section concludes with a set of guidelines for future research.

2 Poisson scaling of interest group positions

In this section, I will briefly present the theoretical consideration underlying my expectations regarding Wordfish performance. They will all take as their departing point the Poisson scaling model.

Wordfish (Slapin & Proksch 2008) is an unsupervised machine learning algorithm that uses the functional form of the Poisson regression estimator. Being unsupervised, it estimates policy positions using only the texts provided and no external information in the form of

virgin texts or anchoring (Grimmer & Stewart 2013: 15). In the simple non-time-series setting the data generating process is assumed 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[3], λ, to the systematic component. This is made out of α, a set of document fixed-effects, ψ, word fixed-effects, β, a word's weight in estimating positions. ω is the parameter of interest – the position of actor i as estimated through its position document.

What makes the estimator different from the Poisson regression is that we want to predict the systematic component using the dependent variable – not the other way around[4]. To do this, the algorithm calculates the fixed-effects. Then the word-weights are estimated using a singular value decomposition (SVD) which collapses the estimates into a uni-dimensional space. The left-right singular vectors from the SVD are used as starting values of both the β and ω parameters. Then, using an expectation-maximization estimator, the model iterates until convergence on a low log-likelihood[5].

Like all other statistical models, this is essentially a model of the data generating process (DGP)[6] – specifically the *text* generating process. If misspecified, estimates will be inconsistent. Thus, pointing out that the texts should result from similar DGPs for Wordfish to work well (as both Slapin & Proksch 2014 and Bunea & Ibenskas 2015 do) is similar to saying that results will be inconsistent, if the model is misspecified. This comparison is helpful, as it can lead us to formulate hypotheses about the conditions under which the Poisson scaling model will perform badly, and, eventually, how to reparameterize it so as to estimate EU consultation documents better.

2.1 Hypothesis 1: Dissimilar Actors as an Omitted Variable

One of the most common ways to misspecify a DGP is by omitting important variables. Bunea & Ibenskas (2015: 4) point out that documents should be formulated "in similar institutional/organizational environments, authored by similar types of actors, serving the same communication purposes, and written in the same terminology". This is intuitive – if

[3]The Poisson distribution only has one moment and, thus, assumes that the mean and variance of the distribution is the same. This may lead to overestimation of certainty, but that is a problem that is outside the range of this text

[4]This actually puts it in the family of Item Response Theory models

[5]For formal details on the likelihood functions that are optimized, I refer to the original article by Slapin & Proksch (2008)

[6]King (1989: 8) makes this point

different types of organization systematically emphasize completely different aspects of a proposal; refrain from commenting on specific issues at all; and use completely dissimilar phrases, they will be incomparable within the Wordfish model. Thus, we can think of dissimilarity between actors involved in a consultation as an omitted variable in the Poisson scaling model, which makes its estimates biased and inconsistent. If we could assume that the omitted factor were only correlated with ω, then we would be able to derive the inconsistency. This would be a strong assumption, however, and the asymptotic bias is best thought to be unpredictable. We can, however, make the prediction that the inconsistency will generally be more severe when the omitted factor matters more. Hence, we would expect that *the performance of the Poisson scaling model will decline, as the included texts grow more different.*

2.2 Hypothesis 2: Informativeness of the Texts

If a statistical model based on the Poisson regression estimator is specified correctly, it can be shown to be consistent – as the sample size increases indefinitely, we would expect its estimates to converge in probability to the true parameters. As Wordfish uses words as observations, this places great demands on the length of the documents in the corpus. In accordance with this, Hjort et al (2015) find that Wordfish does considerably better, as the documents to be estimated increase in length. Similarly, Slapin & Proksch (2009) find that the number of unique words in the corpus increases the algorithms performance. Consequently, I will hypothesize that *the performance of the Poisson scaling model will increase with the length of the documents in the corpus.*

2.3 Hypothesis 3: The number of parameters

Lastly, Proksch & Slapin (2009) point out that Wordfish should perform better, when more documents are available. Increasing the number of documents should simultaneously increase the data available in each estimation and make it easier to distinguish relative positions as more contrasts are introduced (p. 326). However, each new document would require a unique parameter. Since maximum-likelihood estimators are consistent but not unbiased (King 1989: 78), based on statistical theory, we would expect that increasing the number of parameters, when the average document length is short, would actually cause *worse* estimates. Controlling for the *short document*-effect, however, we would expect that *the performance of the Poisson scaling model will increase with the number of documents to be estimated.*

3 Data and method

3.1 Case selection

To test these hypotheses I have selected three cases among the Commissions online consultations: *Reinforcing sanctioning regimes in the financial services sector*, *A New European Regime for Venture Capital* and *Review of the Investor Compensation Scheme Directive*. In the following, I will refer to them, simply, as Sanctions, Venture Capital and ICSD, respectively.

To infer the results to a broader universe of consultations, I have selected the cases so they to constitute hard tests of the Poisson scaling model. If it fares well in a hard test, it is more likely that it will produce good estimates more broadly.

All three consultations represent highly technical issues of how very specific regulatory questions should be handled, and how concrete legislative documents should be worded. Very technical wordings of the documents will make it harder to distinguish the relative positions, because the texts will not contain the very clearly partisan characteristics which Wordfish was developed to seek out. The cases also include a wide range of different actors - in all cases labour unions, corporations, NGOs, INGOs, intergovernmental organizations, consumer groups, individuals, employer associations, national governments and of course the Commmission participate. I expect that the position papers of these very different actors will be the results of diverse DGPs. Lastly, all cases include a large number of discrete regulatory dimensions[7]. In my hand-coding I identified, 5, 6 and 9 discrete issues in Sanctions, Venture Capital an ICSD, respectively. Since the Poisson scaling model was developed to estimate positions in a one-dimensional space, multidimensional issues will represent least likely settings for the algorithm (Bunea & Ibenskas 2015). Lastly, the average document length is relatively short in all cases, but the number of participating actors is quite high (48, 45 and 58). This leaves less information available to estimate each position. Especially the ICSD, where the average number of words in the documents is lower than 700, presents a least likely case for the Wordfish estimator. Table 1 summarizes the least likely selection criteria for the consultations.

To be sure, in any case there will be a wide range of fundamentaly unobservable characteristics that will influence the performance of the Wordfish model. The more dissimilar cases one chooses, the more likely it is that Wordfish will interact with these unobservables in unpredictable ways, thus making it impossible for us to make inferences about the model's performance and the determinants thereof. Therefore, I have chosen consultations that are similar in many aspects.

.

[7] I describe the hand-coding scheme and the dimensions more thoroughly in the appendix

Table 1: Description of cases

Consultation	Subject	No. of organisations	Dimensions	Unique words	Average no. words
National Sanctions	Dealt with how to harmonize sanctions and supervision regimes in the financial sector across the EU. The focus was on technicalities regarding how to develop common EU standard for specific rules.	48	5	2850	803.84
Venture Capital	Dealt with creating a common EU framework for venture capital funds. Primarily, the consultation focused on which specific rules in the UCITS venture funds should be exempt from.	45	6	3089	1313.37
ICSD	Happened in the wake of reforming the deposit compensations scheme. Specifically, this entailed working out which corporations should partake in the scheme, the level of compensation as well as whether monies held by defaulting third parties should be covered.	58	9	2709	691.14

First of all, I have only chosen cases from DG MARKT. There may be arbitrary differences between DGs in the way they word Green Papers, and in how it is custom for the interested parties to address the decision-makers. Also, DG MARKT deals with the most basic issues of EU governance – in essence, it is the DG tasked with promoting free movement of capital, labour and services among member states (DG MARKT 2015). This makes it an ideal reference point when it is only possible to code a small number of cases. Second, all three consultations were held within one year (2009-2010), thus limiting changes over time. These similarities will in some ways limit the generalizability of the results, but it will make unobservable interactions between Wordfish performance and specificities surrounding any single consultation less likely

3.2 Hand coding and scaling of positions

To hand-code the position of each document, I developed a coding scheme inspired by that of the Comparative Manifesto Project (CMP) (Volkens et al 2014). Quasi-sentences has formed the basis of the coding. They are generally understood as a textual unit expressing some policy preference. As such, a quasi-sentence can be either a natural sentence, a part of one or a text unit overlapping two natural sentences (Lowe et al 2010). The coding proceeded on a line-by-line basis, where I categorised each quasi-sentence into pre-defined and mutually exclusive categories. I defined these categories for each consultation individually based on the most important policy dimensions for the specific case. To do this, I first identified the main issues in the Commission's Green Paper. I then chose five position papers from the consultation and applied these codes. Based on what I found here, I then updated the coding list and coded all of the documents in the consultation according to it. In total, I have done a line-by-line coding of 151 documents averaging 9 standard pages.

To identify the overall attitude of each actor towards the policy proposal as such, I identified a common "mother dimension" subsuming all of the consultation-specific dimensions. For the consultations investigated here, I have defined the main dimension to be *re-regulation* vs. *de-regulation*. Thus, for all three cases, the main fracture between the interested parties was identified to be whether *more* or *less* rules should be imposed on actors affected by the specific EU regulation. Consequently, for each of the issues, the coding was done by categorizing the quasi-sentences according to whether the actor was in favour of more or less regulation with regard to that specific dimension. This identification of a single main dimension is of course imperfect, as there will be several important dimensions in each consultation. Table 2 describes how this coding scheme was applied to the Sanctions consultation. For the descriptions of the codes used in the other two consultations, I refer to the appendix.

Table 2: Coding of the Sanctions consultation

Dimension	Description of dimension	Definition of reregulation
Harmonization of rules	Should the principle of maximum or minimum harmonization be used	Quasi-sentences advocating maximum harmonization defined as reregulation
National vs. Supranational regulation	The degree to which the new legal framework should be decided supranationally through regulation or nationally through directives.	Quasi-sentences advocating supranationality defined as reregulation
Retail vs. investment banking	To what extent these new rules should apply to the entire financial sector or only investment banks.	Quasi-sentences advocating broad coverage defined as reregulation
Consumers vs. Producers	If non-commercial consumers should be protected to a larger extent than corporate consumers of financial services.	Quasi-sentences advocating consumer protection defined as reregulation
Prosecution	If both physical and legal persons should be prosecutable, and whether effort should be made to prosecute individuals	Quasi-sentences advocating broad prosecution defined as reregulation

To scale positions, θ, from the hand coded documents, I follow Lowe et al (2010) in using the logged relative balance of quasi-sentences[8]:

$$\theta_d = log(\frac{D_d + .5}{R_d + .5}) \quad (1)$$

Where D and R represent the count of quasi-sentences advocating de-regulation and re-regulation, respectively, in document d. This means that the effect of adding a quasi-sentence to either D or R will decrease on the margin. To obtain uncertainty estimates around each of the hand coded positions, I follow Benoit et al (2009) in using a non-parametric bootstrap function.

3.3 Empirical strategy

To test my predictions, I ran the Poisson scaling model several times on different subsets of the consultation documents. I began by running it on all documents in each consultation, then I randomly removed five to eight documents from the total number of texts. In each consultation, I selected one specific type of organization, which I did not remove. In each case, I chose not to remove the type of actor which was most active in the consultation. In the case of Sanctions, I only removed non-corporations. In Venture Capital, I removed documents from actors that were not venture capital funds. In the case of the ICSD, I 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, my hope is that this will allow me to estimate the causal impact of altering the composition of the different corpora. As the removal is not completely random, controls will have to be included, of course.

This will allow me to estimate the impact of a change in, respectively, the differences in DGPs across actors; the number of parameters; and document length – the central hypotheses of this paper. Both causal estimates and further generalizability could be achieved through Monte Carlo simulations (MCS), but to this, I would have to assume a priori how Wordfish performance is affected by the factors I consider. Because this framework uses real-world texts, I avoid having to make these assumptions about the DGP of consultation texts. This implies, however, that my study could provide valuable information to use in future MCS' of Wordfish performance in EU consultations.

When comparing the hand-coded positions to Wordfish estimates, I will use the rank-order

[8]They argue that this represents a "linguistically superior" approach (Lowe et al 2010: 123) to that used in the Comparative Manifesto Project (Volkens et al 2014), which uses the absolute difference between quasi-sentences advocating left and right policies normalized by the total number og sentences

correlation coefficient Spearman's ρ. I mainly do this, as it is less sensitive to outliers than the Pearson's product-moment correlation coefficient. Hence, if there are a few extreme observations where agreement between hand-coding and Wordfish is very strong (or very weak) these will not affect the overall correlation as much. This is important as in some of the Wordfish estimations, there will be included relatively few documents. As we shall see later, it is also here Wordfish generally performs best, and I want to avoid extreme observations and outliers driving the high correlations. When I in the subsequent sections refer to Wordfish performance, I will mean the rank-order correlation between hand-coded positions and Wordfish estimates.

The operationalizations of the hypotheses regarding the number of parameters and document informativeness is relatively straightforward. I will use, respectively, the number of documents included in the estimation and the average number of words in the documents. To measure the similarity of the DGPs of the actors included in the estimation, I will use the average correlation among the documents. The rationale is that higher average correlations will imply more similar documents and, hence, more similar DGPs. The average correlation subsumes the different aspects, which Bunea & Ibenskas (2015: 4) emphasize generally will imply comparability "with respect to their data generating process"[9]. Thus, I will not have to rely on more superficial measures of, for instance, actor type to approximate dissimilar DGPs.

As means of control, I will include the total number of unique words in each iteration and whether stopwords are included or not. To strengthen the counterfactual interpretation, I include dummies for the consultations so as to only compare performance of the Wordfish algorithm within each of the cases I have studied. As the number of dimensions remain the same within each consultation, this will be controlled for through these consultation fixed-effects. In some estimations, I also include the effective number of different types of organizations as measured through the inverted Hirschman-Herfindahl-indeks (HHI)[10]. As I was not able to calculate it for the Sanctions consultation in time for my deadline, I only use it at as means of control and do not include it in all estimations.

Table 3 provides some descriptive statistics on the variables of interest across iterated removals of documents. Figure 1 gives a graphic impression of their distributions.

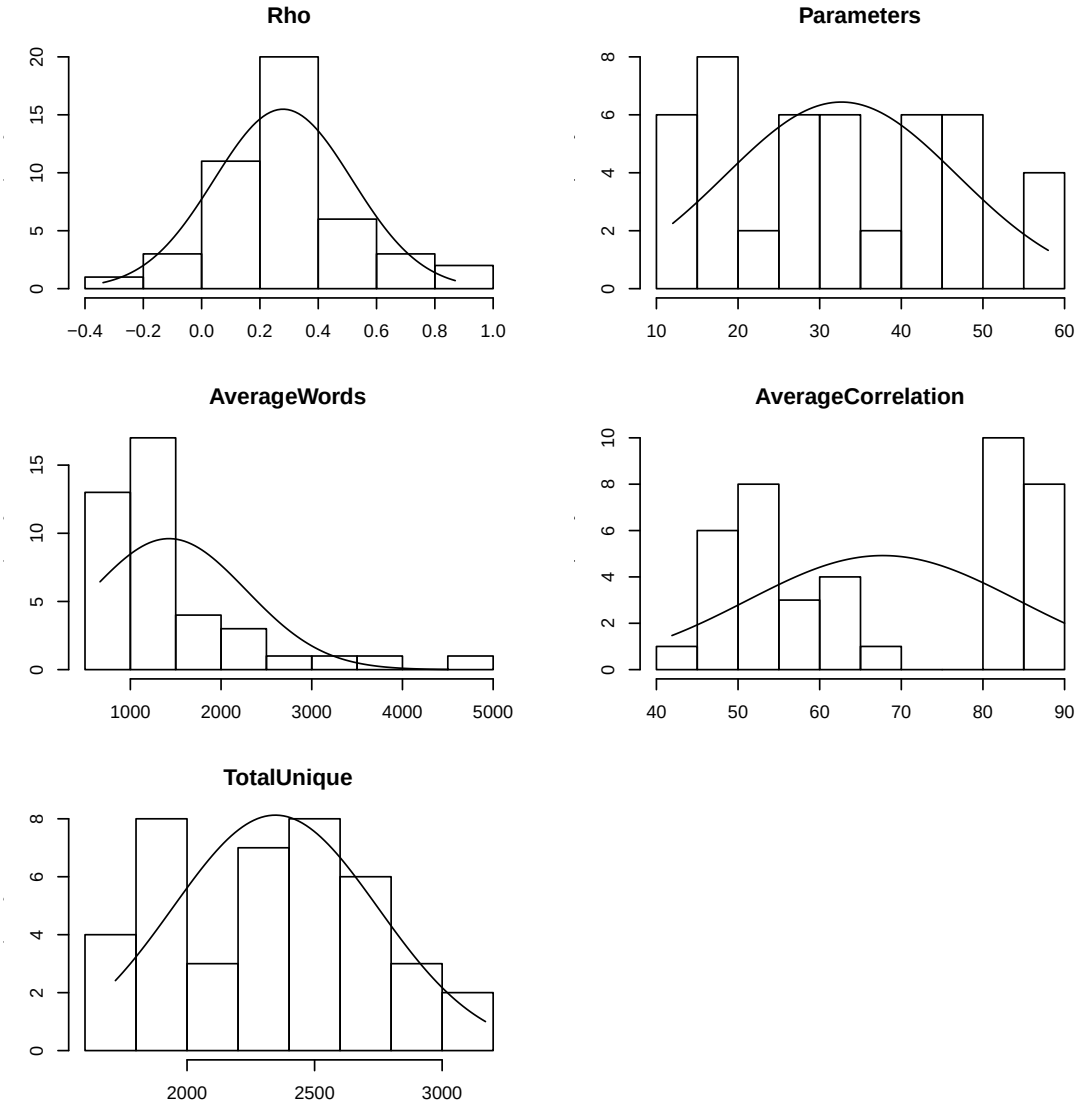
[9]They emphasize institutional or organizational settings, actor types and vocabularies as important aspects of the DGP.

[10]The inverted index is calculated by dividing the squared sum of the count of different organizations by the sum of the squared count: $\sum(O_i)^2 / \sum {O_i}^2$

Table 3: Descriptive Statistics

Statistic	N	Mean	St. Dev.	Min	Max
Rho	46	0.279	0.237	−0.338	0.870
Parameters	46	32.652	14.247	12	58
Average Words	46	1,426.646	851.200	665.480	4,505.667
Average Correlation	46	67.734	16.621	41.900	90.000
Total Unique	46	2,346.146	402.566	1,719	3,170
Stopwords	46	0.500	0.506	0	1
Consult	46	1.652	0.822	1	3

Figure 1: Distributions of main variables



Note: The curves are for comparison between the observed distributions and how a theoretical normal distribution would look in each case.

3.4 A model of Wordfish performance

This leads me to consider the following model, which I will estimate using ordinary least squares (OLS) regression.

$$\rho_c = \delta_0 + \delta_1 P_c + \delta_2 D_c + \delta_3 L_c + \delta_4 X + \theta + \varepsilon_c \tag{2}$$

Where the dependent variable ρ denotes Wordfish performance in a consultation c. P is the number of parameters to be estimated, D is my measure of DGP similarity and L is average document length – our variables of interest. Thus, δ_1 through δ_3 measure the correlation between the main variables and Wordfish performance. X is a vector of controls, θ is a set of consultation fixed-effects, δ_0 is the intercept and ε is the unobserved error term.

The analysis in the subsequent sections will fall in two parts. First, I will investigate whether or not Wordfish can at all achieve high correlations with human coding across my iterated exclusion of documents. Second, I will test my model of Wordfish performance on a dataset comprised of characteristics regarding each iteration.

4 Validating Wordfish Estimates

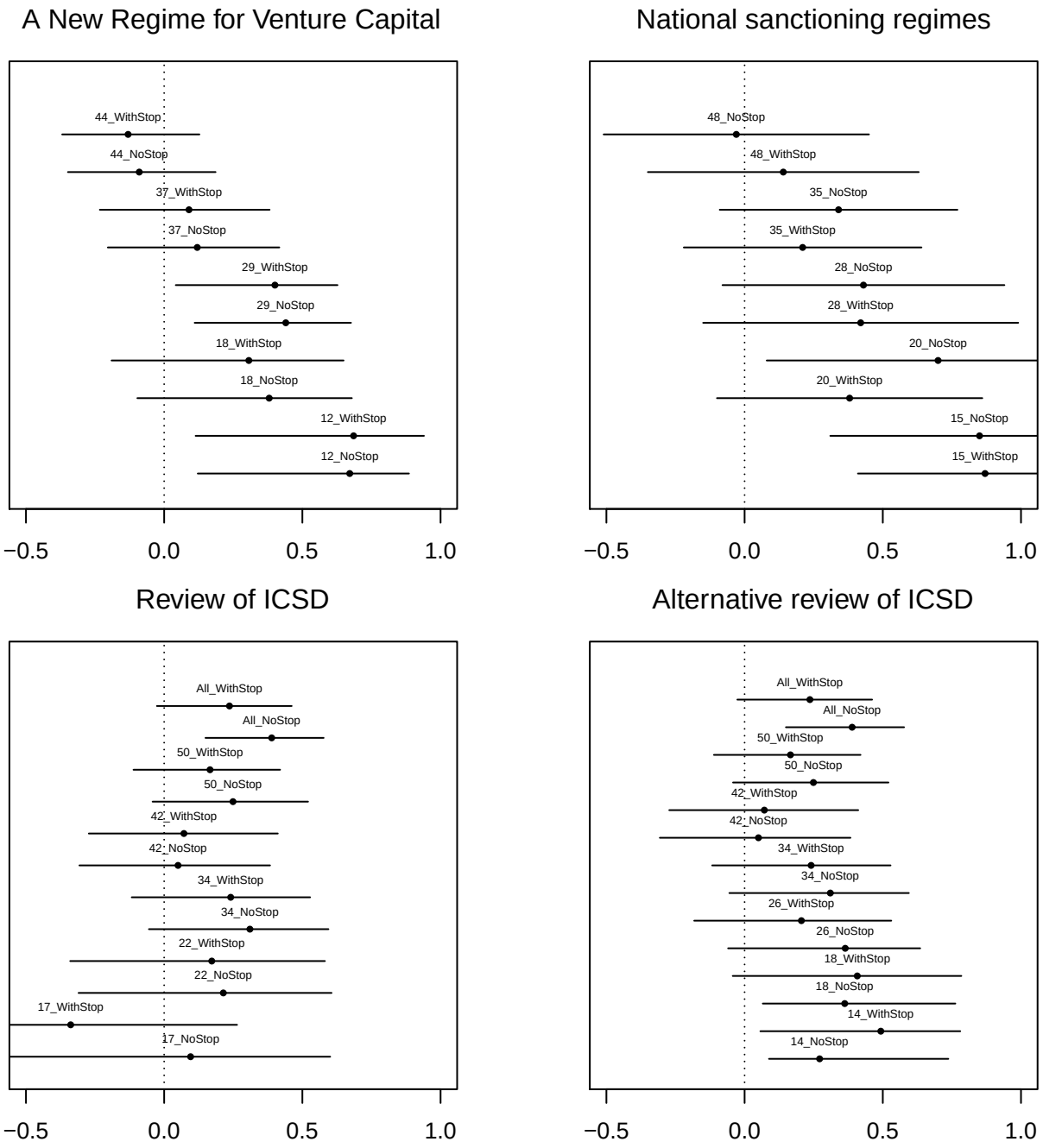
4.1 Overall correlations

In Figure 2 below, I have plotted the correlations between my hand-coding and the Wordfish estimates. I plot each iteration, where I remove documents from the Wordfish estimation. The text above each point estimate identifies how many parameters that have been estimated and whether or not stopwords were excluded or not.

As it can be seen, high correlations are obtained at some point for all cases – the highest ρ for the three consultations is $\in [0.5;\ 0.87]$[11]. But it is also common to all cases that the estimates from the Poisson scaling model for some iterations actually correlate negatively with the hand-coded estimates. The clearest pattern emerges for the consultations on Venture Capital and Sanctions. In both cases, I achieve statistically insignificant and negative correlations, when including all of the documents. The correlations then increase almost linearly for each time I exclude documents from the Wordfish estimation. ρ is 0.87 and 0.7 for in the best-performing iteration of Sanctions and Venture Capital, respectively. For ICSD, however, the pattern is not as straightforward. Here, the estimations including all documents yield relatively good Wordfish performance, with a statistically significant correlation of just above 0.45. When excluding documents at random, the

[11] The overall pattern is robust to using Pearson's product-moment, but when using it the lowest of the high correlations is substantially larger as it > 0.6

Figure 2: Correlations between hand-coding and Wordfish



Note: The Spearman's ρ rank-order correlation coefficient is used. CIs are calculated using 95% non-parametric bootstraps. The percentile method is used.

correlation then decreases and ends up being negative. I suspected that this was because the average number of words in each document in this consultation is very low, as mentioned previously. When excluding documents at random, I run the risk of leaving out essential information. So I proceeded with an alternative strategy, where I – after the iteration with 34 parameters – excluded documents with fewer words than 800. The last

panel plots these alternative iterations. As it can be seen, Wordfish performance in this scenario is lowest for iteration with 42 included documents. It then improves and tops with a statistically significant $\rho > 0.5$.

4.2 A qualitative examination of Wordfish performance

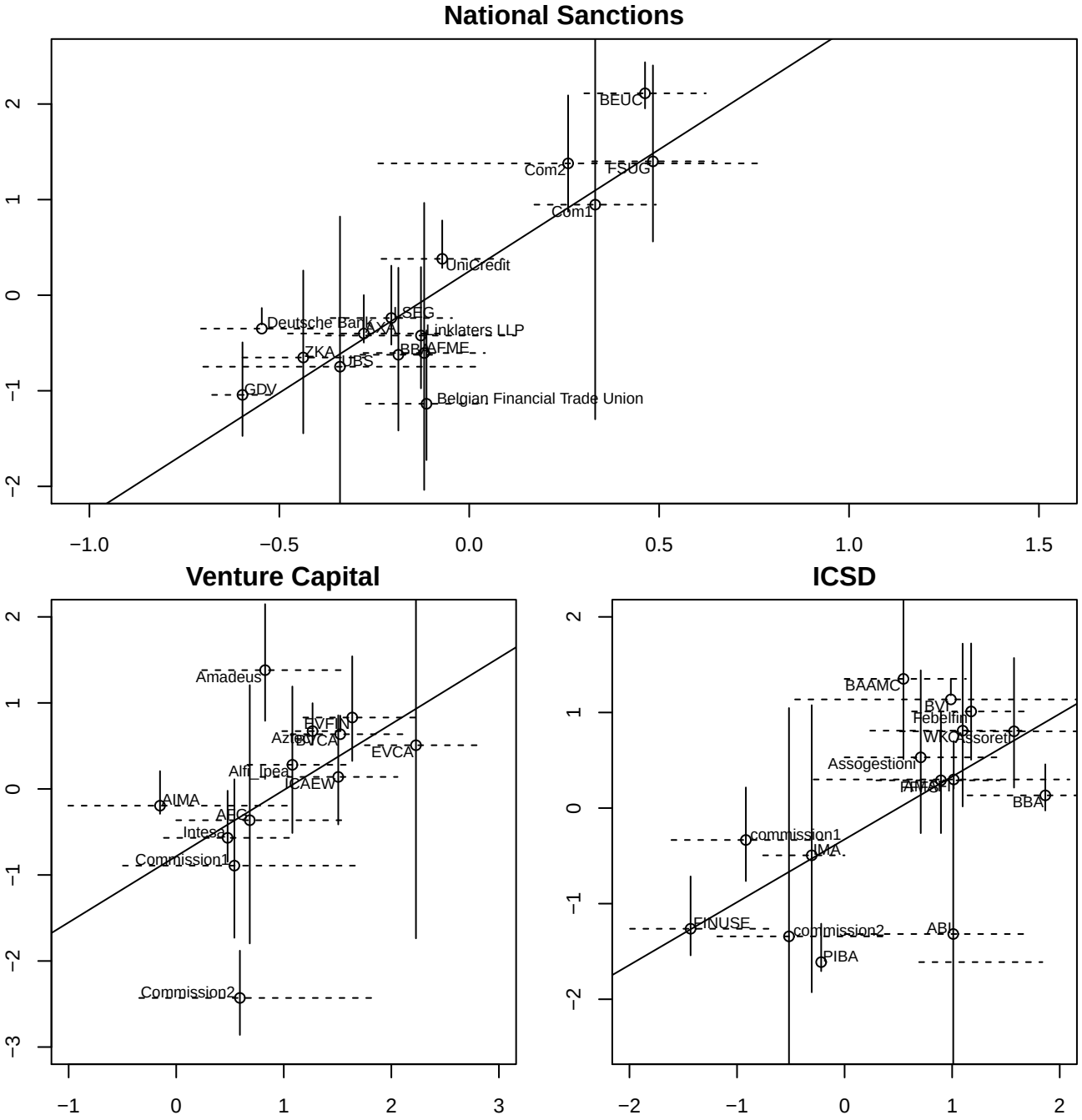
Figure 3 provides a qualitative assessment of how the Poisson scaling model places the interest groups vis-á-vis human coding.

If a document is placed exactly on the fitted line, Wordfish perfectly predicts the hand-coded position and vice versa. Therefore, if either of the CIs overlap the fitted line, we cannot statistically reject, that the two approaches are in agreement. It is clear that for all three consultations, the actors mostly fall neatly around the prediction line. But in all consultations there are two organizations where the disagreement between the two approaches is statistically significant.

For the Sanctions consultation, this is the case for the consumer organization BEUC and the Belgian financial labour union. The disagreement regarding BEUC is however, not very substantial. For the Venture Capital case, we observe statistically significant disagreement regarding the estimates of the Amadeus venture fund and the Commission. Wordfish' misplacement is substantial and especially problematic regarding the Commission, as good estimates of influence demand the Commission's positions to be well estimated. For the ICSD, the Bulgarian Association of Asset Management Companies (BAAMC) and the Irish organization Professional Insurance Brokers Association (PIBA) are statistically significantly misplaced. The point estimate for Association of British Insurers (ABI) is also very different between the two approaches, but the disagreement is too uncertain to be statistically significant. In all cases of misplacement, inspection of the document revealed likely causes of the disagreement. For instance, the BAAMC spent most of their position paper outlining how the ICSD is currently implemented in Bulgaria, and thus left very little information to actually estimate its position regarding the Commissions new proposal. An assessment of each case would, however, be too lengthy for the purposes here.

To sum up this section, Wordfish ended up performing well in estimating positions in all three consultations. A qualitative examination revealed that only a few documents in each consultations were directly misplaced, and – importantly – there is mostly agreement regarding the position of the Commission. This is in spite of the multidimensional nature of all the cases, which Bunea & Ibenskas (2015) emphasize as one of the major problems confronting the use of Wordfish in EU consultations. But Wordfish also performed poorly under most conditions. It seems, however, that performance follows predictable patterns.

Figure 3: Qualitative examination of cases



Note: Wordfish 95% CIs are obtained from parametric bootstrapped with assumed Poisson distribution. Hand-coding uncrtainty is estimated using bootstrapped non-paramtric 95% CIs. The dashed horisontal lines represent uncertainty surrounding the human coding, whereas the solid vertical lines are Wordfish CIs. For both types of bootstrap, the percentile method is used. Positions from the best performing iteration are plotted, so for ICSD the estimates are obtained after non-random exclusion.

5 Testing predictions about Wordfish performance

In Table 4 below, I test my predictions regarding Wordfish performance. Columns one through three provide the bivariate correlations between my three variables of interest

and Wordfish performance. As it can be seen, the number of parameters is negatively correlated with the performance of the Poisson scaler, whereas average document length is positively correlated with performance. This implies that without conditioning on other variables, Wordfish performance would be expected to increase with document length and decrease as more documents are to be estimated. The correlations are statistically significant at the 1 and 10 percent levels, respectively. The average correlation among documents, however, exhibits no discernible association with Wordfish performance, and even enters with the opposite sign of my expectation. Column four compares Wordfish performance in the three consultations under investigation. This shows the same pattern as we have previously seen: Wordfish performance in the Sanctions consultation is substantially better than in the ICSD consultation. Performance in the Venture Capital consultation is somewhat better than in the case of ICSD, but the difference is neither nor substantial nor statistically significant at conventional levels.

Column five models all of the variables together and introduces total unique words and a dummy for stopword removal as controls. This changes the previous results dramatically. The number of parameters is now positively associated with Wordfish performance, and the coefficient is more than double the size of its standard error, which indicates statistical significance at the 5 percent level.

From this model, I would predict that the correlation between Wordfish estimates and hand-coded positions would increase by 0.28 ρ, if the number of parameters to be estimated increased from 18 to 43 (that is, from the first to third sample quantile). This indicates that the negative correlation in the bivariate setting was driven by the fact that when I removed documents from the estimation, these documents tended to be both shorter than and more dissimilar to the remaining ones.

An increase of the average document length from the first quantile (858) to third third (1,500), would lead us to expect and increase in Wordfish performance by 0.26. The correlation has more than quadrupled and has at the same time become more precise. Thus, it is now statistically significant on the one percent level.

Table 4: Testing predictions of Wordfish Performance

	Dependent variable:						
	Wordfish Performance						
	OLS						*Tobit*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Parameters	−0.007***				0.011**	0.013**	0.011***
	(0.002)				(0.004)	(0.005)	(0.004)
Average Words		0.0001*			0.0004***	0.0004***	0.0004***
		(0.00004)			(0.0001)	(0.0001)	(0.0001)
Average Correlation			−0.001		0.027***	0.025***	0.027***
			(0.002)		(0.008)	(0.008)	(0.008)
Total Unique Words					−0.0002	0.00005	−0.0002
					(0.0001)	(0.0002)	(0.0001)
HHI						−0.015	
						(0.014)	
Stopwords removed					1.168***	1.146***	1.168***
					(0.277)	(0.281)	(0.249)
Venture Capital				0.069	−0.459***	−0.590***	−0.459***
				(0.084)	(0.160)	(0.149)	(0.143)
National Sanctions				0.213**	0.277**		0.277***
				(0.084)	(0.102)		(0.092)
Constant	0.521***	0.159**	0.354**	0.218***	−2.670***	−3.006***	−2.670***
	(0.080)	(0.069)	(0.154)	(0.044)	(0.825)	(0.863)	(0.740)
R^2	0.198	0.074	0.009	0.130	0.611	0.651	
Log Likelihood							21.971
Chow test	12.42***	6.7089***	4.1593**		3.0354**	2.6085**	

Note: Standard errors in parentheses. *; **; and *** represent statistical significance at the 10%, 5% and 1% levels, respectively. ICSD is the reference category for the consultation dummies

Also regarding the effect of document similarity we observe an interesting pattern. After controlling for the host of different factors, it now exhibits a very substantial and positive correlation with Wordfish performance. From these results, I would predict that an increase in the average correlation among documents from the first sample quantile (0.52) to the third (0.84), would be accompanied by an increase in the correlation between Wordfish estimation and hand-coded positions by 0.87 ρ. A very substantial correlation indeed. Inspecting the associations between the average correlation among documents and the other explanatory variables in the model, I found a very substantial negative relationship between removing stopwords and document similarity (r $\approx$ -0.94, and $\rho \approx$ -0.86). Including only these two variables yields substantial and statistically significant correlations with Wordfish performance, however not as large as in the primary results reported in column 5[12]. This indicates that the effect of removing stopwords suppresses that of document similarity. My interpretation of this finding is that removing stopwords from the documents reduces the similarity of the corpus, but it does so by eliminating common words that do not help us to distinguish political positions and, consequently, infuse the estimation with noise. Taking the day-to-day vocabulary into account, thus, allows us to estimate the actual effect of including documents that result from (dis)similar DGP's. These results are also reflected in the very substantial correlation betweem removing stopwords and Wordfish performance.

The only statistically insignificant variable in the model, is the total number of unique words. This result is somewhat puzzling given previous findings in the literature. Utilizing Monte Carlo Simulations, Proksch & and Slapin (2009) find that Wordfish does substantially better when there are more unique words in the corpora. In my findings, the number of unique words is statistically insignificant and enters into the equation with a negative sign. I suspect that this is a result of two effects. 1) Proksch & Slapin (2009) only simulate scenarios with relatively few unique word (300 at the most). It could very well be that the positive effect of increasing the number of unique words is only there when there are very few words in the first place. 2) A large amount of unique words can under certain circumstances be indicative of very different DGP's. Highly similar documents would to a larger degree use the same words. The result would be fewer unique words, when including more similar documents – this would actually lead to *better* performance due to fewer differences in the DGPs.

One last notable point is the differences between the dummies for the different consultations. In column four I estimated that Wordfish performed better when applied to the Venture Capital consultation than was the case for the ICSD. In column five, this pattern is dramatically reversed. Now, we would expect Wordfish to perform substantially worse in the Venture Capital consultation – and the difference is very significant statistically

[12]These results are, of course, available, but to save space, I haven't reported them here

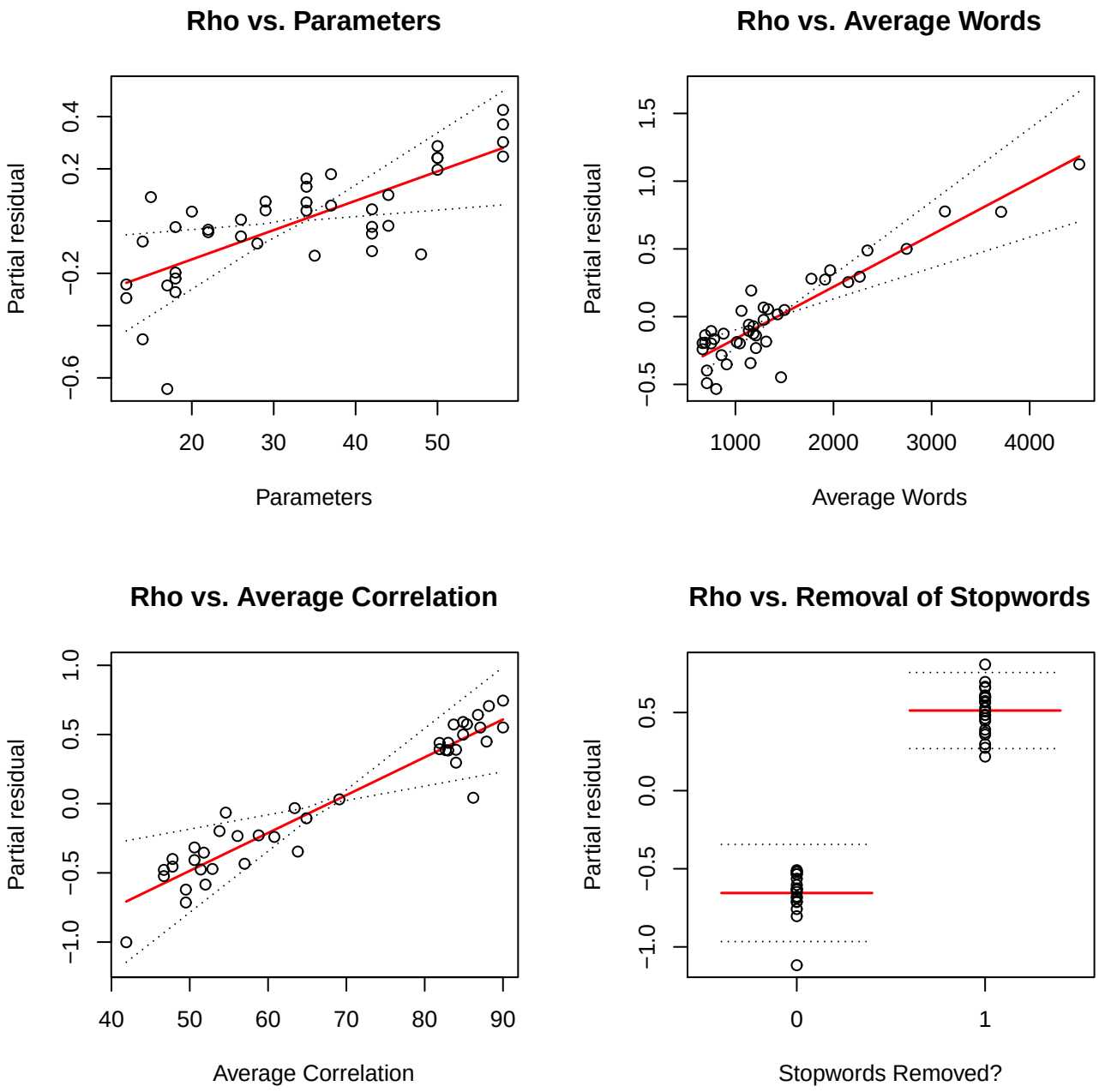
speaking. This indicates that – given the number of parameters as well as document similarity and length – there is nothing inherent in the ICSD consultation, which makes for bad Wordfish estimates. This indicates that the bad performance can be explained by the factors in this model.

The sixth column includes the effective number of different organization types included in each Wordfish estimation. As mentioned, I did not have time to calculate this measure for the Sanctions case, so it drops out of the model. The results as run on this different subset of iterations, however, remain the same. It is interesting that the effective number of different organizations types is far from being statistically significant. The seventh and last column re-runs the model from column five using the *tobit* estimator. Seeing as the dependent variable is naturally bounded at $\rho \in$ [-1; 1], I might get inefficient or overconfident uncertainty measures when using *OLS*, because it assumes that $\rho \in$ [-∞; ∞], when calculating the variance-covariance matrix. The results indicate that there may be some efficiency gains in using the *tobit* estimator.

Lastly, the table also reports R^2, log-likelihood and a chow test. Whereas the loglikelihood does not say much without other models to compare it to, the $R^2 \approx 0.61$ indicates a good fit of the model. This pattern is also obvious when looking at the plots in Figure 4. It shows the associations between each of the four statistically significant variables and Wordfish performance after controlling for all other factors included in the model represented in column 5. As it can be seen, the residuals fall neatly and closely along the fitted lines. The plots also give a better feeling for exactly how substantial the associations are.

I have calculated the chow test to gauge whether or not the estimated slopes differ substantially across the consultations. We would hope that – conditional on the other explanatory variables – the model behaved similarly in all consultations, but as the significant test indicates, this is not the case. Actually, the effect of all variables – except for the average correlation among documents and the stopwords dummy – are stronger for the ICSD consultation. This indicates that the positive effect of removing parameters and increasing the average document length is stronger when the odds are against the Wordfish estimator from the beginning (recall that the ICSD consultation was characterized by short documents and many actors). How far this result can be inferred is unclear, however, seeing as it could be an artefact of the ICSD consultation specifically. It does indicate, however, that I was not successful in selecting cases that were similar enough to avoid interactions with unobservables.

Figure 4: Illustration of model results



Each plot represents the association with all other variables in the model partialled out. Dashed lines represent pointwise CIs for expected values.

6 Gauging consequences of measurement error

As mentioned previously, automated textual analysis is of particular interest to the political study of SIGs because of its potential for estimating political influence. The difference between the estimate obtained through textual analysis and true influence can be conceptualized as measurement error. The consequences of the particular type of measurement

error associated with automated textual analysis are, however, not well investigated. To gauge these consequences, in this section I present the results from a number of Monte Carlo experiments. Since the field so far has mainly tried to explain political influence, I will focus my attention to the case where influence is modelled as the dependent variable.

In the classical analytical approach to gauging the consequences of measurement error, it is normally assumed that the error laden variable is perfectly correlated with the true measure, but also contains some noise (Wooldridge 2013; 308). Letting y^* be the true dependent variable and y be our error laden measure we get

$$y = y^* + e \tag{3}$$

Under these assumption, it can be shown that using the mis-measured dependent variable does not induce bias into our estimators, but only increase their variance (Wooldridge 2013; 308). In this paper, I have shown that given the right methodological choices, Wordfish estimates can approximate the preferences of political actors. The estimates can be highly correlated with the true political positions, but the differences cannot be ascribed purely to sampling error. Consequently, the results from the classical approach to measurement error do not hold.

In my simulations, I have dealt with the case, where influence is measured as a binary variable, as this has been the case, when using Wordfish to study SIGs so far (Klüver, 2011; 2012). To better get at the direction of potential biases, I use two independent variables, one with positive impact, and one with negative. The DGP is assumed to be

$$I_o^* \sim binomial(\mu_o, \sigma^2)$$
$$\mu_o = \beta_0 + \beta_1 X_{1o} + \beta_2 X_{2o} + \phi_o$$

So that the odds that organization o is influential is drawn from binomial distribution with mean μ and variance σ^2. μ depends upon two independent variables and an idiosyncratic error term ϕ. The relations between the variables are

$$\beta_1 = -0.45, \beta_2 = 0.5$$
$$X_{1o} = normal(2, 5), X_{2o} = normal(0, 1)$$
$$\phi_o = uniform(0, 1)$$

I simulate a true measure of influence with 10.000 observations on the basis of this DGP. In an applied setting researchers will try to make sure that their composition of texts is such that Wordfish performs well. In any single case, however, we cannot know whether

or not we are successful. In order to approximate this, I divide the 10.000 observations into subsets of 2.000 each, and generate five proxies that correlate with I_o^* to a varying degree:

$$I_{o1} = -0.2 * I_o^* + \epsilon_1, I_{o2} = -0.1 * I_o^* + \epsilon_2$$
$$I_{o3} = 0 * I_o^* + \epsilon_3, I_{o4} = 0.55 * I_o^* + \epsilon_4$$
$$I_{o5} = 0.9 * I_o^* + \epsilon_5$$

Finally, I let the influence proxy be $I_o = I_{o1} + I_{o2} + I_{o3} + I_{o4} + I_{o5}$. I run the simulations in turn using first the full range of the proxy, then using only the range of the proxy that is non-negatively correlated with actual influence, and finally using only the range of the proxy that is strongly correlated with influence. The results from these simulations are presented in Figure 5.

We observe a clear and strong attenuation bias. For both the estimate of β_1 and β_2 logit strongly underestimates the actual effects of both of the independent variables. In all instances the mean of the simulations lies much closer to zero than to the actual effect, no matter whether the effect is positive or negative. In the case where the proxy is least well correlated with actual influence (the yellow and blue density curves), the effects of both independent variables are virtually indistinguishable from zero. I have tried running simulations with 1.000 observations in each of the 10.000 runs, but this only decreases the variance of the estimator around the same biased mean. These results are available upon request. This indicates that the results are not an artefact of few observations, and the logit estimator is both biased and inconsistent when there is measurement error of this kind in the dependent variable.

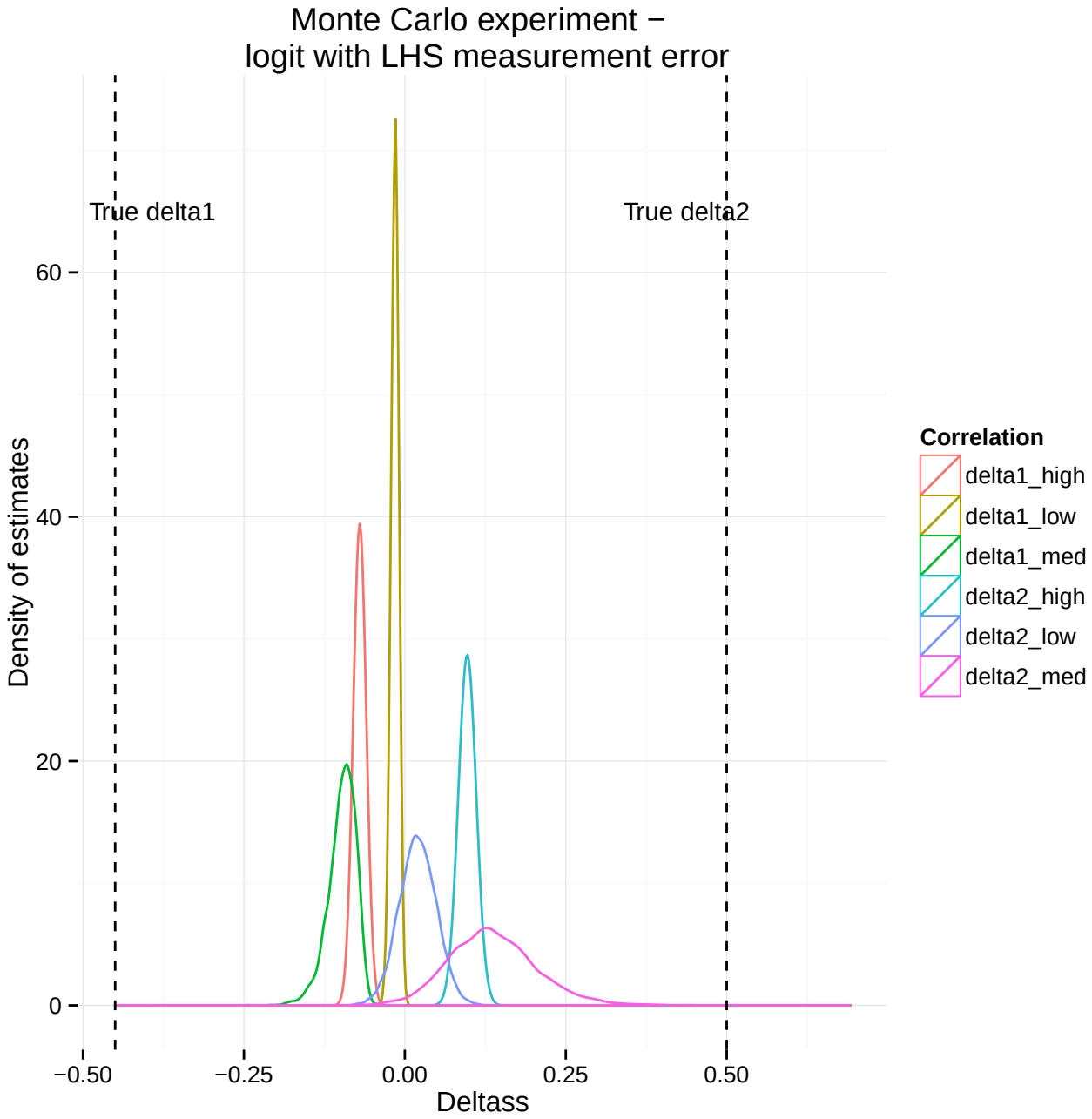
A somewhat puzzling finding is that the logit estimator is *more* biased in the scenario, where the correlation between the proxy and actual influence is $\in [0.55, 0.9]$ (the red and mint density curves) than in the scenario where $\in [0, 0.9]$) (the green and purple density curves). So the scenario where the proxy is more closely related to the actual value is not the scenario, where the estimator performs best. All results hold when using the OLS estimator, and these alternative results are available upon request.

I have no explanation for this ...

The implications remain clear, however – unlike in the case of classical measurement error in the dependent variable, which in most cases only increases the variance of the estimator, this more general type of error leads to attenuation bias. When using Wordfish to estimate political influence of SIGs, this implies that – given the right composition of

the text corpora – we will on average estimate effects of covariates on influence that are much lower than their true impacts.

Figure 5: Results from a Monte Carlo Experiment



Note: Monte Carlo experiments run with 10.000 simulations and 300 observations in each run. True $\beta_1 = -0.45$. True $\beta_2 = 0.5$. The logit estimator is used. Estimations are average marginal effects, when the independent variables are increased from their means

7 Concluding remarks

How brave is the new world of textual analysis with regards to interest group politics? The findings presented here echo those of Bunea & Ibenskas (2015) – careful thought and qualitative assessment should precede the application of the Poisson scaling model to EU consultations. But my results also indicate that some enthusiasm among scholars studying SIGs may well be warranted. If the composition of the text corpora under consideration is right, then Wordfish can be expected to perform well. This, in turn, implies that only certain research questions in the realm of SIG politics would be open to investigation using the Poisson scaling model.

The contribution of this paper is at least three-fold: 1) Through a thorough cross-validation, where I compared positions estimated using the Poisson scaling model to hand-coded positions, I have shown that Wordfish performance can range from dismal to very good. 2) Wordfish performance follows predictable patterns – that is, based on known characteristics of the documents, whose positions we would want to estimate, we can predict whether or not Wordfish will produce viable estimates. 3) The mis-measured proxy of influence, which we obtain through the use of Wordfish, induces attenuation bias into our estimators, when we use our proxy as a dependent variable. This indicates that – if we can assume exogeneity of our independent variables – estimates of the impact of covariates on the political influence of SIGs is likely to be lower-bound, even if we design our text corpora in an optimal way.

Based on these findings, I can conclude this paper with a set of "do"'s and "don't"'s, which would provide viable guidelines for scholars seeking to estimate SIG positions using Wordfish. The following rules of thumb are based on the results reported in figure 4:

1. *Make sure that documents are sufficiently similar.* Very different documents are indicative of diverse underlying DGPs producing the texts. This biases the Wordfish estimator. How to identify the documents that can be included together is not a simple task, however. One way could be to iteratively exclude documents until the ones remaining are sufficiently correlated. If stopwords are excluded, a sensible rule of thumb seems to be, that the average inter-document correlation should not get lower than ≈ 0.55. If stopwords are not excluded, correlation should not be lower than ≈ 0.8. With regard to differences in DGPs, it is an interesting result that the effective number of different types of actors was not associated with Wordfish Performance, as Bunea & Ibenskas (2015) concluded that – among other things – organizational differences present a problem for Wordfish performance.

2. *Always make sure that the documents included in the estimation are of sufficient length.* Specifically, it seemed that including single documents with fewer words than 1,000 lowered Wordfish performance. Also, when average document length in the corpus

gets lower than roughly 1,500 words, it lowers Wordfish performance. These two cut-off points seem to be valid rules of thumb – at least for the sample investigated here.

3. *Think carefully about the number of documents to include.* As we have seen, holding average inter-document correlation and average document length constant, increasing the number of parameters for Wordfish to estimate, increases the algorithms performance. However, it is very important to note that this only holds if the new documents are of sufficient length and similarity. MLE is simply not feasible with too few observations.

4. *Exclude stopwords.* It is common practice to exclude when doing automated textual analysis, and the analysis here indicates that Wordfish performance under most circumstances increases dramatically when it is done.