Scaling things we can count[*]

Will Lowe
Princeton University

DRAFT: April 2016

[*] Thanks to audiences at the universities of Mannheim, Konstanz, and Princeton for pointing out which parts of the paper were incomprehensible, to Thomas Dauebler for pointing which part was false, and to Brandon Stewart for forcing his class to read it. Any remaining errors are probably why I never got a job in the UK.

Abstract

This paper reviews some of many methods political scientists have used to get positions out of textual data and argues that they are the identical to, approximations to, or special cases of a single count data unfolding model that produces low dimensional representations of patterns of relative emphasis. Ideological positioning via a logic of relative emphasis not only generates useful scaling technique for anything we can count, but also vindicates (some of) the key intuitions of the saliency theory of party positioning. I also use the model to motivate biplots as an underused graphical tool for understanding scaling models.

"There should be one – and preferably only one – obvious way to do it"[1]

Political scientists often need to infer policy positions from text, and have over the last 30 years proposed, developed, or borrowed methods and techniques of steadily increasing sophistication. The available methods seem very heterogenous. Many statistical frameworks have been applied, from factor analysis (Warwick 2002), unfolding (Elff 2013), and item-response theory (IRT Bakker 2009; Albright 2008) to heuristic or numerical approaches such as Laver, Benoit, and Garry 2003, Gabel and Huber 2000, and Lowe et al. 2011. Researchers have distinguished between 'a priori' scaling methods that assume that the positions of at least some texts or words are known (Laver, Benoit, and Garry 2003; Pennings and Keman 2002) and 'inductive' methods in which positions and other parameters are simultaneously estimated (Slapin and Proksch 2008; Monroe and Maeda 2004).

These methods have been applied to a range of units, e.g. counts of word types (Slapin and Proksch 2008; Monroe and Maeda 2004; Laver, Benoit, and Garry 2003), sentences and quasi-sentences (Budge, Robertson, and Hearl 1987), dictionary-based content analysis categories (Laver and Garry 2000), metadata tags (König and Luig 2009), and – most popularly in comparative politics – subsets of the category scheme developed by the Budge, Robertson, and Hearl's Comparative Manifesto Project (e.g. Pennings and Keman 2002; Lowe et al. 2011; Elff 2013; König, Marbach, and Osnabrugge 2013). These applictions imply though never quite make explicit, and important desideratum: a suitable scaling models should apply equally well to anything we can count. This issue is taken up in more detail in the appendix.

In the face of this variation researchers might reasonably ask: Is there a common set of substantive and statistical assumptions underlying these text scaling methods? What am I committing to substantively by choosing one method over another? Empirical reviews of the consequences of these different methods exist for particular cases e.g., Dinas and Gemenis (2009) compare Greek and Klemmensen, Hobolt, and Hansen (2007) Danish party positions and Klüver (2009) look at interest groups in relation to the European Commission. But these reviews do not in general attempt to provide insight into the methods themselves.

The first goal of this paper is to provide a unifying theory for existing text scaling methods. I will argue that there is essentially *only one way* scale the kind of count data that textual data implies, and that the methods above are particular implementations, special cases, or computational approximations to it.

This is possible because there is a *common logic* to scaling count data. Specifically, I show that existing methods reflect the idea that document positions are based on low dimensional reconstructions of patterns of relative emphasis by deriving them from a model that explicitly does so.

Low dimensional reconstruction is common theme in latent variable modeling, of which factor analysis and IRT models are perhaps the most widely used instances in political science. However, scaling texts is essen-

[1] from The Zen Of Python. Quote continues "(although that way may not be obvious at first unless you're Dutch.)"

tially a problem of count data, for which these approaches are not appropriate. Methods like factor analysis assume conditionally Normal responses and the IRT models borrowed from the analysis of roll call votes assume that the data is binary. In contrast, for count data it can only be the relative emphasis of one word (or category) over another that signals position. Consequently the empirical quantities that need low dimensional reconstruction in count data scaling are not correlations but associations, so the appropriate model class will be association models and their relatives.

One interesting substantive corollary to the argument will be that all these methods embody – and to the extent they work, also vindicate – some but not all of the key claims for the saliency theory of positioning (Budge 2001). This connection makes it easier to explain what substantive commitments about political text come with text scaling.

A practical consequence the unified theory is to offer new ways to interpret and visualise existing text scaling models and also to define useful new ones. In particular, this paper argues for the use of biplots for visualizing the results of text scaling models.

The paper proceeds as follows: I present a first text scaling model defined as a choice-model over words. I then consider an important application, scaling manifesto positions from CMP data (Budge, Robertson, and Hearl 1987), and note that two recent methods, logit scoring (Lowe et al. 2011) and the CMP's own preferred measures, are special cases. I then show how a computationally convenient way to estimate this model recovers the Wordfish (Slapin and Proksch 2008) and the Rhetorical Ideal-point model (Monroe and Maeda 2004). I then re-derive the model directly as an association model (Goodman 1979) and introduce canonical correlation analysis, perhaps better known as simple correspondence analysis, as an efficient least squares approximation. I note in passing that the 'vanilla method' (Gabel and Huber 2000) is in practise another approximation and derive Wordscores (Laver, Benoit, and Garry 2003) as a special case. Finally I show the advantages of the method for inference about dimensionality and position visualisation.

Text Scaling

Text scaling models almost all represent each document in a collection as a 'bag of words' assumption. This means that a set of documents is represented by an $N \times V$ matrix of counts C where entry C_{ij} is the number of times the j-th textual unit (word, category, or sentence) occurs in the i-th text (or document or speech). Denote the row marginal totals as C_{i+}, the column marginal totals as C_{+j}, and the grand total as C_{++}. For concreteness I will where possible refer to N *documents* and a vocabulary of V *words*, although sometimes these 'words' will be categories, topics or some other countable unit. In this terminology C is a cross-tabulation of documents and word types, C_{i+} is the length of document i in words, and C_{+j} is the frequency of word j in the document collection. The samping scheme for C is product multinomial because words are sampled in a stratification determined by document.[2]

The scaling model

Each row of the cross-tabulation $[C_{i1} \ldots C_{iV}]$ represents the content of document i as a vector of word counts. For text scaling we suppose that each row is generated by an actor attempting to express, and thereby provide information about, her unobserved position θ on one or more policy dimensions. Perhaps the simplest generalized linear statistical model that reflects this relationship between which words she chooses to use and

[2] Only Wordscores makes use of this fact, though it need not have bothered since only quantities that are invariant to row and column margins totals are estimated as positions.

her position is a multinomial IRT model, modeled fairly directly on the Binomial model familiar to roll-call vote analysis.

$$\begin{aligned}[C_{i1} \ldots C_{iV}] &\sim \text{Multinomial}(\pi_{i1} \ldots \pi_{iV}, C_{i+}) \\ \log\left(\frac{\pi_{ij}}{\pi_{ik}}\right) &= \psi_{j/k} + \theta_i\, \beta_{j/k} \end{aligned} \tag{1}$$

where the two sets of word parameters ψ and β are labelled to indicate the word contrasts they apply to and the k-th word operates as a baseline.

The same scaling model

Practically, Eq.(1) can be hard to estimate because conditioning on the document length C_{i+} couples the word parameters so they must be jointly estimated. For large V this can be impractical. Fortunately, a completely equivalent 'surrogate Poisson model' formulation of the model decouples them by assuming independent Poisson processes for each cell count and introducing nuisance parameters α_i to capture the effect of conditioning on each C_{i+}:

$$\begin{aligned} C_{ij} &\sim \text{Poisson}(\mu_{ij}) \\ \log\, \mu_{ij} &= \alpha_i + \psi_j + \theta_i\, \beta_j. \end{aligned} \tag{2}$$

The model parameters from (1) and (2) are linearly related:

$$\begin{aligned} \log\left(\frac{\pi_{ij}}{\pi_{ik}}\right) &= \log\left(\frac{\mu_{ij}/\sum_j \mu_{ij}}{\mu_{ik}/\sum_j \mu_{ij}}\right) \\ &= \log\, \mu_{ij} - \log\, \mu_{ik} \\ &= (\alpha_i - \alpha_i) + (\psi_j - \psi_k) + \theta_i\,(\beta_j - \beta_k) \\ &= \qquad\qquad\quad \psi_{j/k} \quad + \theta_i \quad\quad \beta_{j/k} \end{aligned}$$

This relationship between the two formulations is the multinomial-Poisson transform (Palmgren 1981; Baker 1994).

Eq.(2) has the practical advantage over Eq.(1) that it can be simply fitted by iteratively optimising α and θ conditional on the current values of ψ and β, and then ψ and β conditional on the values of α and θ until a peak in the likelihood is reached, as first suggested by Goodman (1979). For large tables, which includes essentially all text scaling applications, this is a far better choice than Fisher-scoring methods.

The multinomial-Poisson transformation also guarantees that either of the two profile likelihoods can be used for conditional inference about parameters from the other margin. Interestingly, this also means that the model is consistent in increasing N, in fixed or random effects formulation due to parameter orthogonality (Lancaster 2002; Charbonneau 2012)).

A second practical advantage of the equivalence is that if we are willing to assume that the word parameters are well-estimated, the formulation in Eq. (1) allows easy computation of asymptotic standard errors for θ without bootstrapping. Implementers can therefore fit using Eq. (2) and construct conditional standard errors using Eq. (1). Lowe and Benoit (2013) discuss this and alternative methods of uncertainty estimation. Conditional standard errors are implemented in the `austin` R package.

Special cases: Wordfish and Rhetorical Ideal-points

Legislative scholars will recognise Eq.(2) as Wordfish (Slapin and Proksch 2008). With the addition of a grand mean parameter it is also the 'Rhetorical ideal-point' model of Monroe and Maeda (2004) in fixed effects formulation.

The scaling model, again

A illuminating alternative way to motivate Eq.(2) as a text scaling model is not as a cheaper way to estimate Eq.(1) but as a log-multiplicative *extension* to the set of hierarchical log-linear models of C.

Of the five possible log-linear models of the two-way table C, two are relevant to word or topic scaling. These are the independence model

$$\log\,\mu_{ij} = \lambda + \lambda^{R}_{i} + \lambda^{C}_{j} \tag{3}$$

and the saturated model

$$\log\,\mu_{ij} = \lambda + \lambda^{R}_{i} + \lambda^{C}_{j} + \lambda^{RC}_{ij} \tag{4}$$

Under the independence model expected counts depend only on the length of each document and the corpus frequency of the words or topics used within it. Clearly no positions can be expressed or inferred from data assumed to be generated this way. In contrast, the saturated model always recovers C perfectly but provides no analysis of C into positional and non-positional variation.

While both models are substantively unsatisfactory, they do locate the *source* of positioning information: the $(N-1)(V-1)$ odds ratios that determine the $N \times V$ matrix of interaction terms λ^{RC}. Substantively, when scaling topic counts these odds ratios reflect facts such as: party A talks about the economy twice as much as social welfare whereas party B talks about it three times as much. This reflects the idea that it is not how much a party talks about the economy that matters but how much more or less it talks about it in proportion to other policy topics. The full set of $(N-1)(V-1)$ such facts exhausts the information available to determine positioning and the assertion that all such facts can be represented adequately in a much lower dimensional space is the assertion that the spatial model is true of political speech in the same way as it is of other behaviour.

To explore the patterns of relative emphasis represented by full set of odds ratios we define a set of models with complexity greater than the independence model but less than the saturated model by systematically decomposing λ^{RC}.

To motivate this approach intuitively, note that every matrix has a spectral decomposition

$$\begin{aligned}\lambda^{RC} &= U\Sigma V^{T}\\ &= \sum_{m=1}^{M} u_{(m)}\sigma_{(m)}v^{T}_{(m)}\end{aligned}$$

into orthogonal left and right singular vectors and singular values. This offers a principled way to define a sequence of lower dimensional spatial models

$$\log\,\mu_{ij} = \lambda + \lambda^{R}_{i} + \lambda^{C}_{j} + \sum_{m=1}^{M} u_{i(m)}\sigma_{(m)}v^{T}_{i(m)} \tag{5}$$

$$\approx \lambda + \lambda^{R}_{i} + \lambda^{C}_{j} + u_{i}\sigma v_{j} \tag{6}$$

where the second line is a rank one approximation to λ^{RC} constructed by taking the largest singular value and its associated vectors. The general form in Eq.(5) is M-dimensional row-column *association model*, or RC(M) model developed by Goodman (1979, 1985).

When $M = 0$ Eq.(5) defines the independence model so the documents have no positions. When $M = \min(N-1, V-1)$ each document (and word) has a position in M-dimensional space and the model is saturated. Intermediate choices of M define models of intermediate complexity. In applications usually hope that M is small and substantively interpretable. Eq.(6) is the lowest dimensional model that gives documents and words positions.

This parameterisation and identification strategy have three advantages for text scaling. First, the symmetrical normalisation for document positions u and word positions v brings out their symmetric role. Unlike a vote scaling IRT model nothing changes except the substantive interpretation if we work with C^T instead of C.

Second, zero values are interpretable: if $v_j \approx 0$ then the j-th word occurs at rate determined by the independence model and so takes no role in determining document positions.

Third, σ provides a measure of how much of the variation in C is due to positioning and how much is to be expected 'by chance' from the baseline independence model. In addition, when $M > 1$ it provides a relative measure of the number of dimensions inherent in C. This is discussed below.

The same scaling model, yet again

Finally, as perhaps still a fourth formulation of the model, Eq.(5) can be interpreted as a multidimensional unfolding model: a proximity formulation is

$$\log \mu_{ij} = \lambda + \lambda_i^R + \lambda_j^C - \sigma(u_i - v_j)^2 \tag{7}$$

where σ is an shared inverse variance. However, without any extra information about u and v the 'proximity' model in Eq.(7) has Eq.(6), an explicitly 'directional' formulation, as its reduced form.[3] (Expand the quadratic terms, reduce, and flip some signs.)

Eq.(7) can be seen as i's unobserved utility for deploying a word or topic with β_j, which depends quadratically on distance to θ_i, in the same way as i must determine whether the status quo β_{nay} or bill's position β_{yea} is closer to θ_i in a roll call. Whereas roll calls consist under the roll call analysis model as a sequence of simple choices between each of which generate a single decision boundary and a logistic regression with coefficient $\beta_{yea/nay}$, a speech consists of V possible decisions made N times, implying the multinomial logistic regression structure of Eq.(1).

Special case: Wordfish and Rhetorical Ideal-points

To recover Wordfish Slapin and Proksch's Wordfish from the RC(M) parameterisation set

$$\alpha \leftarrow \lambda + \lambda^R \quad \psi \leftarrow \lambda^C \quad \beta \leftarrow \sigma v \quad \theta \leftarrow u.$$

For Monroe and Maeda's Rhetorical Ideal-points, set β to σv. The appendix describes some identification issues with these models and their solutions.

[3] It is possible, and perhaps desirable to assert a *non*-quadratic utility structure, but that is not what existing models do.

Special cases: Logit scores and proportional differences

If C is collapsed over left and right categories, or if there are to begin with only two categories in the column set, then the $\hat{\theta}$ of Eq.(11) are just linear transformations of the estimated θ from Eq.(2). This makes Wordfish is the multi-category generalisation of the logit scores described in Lowe et al. 2011. Proofs are straighforward and summarised in (Meyer and Wagner 2014, Appendix).

Approximation: Canonical correlation analysis, correspondence analysis

The intuition embodied in the Poisson formulations above is that positions can only be uncovered after factoring out the counts that would be expected just on the basis of document length and word frequencies. This is realised by a reduced rank model of the interaction terms λ^{RC} embedded in an otherwise log-linear model. However, a computationally cheaper and almost equivalent realisation of the same intuition is to analyse the residuals from the independence model directly.

Define P_{ij} as C_{ij}/C_{++}. The predicted cell probabilities under the independence model are each $P_{i+}P_{+j}$ with Chi-squared residuals

$$R_{ij} = \frac{P_{ij} - P_{i+}P_{+j}}{\sqrt{P_{i+}P_{+j}}}$$

As before, the matrix of residuals has singular value decomposition

$$\begin{aligned} R &= U\Sigma V^T \\ &= \sum_{m=1}^{M} u_{(m)}\sigma_{(m)}v_{(m)}^T \end{aligned}$$

When a smaller number of u and m are retained this is known as canonical correlation analysis (see e.g. Agresti 2002, sec.9.6.3, $\sigma_{(m)}$ is the 'correlation'), and in the case where $M = 2$ and u and v are plotted in the same space, simple correspondence analysis (see M. Greenacre 2007, for a review).

The status of this procedure as an approximation to the association model in Eq.(5) is clearer when it is formulated as an explicit model of the elements of P

$$P_{ij} = P_{i+}\,P_{+j}\,(1 + \sum_{m=1}^{M} u_{i(m)}\,\sigma_{(m)}\,v_{j(m)}^T) \tag{8}$$

$$\approx P_{i+}\,P_{+j}\,(1 + u_i\,\sigma\,v_j) \tag{9}$$

where the second line is a rank one approximation that is optimal in the least-squares sense (Eckart and Young 1936). This is referred to as the 'reconstitution formula' in the correspondence analysis literature.

As the notation suggests, the u and v of Eqs (8) and (9) play the same role as, and are highly correlated with, Eqs.(5) and (6). The approximation will be best when relatively little variation in the data is due to positioning. This is because $e^x \approx 1 + x$ when x is close to zero. Goodman 1985 provides a exhaustive description of the similarities and differences between canonical correlation / correspondence analysis and the RC(M) association model.

In practical applications θ from Eq.(6) and u from Eq.(9) are very highly correlated, as we might expect an ML model and its least squares approximation. Perhaps not surprisingly for a linear approximation, there are values of u and v that reconstruct entries in P that are less than zero or larger than one. This is a function of the geometric roots of correspondence analysis rather than the generative roots of association models.

Substantively, Eq.(9) reflects the same logic of positioning by relative emphasis and should be used as such.

Application: Manifesto positions from coded manifestos

In applications based on CMP data (e.g. Budge, Robertson, and Hearl 1987; Bakker 2009) C_{ij} contains *percentages* of quasi-sentences counts in manifesto i devoted to category j rather than counts. With this important difference to the definition of C, researchers have suggested several ways to estimate party policy positions. The CMP project itself does so by identifying a set L of the column indices corresponding to 'left' categories and another set R for 'right' categories e.g. Budge 2001, table 1, asserting these are equally informative and substantively equivalent, and then collapsing C over R and L to create a new matrix with two columns from which row positions are derived. Different functions contrasting the two columns correspond to different ways to estimate positions.

A straighforward choice is the difference in the proportion of each document spent expressing left versus right policy categories

$$\tilde{\theta}_i \propto \sum_{j \in R} C_{ij} - \sum_{j \in L} C_{ij} \tag{10}$$

When divisor is C_{i+} this is the the preferred CMP estimate of left-right position 'RILE'. Kim and Fording (2002) and Laver and Garry (2000) have suggested that a better denominator might be the sum total of left and right category counts. Lowe et al. (2011) have argued, partly on psychophysical grounds, that sentences categorised on either side of an issue should have decreasing rather than constant marginal effects on a position estimate. This leads to so-called 'logit scores'

$$\hat{\theta}_i = \log \frac{\sum_{j \in R} C_{ij}}{\sum_{j \in L} C_{ij}} \tag{11}$$

Notice that Eqs.(1) and (11) map two modeling limits. Eq.(11) places almost no constraints on possible positions and in this respect resembles a saturated Binomial version of Eq.(1) with as many parameters as positions. The lack of an explicit model has the undesirable consequence that zero counts must be dealt with in an ad-hoc manner, by adding a constant. In contrast, Eq.(1) with scalar θ is nearly the strongest possible model because it forces its linear predictors into a much lower dimensional structure than the original count data. As the *dimension* of θ increases, Eq.(1) provides fewer constraints on the fitted counts until there are as many parameters as positions and the positions themselves are just the empirical logits in (11).

If left and right category counts are fairly balanced then $\tilde{\theta} \approx \hat{\theta}$. Budge and McDonald (2012) confirm that this is the case in applications.

We can make an independent check on whether the choice of left and right categories that go into RILE are indeed substantively equivalent by fitting the scaling model and examining the item parameters β. Fig. 1 shows the results for post war manifestos in Germany. Here all the basic CMP categories and plotted them in order of estimated position in blocks. The figure shows first that while the left and right categories do seem to support these interpretations in Germany, some categories are further left or right than others, so it is not strictly true that they are substantively equivalent, and aggregating them is weakly motivated.

Also, other CMP categories from the middle block could have been included in the left right measure but were not. The scaling model therefore provides some validation of the basic choices but opens the way to a more complete and nuanced scale construction if needed: θ estimates would perhaps be a better choice of scale than either aggregated differences or their log ratio.

Special case: the Vanilla Method

Correlation models compute the SVD of R which is a transformation of C from which expected variation due to differing document lengths *and* word frequencies has been removed. Practically this is done by subtracting expected means and dividing by expected standard deviations under independence. But what happens if SVD is applied to C directly rather than to R as in Gabel and Huber's 'vanilla method'?

In this case the first left and right singular vectors will capture the information provided by the independence model, essentially the relative document lengths and relative word frequencies and the *second and subsequent* left and right singular vectors will have positional interpretations.

The 'vanilla method' for estimating manifesto positions from CMP data is to apply principal component analysis (PCA) and take the first principal component score as the manifesto's position. Although there is a close connection between SVD and PCA – if the SVD of a matrix A is $U\Sigma V^T$ then U are also the eigenvectors of AA^T, V the eigenvectors of A^TA, and σ the square roots of the eigenvalues of AA^T and A^TA – we would not necessarily expect the first principal component of the data matrix to be close to the u of Eq.(9), or even to be particularly interpretable.

However, the CMP represents each manifesto as a vector of policy area percentages rather than counts.[4] Because these should, and mostly do sum to one hundred in each row, document length information is *already* effectively removed from the data. Moreover, standard routines for PCA suggest standardising by column to remove the effects of having variables (columns) on different scales.[5] After these row and column normalisations, the manifesto positions constructed by the vanilla method will be therefore be extremely close to those constructed using Eq.(9) or indeed Eq.(6). One substantive implication of this is that researchers should probably not use the 'vanilla method' on general C matrices, but only on the CMP's prefered format.

Implementations

Practical implementations of Eq.(8) require a method for computing the singular value decomposition. Efficient matrix methods are now widely available, but in the past a reciprocal averaging scheme was used. This consists of alternating

$$u_i^{(t)} \leftarrow \sum_j C_{ij} v_j^{(t-1)} / C_{+j} \quad v_j^{(t)} \leftarrow \sum_i C_{ij} u_i^{(t-1)} / C_{i+} \tag{12}$$

As t increases $u^{(t)} \rightarrow u$ and $v^{(t)} \rightarrow v$, from any non-identical starting values Hill 1974, prop.1. At the end of each iteration a normalisation of u, e.g., Eq.(13), must be applied to prevent u and v collapsing to vectors of ones.

[4] Original counts can be recovered by multiplication. Mostly. This would work better if the project believed in distributing policy area counts in even single precision.

[5] For example R's `prcomp` function, which describes its argument `scale` as follows: "a logical value indicating whether the variables should be scaled to have unit variance before the analysis takes place. The default is 'FALSE' for consistency with S, but *in general scaling is advisable*." (emphasis added). Alternatively PCA operates on the correlation matrix of the data which clearly performs both row and column normalisation.

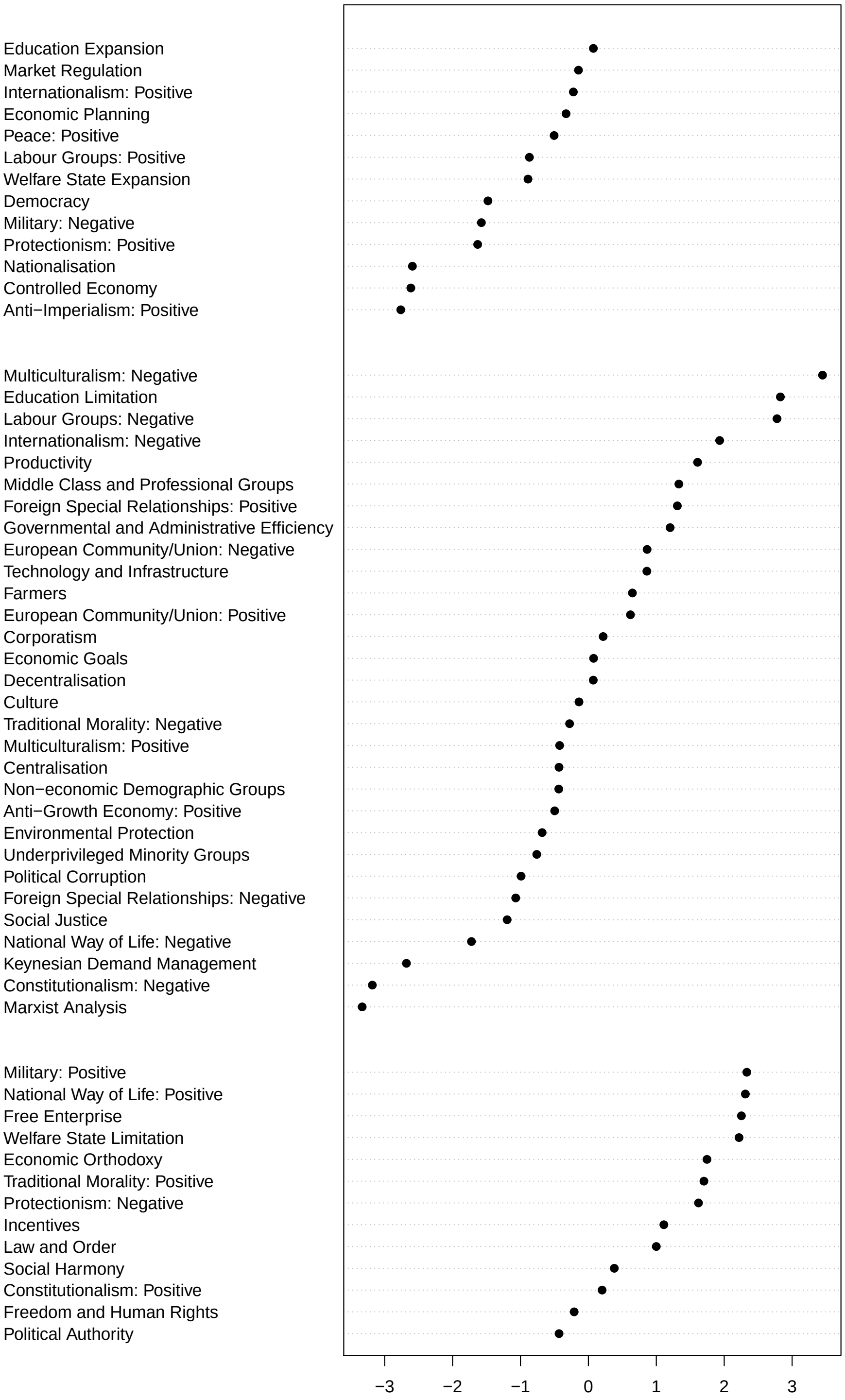

Figure 1: CMP category scores for manifestos from post-war Germany in groups of by left RILE (top), unused (middle), and right RILE (bottom).

Special case: Wordscores

When all document positions u are considered to be known, only the word positions v need to be estimated. Wordscores applies just one application of the right hand side of Eq.(12) and constructs positions for out-of-sample documents using the left hand side. This defines the Wordscores algorithm[6] (Laver, Benoit, and Garry 2003). A more detailed description of this connection with reference to Eq.(9) in the form of correspondence analysis is given in Lowe (2008).

Wordscores appears not to require the imposition of any normalisation step such as (13) because the assumption of known u implies that no iterations are necessary. However, a related issue does arise when predicting the new document positions(Benoit and Laver 2008; Martin and Vanberg 2007). This identification issue cannot be avoided.

The connection between canonical correlation and Wordscores has one more substantive implication: The linearity of (9)'s approximation to (6) implies that if the true data are generated by Eq.(5) then the word and document position parameters in Eq.(8) will be shrunk towards the centre at the edges. This is related to the endpoint shrinkage discussed by Lowe (2008) and in more detail by ter Braak and Looman (1986).

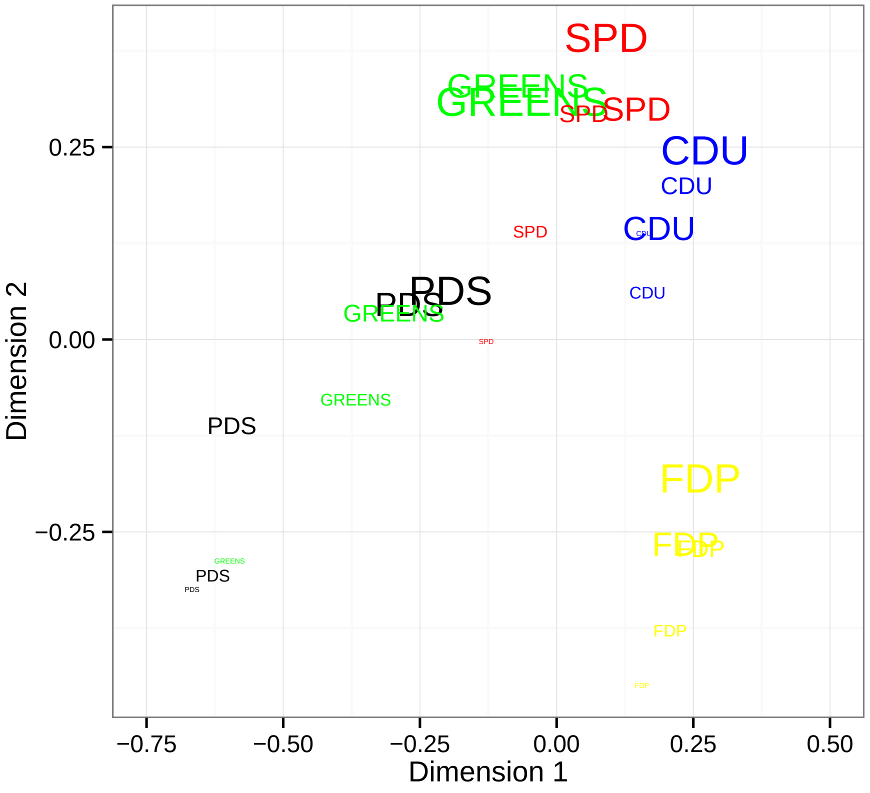

Figure 2: Party positions in German federal elections of 1990, 1994, 1998, 2004, and 2008 from a two-dimensional canonical correlation model. Positions are labelled with party initials. Labels are sized from the 1990 (smallest) to 2008 (largest).

[6] Strictly, the left recursion in Wordscores is performed on a column-normalised version of C but dividing each column by a constant makes no difference, except through rounding error, to the final u and v.

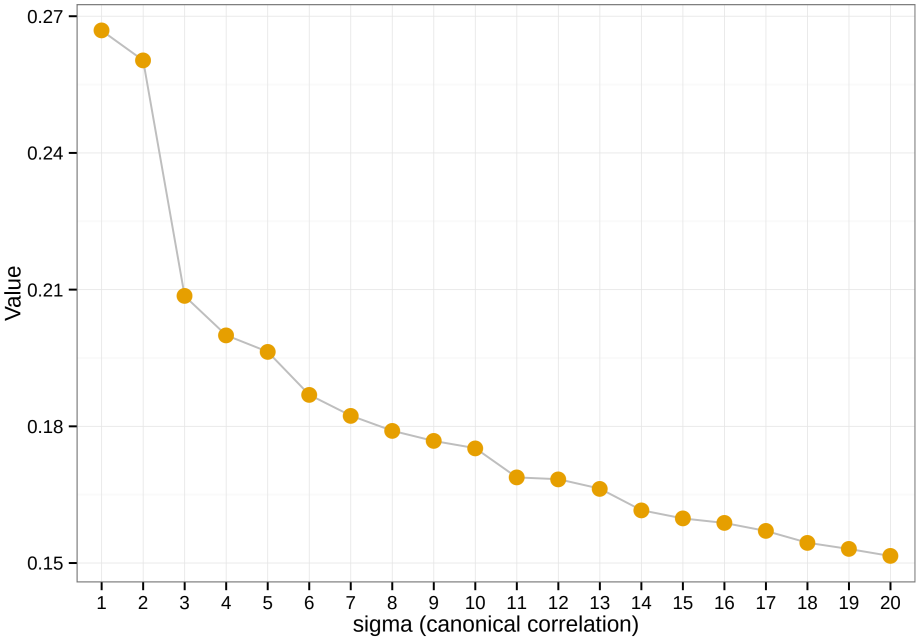

Figure 3: Canonical correlation parameters σ for the first 20 dimensions of the german economic manifesto data.

Practical Advantages

In addition to providing a unifying conceptual scheme for text scaling models, and perhaps stemming the tide of new formulations of the same (good) intuition, there are practical advantages to working with association models or their canonical correlation approximations for text scaling.

Easily fitted multidimensional models

A considerable practical advantage of the canonical correlation is the ease and speed of fitting models with multiple dimensions, to which models like Wordfish have not yet been extended.

For example, Fig. 2 shows positions on the first two dimensions for each the German manifesto data used in the original Wordfish paper (The combined economic sections of five main parties in five German elections Slapin and Proksch 2008). The positions have been coloured and labelled by party and larger labels indicate more recent elections. Fig. 3 shows the estimated values of σ for the 20 largest dimensions and suggests that a model where $M = 2$ might be reasonable.

Fig. 3 suggests that there are two main dimensions of variation in this data. So we can now ask: what substantive meaning, if any, do the dimensions in Fig. 2 have? Knowledge of recent German politics suggests that there are indeed two interpretable dimensions embedded in the positions, but that these are not (quite) aligned with the axes. This will in general be true – there is no reason to expect substantive dimensions to be orthogonal (Albright 2010).

If the positions are projected onto a line extending from middle left to bottom right then the parties order on a free market economics dimension from PDS on the left to FDP on the right. At 90 degrees to this the parties order by time, with earlier elections towards the bottom left and later ones towards to the top right of the figure. A one dimensional model would have recovered only the first dimension which would mostly remove the distinctions in economic policy between the centre-right CDU and the explicitly free market-oriented FDP.

An interesting consequence of estimating the second dimension is the movement of some but not all party positions towards the centre of the substantive first dimension and the large changes over time of the left wing parties' ways of talking about economics, or more likely the choice of economic topics that their manifestos covered over this set of elections.

Better visualisation

Fig. 4 shows a two dimensional canonical correlation model of CMP category counts for the major post-war German party manifestos (inflated to suitable C matrices). This plot is a *biplot* (M. J. Greenacre 2010) and its substantive interpretation is as follows:

Parties appearing close together hae similar profiles of topic emphasis and topics close together are emphasised by similar sets of parties. Points appearing at the origin use each column category at rates that would be expected by chance, that is, according to the independence model. That is, a party appearing at the origin uses each policy topic exactly as much as would be expected from the frequency of references to these topics in manifestos of the period, and a topic at the origin is one used by each party equally often. Deviations from the origin represent the *amount* more relative emphasis each manifesto puts on each category.

Finally, the angle between the a vector from the origin to a party manifesto and a vector from the origin to a policy topic indicates how much a party emphasised that topic over others. Thus a topic such as 'Freedom and Human Rights' (lower left quadrant) is emphasised almost equally as strongly by the FDP (lower right quadrant) and the Greens and left parties (left side). 'Education Expansion' has a similar party profile but its emphasis is weaker. In contrast, 'Welfare State Limitation' is almost exclusively an FDP policy topic.

The biplot also nicely shows the two non-orthogonal dimensions of policy dispute in the party system: a liberal-conservative ordering on social issues from left to top right, and a redistributive-neoliberal dimension from left to bottom right.

The biplot shows only the policy topics recommended for a left right dimension by the CMP, but e can also scale and project all measured policy topics in the same space. Figure 5 shows all topics.

The saliency theory of positioning

Budge (2001) lays out the basic assumptions of the saliency theory of party competition. The two most relevant for this paper are that

1. "policy differences between parties thus consist of contrasting emphases on different policy areas (thus, one party often mentions taxes, another benefits)".
2. "party strategists see electors as overwhelmingly favouring one course of action on most issues. Hence all programmes endorse that position, with only minor exceptions"
3. In certain special cases "sets of policy emphases which go together can be added numerically and contrasted with sets of opposing emphases to form a unified directional index such as Right versus Left"

(Budge 2001)

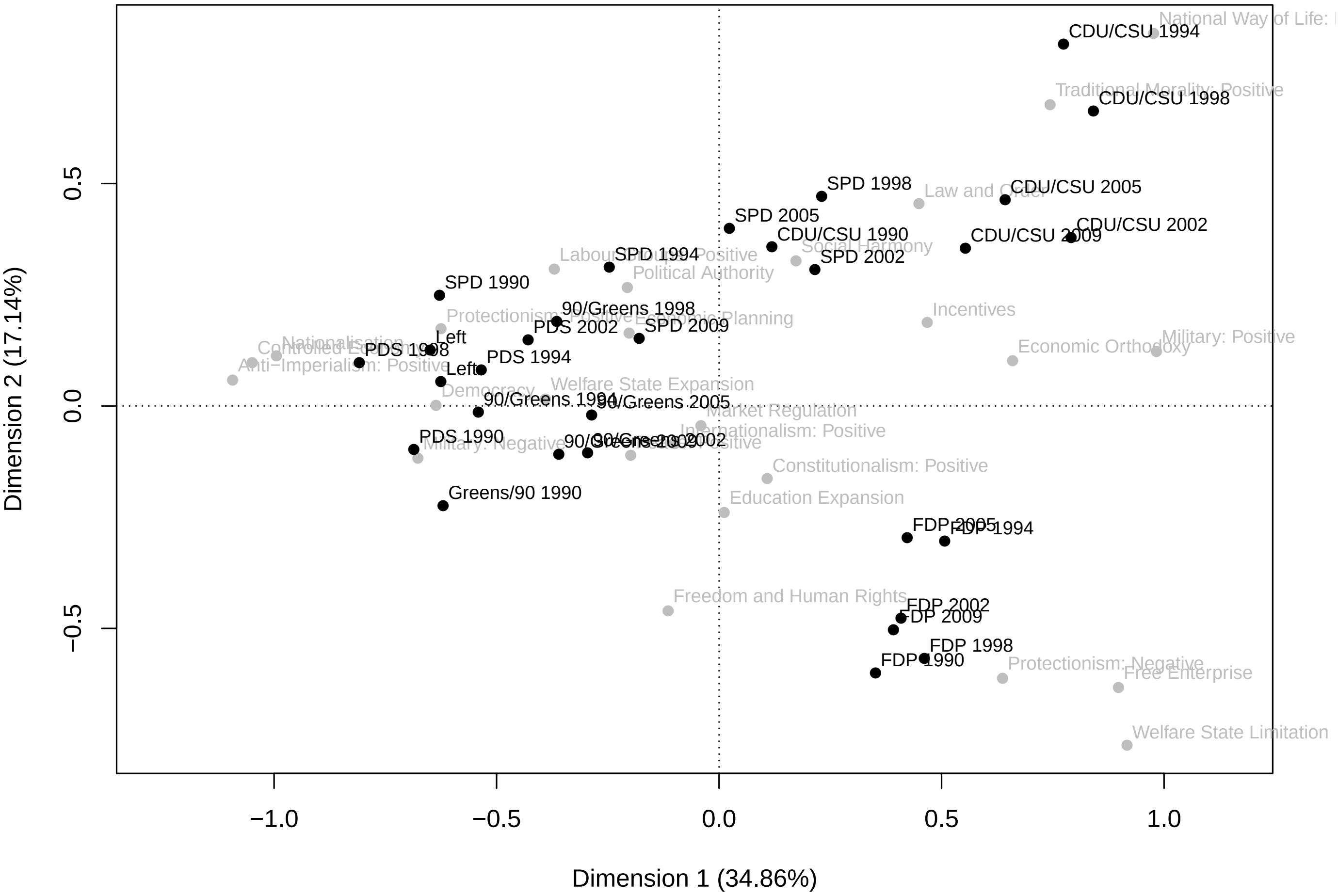

Figure 4: Biplot of post war German manifestos and their use of left-right CMP categories.

The assumptions of the models discussed above reflect exactly the first assumption. Relative emphasis on a category in a manifesto means, in the CMP quasi-sentence operationalisation, talking more about relative to other categories.

The models discussed here operationalise this idea more precisely because they require a speaker or writer to use more words or sentences on one topic than another relative to what would be expected by chance according to the independence model. Thus the actual *number* of words or quasi-sentences needed to do this differs according to the chance baseline, the prevalence of the other categories, and the length of the document. Nevertheless, the intuition is the same as 1) above.

In particular cases, some categories are naturally contrasted (or bundled) and are good candidates for an index measure. When the bundles of categories are known then logit or linear indexes are a natural option. When bundle contents are unknown they can be estimated as in Fig 1.

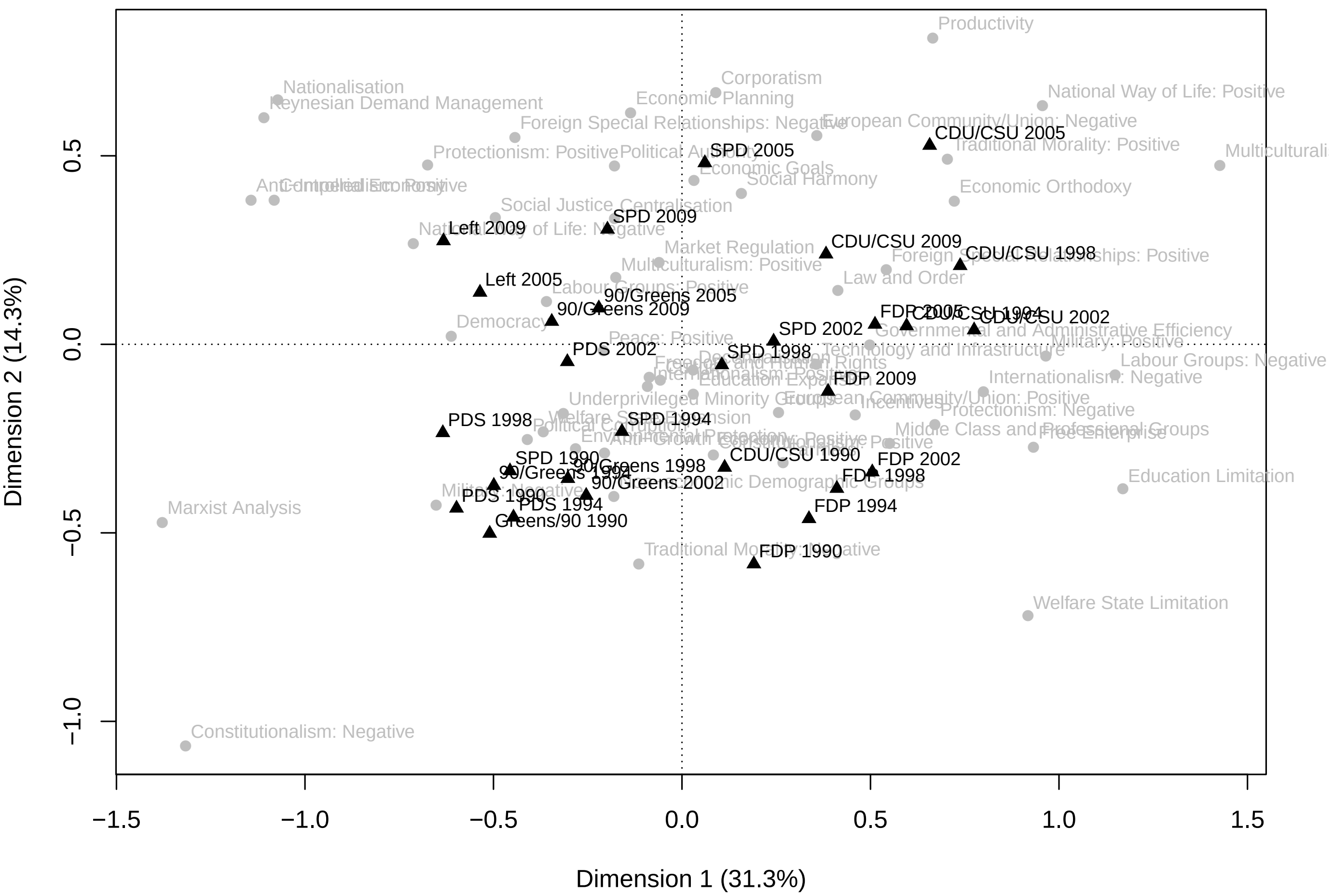

Figure 5: Biplot of post war German manifestos and their use of all CMP categories.

The models above also provide a counter to the claim that all programmes have the same position. Here, position just *is* the low dimensional summary of relative emphases over the appropriate unit so it is not possible to differ in relative emphasis while maintaining the same position. To retain the intuition in 2) it might be reasonable to say that electors all see the policy advantages of e.g. reducing state intervention in the economy and also the advantages of increasing the state's role, but weigh the benefits of these differently.

Finally, the model above not only locates positioning information, but also more traditional salience information. In the association model fitted to category data, e.g., fitted to CMP category counts from a single country, λ_L offers a view of the relative amounts of emphasis spent on each policy topic. This might be a reasonable operationalisation of policy salience. An interesting implication of the model is that salience is nearly, but not quite independent of position, because increasing salience by raising the base rate of discussion about a policy area also changes the exact amount of emphasis, in sentence or word units, necessary for an actor to maintain the same positions on potentially all issues.

To be clear, this discussion of saliency theory is not intended to criticise (or support) it as a substantive theory of party competition, but only to point out that two of the three main assumptions of saliency theory are in fact realised by the models discussed in this paper. This makes CMP objections to text scaling somewhat ironic.

There are always positions

The logic of positioning by relative emphasis implies that anything we can count can always be scaled; the worst that can happen is that no low-dimensional projection of counts will clearly account for a large amount of the observed variation. So the question for political scientists is not: 'Are there positions?' but rather: 'Are these the patterns of relative emphasis we are looking for?'

Conclusion

This paper has presented a unified theory of scaling count data and showed many existing methods to be special cases or approximation of it. The theory rests on the idea that a logic of relative emphasis can be used to model and visualise contingency table data of all kinds. It also shows that models based on this theory recover and allow us to extend previous manual coding-based approaches. While more sophisticated probabilistic modeing will continue to be developed for particular political science applications we should be able to agree on the basic logic of positioning and concentrate on harnessing it to address substantive questions.

A Identification

The focus of this paper is on the logic and models of scaling rather than particular estimation strategies, but it is important to note some practical considerations concerning identification. When θ is modeled as a fixed but unobserved quantity it must be constrained for the model to be identified. For one-dimensional θ a common approach is to require that

$$\sum \theta_i = 0 \qquad \sum \theta_i^2 = 1 \qquad \theta_a > \theta_b \tag{13}$$

where a and b are any two row indices and the third constraint sets the left-right interpretation. For comparisons between the association model and correspondence analysis formulations of the models it is sometimes convenient to make these weighted averages with weights given by the margins of C. For identification it is ony necessary there be two linear constraints.

Multidimensional θ will also require constraints on covariance between positions on different dimensions. Setting them all to zero is a convenient choice.

If instead θ is instead considered as a sample from a population of possible policy positions $\theta \sim \text{Normal}(0, 1)$ is also sufficient to identify the model, though any sufficiently informative prior will work. The normal prior defines a latent trait model (Moustaki and Knott 2000) or equivalently, makes the connection to multinomial IRT models (Clinton, Jackman, and Rivers 2004).

It is also possible to define models that condition θ on external information e.g. about the speaker, which can also serve to identify positions.

Regrdless of the strategy for identifying θ the other model parameters also need identifying. Interestingly, the Wordfish model is not likelihood identified. Any change in $\bar{\beta} = \sum_j \beta_j / V$ can be compensated for by changes in α without changing the probability of the data under the model. Specifically, let m and s be the average and standard deviation of β. The translations between the two models are

Wordfish to Association Model	Association Model to Wordfish
$u \leftarrow \theta$	$\theta \leftarrow u$
$v \leftarrow (\beta - m)/s$	$\beta \leftarrow v\sigma + m$
$\sigma \leftarrow s$	
$r \leftarrow \alpha + \theta m$	$a \leftarrow \lambda + \lambda^R - \theta m$
$\lambda^R \leftarrow r - \bar{r}$	$\alpha \leftarrow a - a_1$
$\lambda^C \leftarrow \lambda^C - \bar{\lambda}^C$	$\psi \leftarrow \lambda^C + a_1$
$\lambda \leftarrow \bar{r} + \bar{\lambda}^C$	

where a bar indicates arithmetic average. In the second column, m may be chosen arbitrarily; the association model fixes it to be zero.

Despite this lack of identification, the Wordfish model as a whole *is* identified because of the extra constraint provided by a ridge prior on β. Contrary to the original paper's claim, this does not fix a 'technical issue' by adding a regularisation term but rather completes the identification[7].

[7] so don't leave it out…Alternating conditional maximization is not actually an EM algorithm either, despite the paper's claims,

B Scaling in the space of text analysis models

Why should a model apply to counts at all levels? It is easier to see why this should be true from a measurement perspective. Starting with at the lowest level, if words are modeled as conditionally independent given an latent variable, e.g. a position, then we can scale words to recover positions directly.

If, as is more plausible, positions are patterns of relative emphasis over policy topics rather than individul word types, then scaling topic counts rather than word counts will recover positions.

But if, as topic modelers assume, topics are themselves distributions of word types conditioned on unobserved topic indicators, then scaling words directly will recover noisy versions of the topic-defined positions without our needing to know or inferring the topic indicators, but with extra uncertainty due to not making use of the subspace of word counts induced by convex combinations of underlying topics.

It is sometimes helpful to think of K topics as inducing a $K - 1$-dimensional subspace in the N-dimensional space of word counts, and of scaling models as pursuing an $M \ll K$ dimensional summary of relative emphasis across K topics. In this generative scheme knowing intermediate topic structure decreases variance in position estimation and increases bias and substantive interpretability.

but that detail does not compromise our understanding of the form and substantive implications of the model.

References

Agresti, Alan. 2002. *Categorical data analysis.* 2nd ed. New York: Wiley-Interscience.

Albright, Jeremy J. 2008. *Bayesian estimates of party left-right scores.* Working Paper, Society for Political Methodology 801. July 11.

———. 2010. "The multidimensional nature of party competition." *Party Politics* 16, no. 6 (November 1): 699–719. Accessed June 18, 2015.

Baker, Stuart G. 1994. "The multinomial-Poisson transformation." *Journal of the Royal Statistical Society. Series D (The Statistician)* 43 (4): 495–504. JSTOR: 2348134.

Bakker, Ryan. 2009. "Re-measuring left–right: A comparison of SEM and Bayesian measurement models for extracting left–right party placements." *Electoral Studies* 28, no. 3 (September): 413–421.

Benoit, Kenneth R., and Michael Laver. 2008. "Compared to what? A comment on 'A robust transformation procedure for interpreting political text' by Martin and Vanberg." *Political Analysis* 16 (1): 101–111.

Budge, Ian. 2001. "Validating party policy placements." *British Journal of Political Science* 31 (1): 210–223. JSTOR: 10.2307/3593282.

Budge, Ian, and Michael D McDonald. 2012. "Conceptualising and measuring 'centrism' correctly on the Left–Right scale (RILE) - without systematic bias. A general response by MARPOR." *Electoral Studies* 31 (3): 609–612.

Budge, Ian, David Robertson, and Derek Hearl, eds. 1987. *Ideology, strategy and party change: Spatial analyses of post-war election programmes in 19 democracies.* Cambridge UK: Cambridge University Press.

Charbonneau, Karyne. 2012. "Multiple fixed effects in nonlinear panel data models." Accessed April 7, 2016.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. "The statistical analysis of roll call data." *American Political Science Review* 98, no. 02 (June): 1–16.

Dinas, Elias, and Kostas Gemenis. 2009. "Measuring parties' ideological positions with manifesto data: A critical evaluation of the competing methods": 1–25.

Eckart, Carl, and Gale Young. 1936. "The approximation of one matrix by another of lower rank." *Psychometrika* 1, no. 3 (September): 211–218. Accessed June 18, 2015.

Elff, Martin. 2013. "A dynamic state-space model of coded political texts." *Political Analysis* 21 (2): 217–232.

Gabel, Matthew J., and John D. Huber. 2000. "Putting parties in their place: Inferring party left-right ideological positions from party manifestos data." *American Journal of Political Science* 44, no. 1 (January): 94. JSTOR: 2669295?origin=crossref.

Goodman, Leo A. 1979. "Simple models for the analysis of association in cross-classifications having ordered categories." *Journal of the American Statistical Association* 74 (367): 537–552.

———. 1985. "The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries." *The Annals of Statistics* 13 (1): 10–69. JSTOR: 10.2307/2241143.

Greenacre, Michael. 2007. *Correspondence analysis in practice.* Chapman and Hall/CRC.

Greenacre, Michael J. 2010. "Correspondence analysis." *WIREs Computation Statistics.*

Hill, Mark O. 1974. “Correspondence analysis: A neglected multivariate method.” *Applied Statistics* 23 (3): 340–354.

Kim, Heemin, and Richard C. Fording. 2002. “Government partisanship in Western democracies, 1945-1998.” *European Journal of Political Research* 41 (2): 187–206.

Klemmensen, Robert, Sara Binzer Hobolt, and Martin Ejnar Hansen. 2007. “Estimating policy positions using political texts: An evaluation of the Wordscores approach.” *Electoral Studies* 26, no. 4 (December): 746–755.

Klüver, Heike. 2009. “Measuring interest group influence using quantitative text analysis.” *European Union Politics* 10 (4): 535–549.

König, Thomas, and Bernd Luig. 2009. “German ’LexIconSpace’: policy positions and their legislative context.” *German Politics* 18, no. 3 (September): 345–364.

König, Thomas, Moritz Marbach, and Moritz Osnabrugge. 2013. “Estimating party positions across countries and time–a dynamic latent variable model for manifesto data.” *Political Analysis* 21, no. 4 (October 1): 468–491. Accessed June 18, 2015.

Lancaster, Tony. 2002. “Orthogonal parameters and panel data.” *The Review of Economic Studies* 69 (3): 647–666.

Laver, Michael, Kenneth R. Benoit, and John Garry. 2003. “Extracting policy positions from political texts using words as data.” *American Political Science Review* 97, no. 2 (May): 311–331.

Laver, Michael, and John Garry. 2000. “Estimating policy positions from political texts.” *American Journal of Political Science* 44 (3): 619–634.

Lowe, Will. 2008. “Understanding Wordscores.” *Political Analysis* 16 (4): 356–371.

Lowe, Will, and Kenneth R. Benoit. 2013. “Validating estimates of latent traits from textual data using human judgment as a benchmark.” *Political Analysis* 21 (3): 298–313.

Lowe, Will, Kenneth R. Benoit, Slava Mikhaylov, and Michael Laver. 2011. “Scaling policy preferences from coded political texts.” *Legislative Studies Quarterly* 36 (1): 123–155.

Martin, L W, and G Vanberg. 2007. “A robust transformation procedure for interpreting political text.” *Political Analysis* 16 (1): 93–100.

Meyer, Thomas M., and Markus Wagner. 2014. “How parties and candidates move: Shifting emphasis or positions?” October.

Monroe, Burt L., and Ko Maeda. 2004. “Talk’s cheap: Text-based estimation of rhetorical ideal-points.”

Moustaki, Irini, and Martin Knott. 2000. “Generalized latent trait models.” *Psychometrika* 65, no. 3 (September): 391–411.

Palmgren, Juni. 1981. “The Fisher information matrix for log linear models arguing conditionally on observed explanatory variables.” *Biometrika* 68, no. 2 (August): 563–566.

Pennings, Paul, and Hans Keman. 2002. “Towards a new methodology of estimating party policy positions.” *Quality and Quantity* 36 (1): 55–79.

Slapin, Jonathan B., and Sven-Oliver Proksch. 2008. “A scaling model for estimating time-series party positions from texts.” *American Journal of Political Science* 52, no. 3 (July): 705–722.

Ter Braak, Cajo J. F., and Caspar W. N. Looman. 1986. “Weighted averaging, logistic regression and the Gaussian response model.” *Plant Ecology* 65, no. 1 (January): 3–11.

Warwick, Paul V. 2002. “Toward a common dimensionality in West European policy spaces.” *Party Politics* 8, no. 1 (January): 101–122.