Estimating Ideal Points from Votes and Text *

In Song Kim†  John Londregan‡  Marc Ratkovic§

Forthcoming in Political Analysis

Abstract

We introduce a framework for combining vote data and text data within a single formal and statistical framework. Formally, we model vote choice and word choice in terms of a common set of underlying preference parameters. Statistically, we implement a method for recovering these preference and location parameters. The method estimates the number of underlying ideological dimensions, models zero inflation, and is robust to extreme outliers. We apply the method to rollcall and floor speech from recent US Senates. We find two stable dimensions, one ideological and the other capturing leadership. We then show how the method can leverage common speech in order to impute missing data, to estimate rank-and-file ideal points using only their words and the vote history of party leaders, and even to scale newspaper editorials.

---

*We thank Ben Johnson, Tolya Levin, and Michael Peress for comments on this and an earlier draft.
†Assistant Professor, Department of Political Science, Massachusetts Institute of Technology, Cambridge, MA, 02139. Email: insong@mit.EDU URL: http://web.mit.edu/insong/www/
‡Professor of Politics and International Affairs, Woodrow Wilson School, Princeton University, Princeton NJ 08544. Phone: 609-258-4854, Email: jbl@princeton.edu
§Assistant Professor, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 608-658-9665, Email: ratkovic@princeton.edu URL: http://www.princeton.edu/~ratkovic
1 Introduction

Social scientists have grown quite adept at recovering latent preference estimates from rollcall data (e.g., Clinton, Jackman, and Rivers [2004] Poole and Rosenthal [1997]). A more recent literature has worked to extract ideological location from text (e.g., Slapin and Proksch [2008] Laver, Benoit, and Garry [2003]). Several recent studies have scaled binary choice data, ignoring readily available text data (Barbera [2016] Ho and Quinn [2008]). Others have modeled text, ignoring readily available vote data (Lo, Proksch, and Slapin [2014] Elff [2013] Quinn et al. [2010]). A third set has modeled preferences conditional on observed text (Lauderdale and Clark [2014]). In these works, ideal points recovered only after textual topics have been estimated. Words do not enter into the political actors’ choice set.

Political actors, though, decide how to vote and they choose the words that they use. We place this process within a unified and practical framework. By unified, we mean that we model both votes and words as arising from a single preference structure. We lay out the theory and underlying assumptions necessary, so that the researcher may assess their validity in a given context. Our model inherit both the strengths and shortcomings of the standard spatial model (e.g. Ladha [1991] Clinton, Jackman, and Rivers [2004]), both of which we explicate below.

By a practical framework, we mean that the structural parameters of the formal model are identified and estimable from data. We offer two statistical advances. First, we estimate rather than assume the number of latent dimensions. We put a sparsity prior over the dimension weights, shrinking most dimensions to zero. Second, we provide a robust and flexible model of word counts. Unlike existing methods, we pay particular attention to zero-inflation, a pronounced attribute of text data. The model is also robust to extreme word counts, another feature common to text data.
The method is implemented through our proposed statistical method, Sparse Factor Analysis (SFA). SFA offers three advances for the applied user. First, SFA eases interpretation through scaling both actors and their words. As we show, dimensions can be interpreted off the estimated word locations. Second, we compare SFA to the increasingly popular, and quite flexible, topic model. We attempt to help the researcher decide which framework may be useful for a given scenario. Third, we make publicly available software for implementing the method in the R package BLINDED.

For illustration, we apply SFA to data from the last several US Senates. After assessing the validity of SFA’s assumptions in this case, we fit the model. We recover two stable dimensions, one running left-right and the second from party leaders to heterodox members. We then assess the method’s internal validity through an in-sample predictive exercise. Lastly, we assess the methods’ external validity through scaling newspaper editorials.

The paper progresses in three parts. First, we discuss the formal model and statistical implementation. Second, we apply the method to recent US Congressional data. A conclusion follows.

2 SFA: The Proposed Method

Voting and speaking are two of the most studied and illuminating political acts. Often, the same actors will do both. Yet, it is common for exemplary papers to take advantage of one or the other types of data, but not both. These are works that either scale a binary choice and leave readily available text alone or model text, but do not connect the text to readily available voting data. As an example of the former, Barbera (2016) scales the choice to follow Twitter users but does not include the content of tweets in the scaling. Similarly, Ho and Quinn (2008) scales newspaper editorials that register opinions on Supreme Court cases, but the study does not include the content of the editorials. As an example of the latter, Lo, Proksch, and Slapin (2014) scales the text of election
manifestos through a variant of the Wordfish model, while not including any vote data. Recent
works that combine both words and votes treat them as coming from separate data generating
processes, rather than a single preference structure (Lauderdale and Clark, 2014).

In this section, we develop a choice theoretic spatial model that establishes a basic homology
between voting and speech: both are anchored to the very same legislator ideal points. Votes
and words are jointly generated from the same ideal points in the same ideological space. Next, we
introduce Sparse Factor Analysis (SFA) for estimating these ideal points using vote data, word data,
or a combination of the two. The subsequent section discusses some of the key ideas embedded in
SFA.

2.1 The Model

We discuss the formal model in this section and estimation strategy in the next.

**Observed data.** We observe for each member \( l \), \( l \in \{1, 2, \ldots, L\} \) both a vote \( V_{lp} \in \{0, 1\} \) on
proposal \( p \in \{1, 2, \ldots, P\} \) and a count for the number of times she utters term \( w \), \( T_{lw} \in \{0, 1, 2, \ldots\} \).
The observed vote corresponds with each actor registering a Yay (1) or Nay (0) vote. The observed
term count, later operationalized as stemmed bigrams, is how many times the actor uses a word or
set of words.

We associate with each observed vote, \( V_{lp} \), a latent variable \( V_{lp}^* \) such that larger values of the
latent variable means the member is more likely to favor the proposal. We also associate with each
observed term count \( T_{lw} \), a latent variable \( T_{lw}^* \), such that larger values of this variable indicate a
member’s proclivity to use this term with a higher frequency. These latent variables are on the
same scale, but they will be mapped back to observed votes or word counts.

We assume that preferences exist in a \( D \)-dimensional space. In each dimension, member \( l \) has
a preferred outcome $x_{id}$, $d \in \{1, 2, \ldots, D\}$. Each dimension has a weight $a_d \geq 0$, common across members, with a higher value of $a_d$ signifying a more relevant dimension. Dimensions with a weight of 0 are irrelevant.

**Modeling assumptions.** Before considering any statistical method, the researcher should check that the assumptions of the model seem plausible in the study at hand. SFA is formulated within the standard quadratic-loss spatial random utility model; see Ladha (1991); Clinton, Jackman, and Rivers (2004) for seminal work.\(^1\)

Behaviorally, SFA assumes actors are sincere in several senses. First, actors must not be deceiving others with their votes or speech. Both should be accurate reflections of underlying preferences. Second, the actors must not be strategic. This means that actors are not voting in order to change the votes of others, and an actor’s vote will not change because of the behavior of others. Similarly, an actor’s speech must accurately reflect her beliefs, and the speech must be expressive rather than deliberative or persuasive.

SFA also requires two underlying structural assumptions. First, like the standard vote model, the SFA vote model requires an exogenous agenda setter. The term model requires that the relevance of different terms be set exogenously. In both cases, the model assumes that all actors are “agenda-takers” rather than “agenda-makers.” Secondly, also like the standard vote model, the SFA vote model assumes that all proposals have exogenous and status quo and alternative positions in the ideological space. Analogously, the term model requires that words have an exogenous and fixed position in the ideological space. The term, status quo, and alternative positions may be perceived with error, though we assume that actors select to maximize their utility given the perceived locations (Ladha, 1991 esp. Sec 2). Basically, words have to mean something, everyone has to

\(^1\)Note that this framework differs from that of Poole and Rosenthal (1997), who assume a Gaussian utility.
approximately agree on what they mean, and that meaning cannot change over the period under study.

Even when these assumptions hold, SFA may estimate a scale attributable to non-ideological preference. The estimated locations are no less a reflection of preference for being over non-positional issues. For example, in our analysis below, our second dimension is anchored on one end by party leaders. These leaders have preferences over certain words in a manner largely uncorrelated with their vote choices. Their most preferred outcomes, then, are the words and votes that are closest to their spatial location in each dimension.

Of course, these assumptions will not hold exactly outside of controlled or contrived situations. SFA, or any scaling method for that matter, should only be applied in situations where the assumptions seem reasonable given substantive knowledge. For example, we apply the method to US Senate floor speeches. We rely on qualitative studies to argue that floor speech is primarily expressive, meeting the assumptions above. We would be less comfortable applying the method to committee hearings, and even less so to judicial argument. We also discuss and implement several methods for assessing internal and external validity. We strongly recommend each of these checks when using SFA.

**Vote Choice** In making a vote choice, the member compares the distance between her ideal point and the spatial consequence of voting either for or against the proposal. The proposal and status quo are assumed to have a location in the same space as the ideal points, with locations denoted \( r_{ld} \) and \( q_{ld} \), respectively. We assume a quadratic utility loss, implying that members prefer the proposal
to the status quo as

\[ U_{l}^{\text{vote}}(\text{Aye}; \{x_{id}\}_{d=1}^{D}, \{z_{pd}^{\text{aye}}\}_{d=1}^{D}) - U_{l}^{\text{vote}}(\text{Nay}; \{x_{id}\}_{d=1}^{D}, \{z_{pd}^{\text{nay}}\}_{d=1}^{D}) \]  

(1)

\[ = -\frac{1}{2} \sum_{d=1}^{D} a_d (z_{pd}^{\text{aye}} - x_{id})^2 - \left( -\frac{1}{2} \sum_{d=1}^{D} a_d (z_{pd}^{\text{nay}} - x_{id})^2 \right) - \epsilon_{lp}^{\text{vote}} \]  

(2)

with \( \epsilon_{lp} \) a standard normal random variable. Simplifying and combining terms gives a representation of the vote choice as\(^2\)

\[ V_{lp}^* = c_{l}^{\text{vote}} + b_{p}^{\text{vote}} + \sum_{d=1}^{D} a_d x_{id} g_{pd}^{\text{vote}} - \epsilon_{lp}^{\text{vote}} \]  

(3)

where \( c_{l}^{\text{vote}} \) and \( b_{p}^{\text{vote}} \) are individual- and proposal-specific effects, \( a_d \) and \( x_{id} \) are the dimension weights and ideal points described above, and \( g_{pd}^{\text{vote}} \) is the signed distance between the Yay and Nay alternatives. The terms \( c_{l}^{\text{vote}} \) and \( b_{p}^{\text{vote}} \) are amalgams of structural parameters of interest and are rarely of direct interest themselves; they serve to model the baseline propensity for a given proposal to receive support and a given member to support any proposal.

**Term choice.** We assume a similar model for the term choice. We take as the choice variable the a latent intensity variable that maps to an observed count for each term, \( T_{lw}^* \). The member selects \( T_{lw}^* \) with two considerations: the proximity of the term to her ideal point and the pertinence of the term to the issues of the day. Ideological proximity is the distance from her ideal point to the term’s spatial location. If the member only selected terms based off ideology, then she would simply utter her most preferred terms \textit{ad infinitum}, regardless of external circumstance. As members do not choose terms this way, we model a second, countervailing force: \textit{pertinence}. Pertinence captures the notion that some terms are more appropriate in some years than others; for example, we find discussion of mortgage backed securities in 2009 that were not relevant in 1999. Pertinence is a function of a legislator’s baseline verbosity (\( v_l \)), the aptness of the word to the debate at hand (\( s_w \)),

\(^2\)For a specification of the utility functions and a full derivation, see the technical appendix.
and the diminishing return from overusing a term. Formally, the choice is taken from maximizing

$$U_{lw}^{term} \left( T_{lw}^*, \{x_{ld}\}_{d=1}^D, \{z_{wd}^{term}\}_{d=1}^D \right) = -\frac{1}{2} T_{lw}^* \sum_{d=1}^D a_d (x_{ld} - g_{wd}^{term})^2 + T_{lw}^* \left( v_{l} + s_w - \frac{1}{2} T_{lw}^* - \epsilon_{lw}^{term} \right)$$

(4)

Rearrangement and substitution gives an optimal choice of the form

$$T_{lw}^* = c_{l}^{term} + b_{w}^{term} + \sum_{d=1}^D a_d x_{ld} g_{wd}^{term} - \epsilon_{lw}^{term}$$

(5)

which shares a similar structure to the model of vote choice.

**Placing votes and words in a common space.** Central to SFA is the assumption that both word and vote choice arise from the same preference structure. Intuitively, we assume that people are voting on what they’re talking about and talking about what they’re voting on. Previous works has modeled vote choice in a latent $z$-space ([Clinton, Jackman, and Rivers, 2004](#)), and other work has placed term choice in a latent Poisson space (e.g. [Elff, 2013](#)). If both observed words and observed votes are being driven by the same preference structure, as in when a legislator gives floor speeches and roll call votes, jointly modeling a single preference structure generating both on the same scale and in the same latent space.

In order to do so, we assume that the error term, $\epsilon_{lw}^{term}$, like $\epsilon_{lp}^{vote}$, are independent and follow a standard normal distribution. This places the term choice and vote choice in the same latent space. The parameters $c_{l}^{term}$ and $b_{w}^{vote}$ are individual- and term-specific effects. The ideal points and dimension weights, $a_d$ and $x_{ld}$, are precisely those from the vote equation. We assume a set of cutpoints, $\{\tau_k\}_{k=-1}^\infty$ such that the probability of observing a given term count is the probability of the latent variable falling between two adjacent cutpoints. This connects the latent space to the
observed term count as

\[
\Pr(T_{lw} = k) = \Pr(\tau_{k-1} \leq T_{lw}^* < \tau_k) = \Phi \left( \tau_k - c_{l}^{\text{term}} - b_{w}^{\text{term}} - \sum_{d=1}^{D} a_{d} \epsilon_{ld} \right) - \Phi \left( \tau_{k-1} - c_{l}^{\text{term}} - b_{w}^{\text{term}} - \sum_{d=1}^{D} a_{d} \epsilon_{ld} \right)
\]

with the convention that \( \tau_{-1} = \infty \) and \( \Phi(\cdot) \) denotes the distribution of the standard normal density.

We note that, under this framework, the posterior density is log-convex, meaning there is a single mode. This provides an advantage over topic models, where different starting values may lead to different results (but see Roberts, Stewart, and Tingley, 2015).

### 2.2 Estimation

Much of SFA can be estimated as a standard Bayesian item response theory (IRT) model (see Jackman (2009)). For all parameters except the cutpoints, \( \{\tau_{k}\}_{k=-1}^{\infty} \), and dimension weights, \( a_{d} \), we assume conjugate priors that are normal for mean parameters and inverse-gamma for variance parameters. As we rely on a latent probit specification (Clinton, Jackman, and Rivers, 2004; Albert and Chib, 1993), the error terms \( \epsilon_{lw} \) and \( \epsilon_{vote} \) are assumed independent and identically distributed standard normal variables. Since the number of term- and individual specific effects grow in the sample size, we place separate mean-zero normal priors over each set of effects so as to avoid an incidental parameter problem (Neyman and Scott, 1948) and a Jeffreys hyperprior over the variance.

Lastly, we offer two implementations of the software. The first is a full MCMC implementation, which will generate samples from the full posterior given the data. This allows the researcher to not only estimate the spatial locations as well as the number of latent dimensions, but also to characterize the uncertainty estimates. The second is an EM implementation, which, while it only returns point estimates, is faster than the MCMC implementation and often useful for preliminary results during practical modeling.
As much of the estimation is standard, we defer it to a technical appendix. A generic concern with Bayesian modeling, which we share, is the sensitivity of posterior inference to prior parameter specification. When possible, we use noninformative priors or rely on the existing scaling literature for guidance (e.g., Lauderdale and Clark [2014], Sec 3.2). One component of our model, the Bayesian LASSO, uses two prior parameters set by the researcher. We show in our Supplemental Results that our main findings are robust to shifts in these prior parameters. Our software allows for setting these parameters, and we emphasize that assessing prior sensitivity should be part of any analysis.

**Statistical considerations.** We constructed the statistical model so as to address two concerns when scaling vote and text data. First, we estimate the number of latent dimensions endogenously. Second, we map the observed term counts to and from the latent space using a model designed to accommodate several key features of text data, namely zero-inflation and extreme outliers in the word counts.

In order to estimate the number of latent dimensions, we implement a Bayesian variable selection prior (Park et al., 2008; Tibshirani, 1996). As an alternative, we may have fit the model using maximum likelihood and selecting the number of dimensions via some ancillary criterion. Doing so would have raised the meta-question of which statistic to choose, whether a likelihood-ratio test with a $p$-value cutoff, AIC, BIC, GCV, or cross-validation statistic. Each may give different results. Instead, we place a prior over the dimension weights that will *ex post* estimate some weights as zero, as we describe below. For earlier work using a variable selection prior for matrix subspace selection, see in particular Mazumder, Hastie, and Tibshirani (2010, esp. comparing their Eq. (9) to our Eq. 8 below) and Witten, Tibshirani, and Hastie (2009) for extensions. For variable selection on latent dimensions in a Gaussian copula framework, see Murray et al. (2013); Pitt, Chan, and
We introduce a model for the cutpoints in order to avoid having to estimate thousands of parameters. Our model for the cutpoints addresses three concerns. First, the cutpoints are increasing in the term counts, as higher observed counts imply that the latent variable in our model fell above a higher cutpoint. Second, we want a method that is robust to extreme outliers. In a term document matrix, the largest value may be hundreds of times larger than the mean. For example, in the data we analyze below, the largest value of the term-document matrix ranges from more than 2800 to more than 5300, and we do not want the model’s results to be driven wholly by these shifts. Third, while SFA is robust to shifts at the high end, it is also robust to shifts in the proportion of the term-document matrix that is zero. A key step in analyzing text data involves removing sparse and rarely-used terms. Different choices at this stage result in a different total proportion of zeroes in the term-document matrix. This choice is precisely what Simmons, Nelson, and Simonsohn (2011) mean as a “researcher degree of freedom:” a choice made while collecting formatting the data that is given only cursory discussion and may have a substantial impact on the results. In order to help account for this choice, we include a parameter in our cutpoint model that adapts to the total number of zeroes in the term-document matrix.

**Estimating the number of dimensions.** The method offers two non-standard estimation strategies. First, we place a Laplacian (LASSO) prior over the dimension weights (Park et al., 2008; Tibshirani, 1996):

$$
\text{Pr}(a_d) \sim \frac{\lambda}{2} exp(-\lambda|a_d|) \tag{7}
$$

\[3\] We differ, in particular, from Murray et al. (2013) in that the authors simply use the rank-likelihood and do not estimate underlying cutpoints but take them as determined wholly by the data marginals (see esp. Murray et al. (2013, Sec 2))
This prior provides a principled means of setting most dimensions to zero. As part of our estimation, we naturally recover the maximum likelihood estimate of $a_d, \hat{a}_d^{ML}$. Given $\lambda$, the maximum a posteriori estimate (MAP) estimate for $a_d$ is

$$\hat{a}_d^{MAP} = \begin{cases} 
\hat{a}_d^{MAP} - \lambda & a_d^{MAP} > \lambda \\
0 & a_d^{MAP} \leq \lambda 
\end{cases} \tag{8}$$

The threshold parameter $\lambda$ is estimated within a Gibbs sampler (Park et al., 2008).

Zero-inflation and robust cutpoint estimation. We model the cutpoints as a function of the word count. The likelihood in equation 6 suggests an ordered probit formulation for the cutpoints.\footnote{This likelihood differs from that of an ordered probit as there is no maximal level, see McKelvey and Zavoina (1975) or Greene (2000, pp. 875–879).}

Given the data are counts, with values from 0 to several thousand, we do not want to fit a cutpoint for each value. Instead, we model the cutpoints so as to handle three attributes common to text data. First, the data is zero-inflated: most members do not use most terms in a given year. Second, the data is highly skewed: the observed counts range from 0 to the hundreds. Third, the largest values are highly variable from year to year, and we model cutpoints that are robust to shifts at the high end. We turn next to how we handle each attribute of the data.

To generate cutpoints robust to outliers, we model the cutpoints as function of the empirical CDF. The empirical CDF for a given time period is defined as

$$\hat{F}(c) = \frac{1}{LW} \sum_{l=1}^{L} \sum_{w=1}^{W} 1(c \leq T_{lw}). \tag{9}$$

Using the empirical CDF is equivalent to working with ranks, rescaled from zero to one.

Without additional structure, our model requires estimating hundreds or thousands of cutpoints. In order to avoid the inefficiencies from...
We model the $c^{th}$ cutpoint as

$$
\tau_c|\beta_0, \beta_1, \beta_2 = \beta_0 + \beta_1 \hat{F}(c - 1)^{\beta_2}
$$

where

$$
\hat{F}(-1) = 0
$$

$$
\beta_0 = \tau_0 = \hat{F}(0)
$$

$$
\beta_1, \beta_2 > 0
$$

Modeling the cutpoints in terms of $\hat{F}(c)$ instead of $c$ leaves them less sensitive to extreme outliers. Forcing the intercept, and hence first cutpoint, to be $\hat{F}(0)$ models the zero-inflation directly: the intercept is equal to the proportion of zeroes in the data. Finally, the quasi-linear form of $\beta_1$ and $\beta_2$ allows some flexibility in modeling the cutpoints, while still ensuring that they are an increasing function of $c$. The values of $\beta_1$ and $\beta_2$ are estimated via Hamiltonian Monte Carlo (Neal, 2011); see the technical appendix for details.

**Balancing words and votes.** As there are often an order of magnitude more terms than votes, the researcher may fear that the term data is swamping the vote data. We therefore introduce a parameter, $\alpha$, that controls the relative information coming from each source.

At $\alpha = 0$, all information on the scaled locations comes from votes; at $\alpha = 1$, all information on the scaled locations comes from words. Existing methods allow for the scaling of either text or votes. Even when using data from only text or only votes, SFA offers additional insight over existing methods. When scaling off only the votes, SFA will place words in the same latent space as the votes, giving an ordering to terms driven by the vote information. This can help the researcher interpret

---

5As a Bayesian model, the sources should be averaged and weighted by their precisions. Since the latent space is standard normal, the precision is 1 for each source.
the latent dimension, through finding words that are extreme in the vote dimension. Scaling the two datasets separately does not allow one dataset to inform the other.

We suggest two ways to select \( \alpha \). The first involves fitting \( \alpha \) at a range of values and present the results, showing how they change along these shifts. This is the strategy we follow in our example below, presenting results for \( \alpha \in \{0, 1/2, 1\} \).

We also suggest a data-driven method to select \( \alpha \) such that the ideal points are maximally discriminatory\(^6\). Denote the dimension weights and ideal points as a function of \( a_d(\alpha) \) and \( x_{\ell d}(\alpha) \). Our criterion favors strong dimensions (large values of \( \{a_d(\alpha)\}_{d=1}^D \)) as well as ideal points \( \{(x_{\ell d})_{(L,d)=(1,1)}^{(L,D)} \} \) that provide maximal discrimination among individuals.

In the worst case scenario, dimension weights of zero \( a_d = 0 \) and \( x_{\ell d} = 0 \) across individuals and dimensions would return a value of zero. The criterion selects a point as far from this outcome as possible. The criterion we suggest is

\[
\text{disc}(\alpha) = \sum_{d=1}^D \sum_{l=1}^L \sum_{l'=1}^L a_d(\alpha)^2 (x_{\ell d}(\alpha) - x_{l,d}(\alpha))^2. \tag{14}
\]

All elements of the discrimination statistics are returned from the MCMC output, so our software returns the full posterior density of this statistic, and the optimal value can be selected off the mean. We illustrate the use of this statistic in the Supplemental Results.

**Additional uses.** We have focused on a situation where both votes and term counts are present. There are cases where we observe legislative speech, but we either lack their votes or the votes are so heavily whipped we do not trust them. In this case, as we show below, SFA can leverage words to recover reliable ideal point estimates even in the absence of non-reliable vote data for non-party leaders.

\(^6\)We are rather grateful to Brandon Stewart for suggesting this approach.
Second, SFA recovers the spatial location of words. As these results come from an underlying measurement model, they have a firmer basis than methods that generate right/left measures due to how often each is used by actors with known preferences \cite{Gentzkow2010, LaverBenoitGarry2003}.

Third, SFA estimates the ideological content of terms. Existing studies estimate the political affect of terms by attributing left leaning content to those used by Democrats, and rightward import to those spoken by Republicans \cite{Gentzkow2010}, or modelers posit that terms and votes arise from two conditionally independent data generating processes \cite{LauderdaleClark2014, GerrishBlei2012}. SFA scales terms and votes simultaneously, providing natural structural estimates of word affect.

### 2.3 Comparison to Topic Models and Related Methods

We next compare our method to the popular topic model \cite{Roberts2014, Grimmer2010, BleiNgJordan2003}. In brief, when a researcher is looking for an underlying latent structure ordering actors and their choices, and they are comfortable making the model’s assumptions, then SFA should be implemented. In the absence of any structural or behavioral assumptions, topic models will always return a summary of clusters within the data. We emphasize that this is not an either/or distinction; easy to implement software is available for both. Lastly, we situate SFA in the context of several other related methods.

**How do dimensions and topics differ?** Given text data, the applied researcher will normally reach to a topic model in order to summarize the data. We consider here a simple example in order to help the researcher in deciding between SFA and a topic model, and when either might be appropriate.
Figure 1: **Simulated data setup.** Legislators are arrayed across rows and votes across columns. The darker the square, the more likely the legislator to vote Aye on that particular vote.

The basic difference between SFA and a topic model is analogous to the difference between cluster models and factor analysis. Topic models will return clusters of co-occurring words. SFA will return latent factors, locating actors and terms in a latent space. As an example, assume ten legislators facing six votes. The true underlying probability of voting Aye comes from an underlying process with one ideological dimension, as presented in Figure 1. Legislators are arrayed across rows and votes across columns. The darker the square, the more likely the legislator to vote Aye on that particular proposal. Legislators 1–5 are more likely to vote Aye on the first 3 votes and more likely to vote Nay on the last 3. Legislators 6–10 are more likely to vote Nay on the first 3 votes, and Aye on the last 3. Legislators 5 and 6 are relative moderates, while bills 3 and 4 are relatively noncontroversial.\(^7\)

\(^7\) Specifically, let \(s_i = \{-4.5, -3.5, \ldots, 4.5\}\) and \(w_j = \{-2.5, -1.5, \ldots, 2.5\}\). We drew \(Y_{ij} \sim Bern(\Phi(s_i w_j/2))\).
### Table 1: Results from SFA and a Topic Model on the Simulated Dataset.

We fit both SFA and a topic model to a draw of the vote data. In order to fit a topic model, we assume each legislator uttered six “terms” representing their vote and the bill number. For example, if the legislator voted Nay on vote 4, we assume that they said “Nay on 4,” and enter that into a topic model as a unique word. Each legislator then uttered six elements from the set \{“Aye on 1”, “Nay on 1”, “Aye on 2”, “Nay on 2”, ..., “Aye on 6”, “Nay on 6”\}, one of either “Aye” or “Nay” on votes 1-6. We implemented the EM version of SFA and also gave the same data to the a Structural Topic Model, as implemented in \texttt{stm}. We fit a three-topic model to the data. Four-, five-, and six-topic models returned qualitatively similar results.

The left three columns of Table 1 contain the results from SFA. The first column contains the estimated posterior mode, and only the first dimension has a non-zero mode. The next two columns contain each legislator’s ideal points and the bill estimates. SFA returns estimates of the underlying structure, correctly recovering the unidimensional structure of the data generating process, and identifying legislators 1-4 and 7-10 as relative extremists at opposite ends of the spectrum. SFA also successfully identifies the relatively moderate legislators, 5 and 6, and correctly notes which proposals will draw support from which legislators.
The rightmost three columns of Table 1 report the topic model estimates, presenting the first four terms of the three fitted topics. Consider the first topic. Legislators that vote Nay on votes 1 and 2 are likely to vote “Aye” on votes 4 and 6. Similarly, considering the second topic, legislators who vote “Nay” on votes 4 and 6 are likely to vote “Aye” on votes 2 and 3.

We make three observations from this example. First, both the topic model and SFA produced consistent results. In the course of practical modeling, both should be tried even if one is favored. If their results corroborate, the researcher should be more confident that some systematic attribute of the data is being discovered. Second, SFA returns a low-dimensional ordering and scaling of the data. This stands in contrast to the dozens or scores of topics commonly returned by topic models. Third, SFA presupposes both a structure and sincerity in voting and speaking. Like all assumptions, they can bring insight when correct, but mislead when not. For the researcher interested in summarizing word co-occurrence in a primarily descriptive way, topic models are the appropriate tool.

Existing methods integrating topic models and scaling. A last set of works have combined vote data with a topic model [Lauderdale and Clark 2014] [Wang et al. 2013] [Gerrish and Blei 2012]. The methods, particularly when applied to courts, offer great insight: we may think of courts as voting on a wide array of topics, and judicial preference may vary greatly from one to the next. These models differ from SFA in two crucial aspects. First, the topic model methods do not give an ideological position to terms. Terms are not placed on, say, a left-right dimension shared by the actors. The topic models capture what actors are voting about, but not the ideological structure of the topics. Second, the models offer a disjointed relationship between words and votes. Actors are choosing both words and votes, but only vote choice is grounded in a spatial model. SFA differs by offering a tight, and formal, coupling of the ideological preferences that generate terms and votes.
SFA is a model for term selection that is fully compatible with the standard and accepted models of vote choice.

**Comparison with additional methods.** SFA is related to several existing methods for scaling votes, scaling text, and combining multiple outcomes in a single factor analytic model. Our formal model is an extension of the standard model of Ladha (1991) to term choice. The statistical model is a similar extension of the latent probit model of Clinton, Jackman, and Rivers (2004) to include count as well as binary outcomes. Our estimation technique relies on probabilistic principal components analysis (Tipping and Bishop, 1999), whereby singular vectors and latent factors correspond if the errors are assumed independent and identically Gaussian.

Rather than matrix decompositions in a latent space, several works have turned to a Poisson or negative binomial model in order to model term counts. For example, *Wordfish* of Slapin and Proksch (2008); Lo, Proksch, and Slapin (2014) is similar to SFA, for terms, but under a Poisson or negative binomial link instead of a probit link. SFA leads to an identical formulation for the latent systematic component of word choice ($\theta_{lj}^{term}$), except the latent component is exponentiated in order to guarantee positivity (see also Elff, 2013; Bonica, 2014). SFA differs in that it places both word and vote choice in the same latent z-space, allowing both types of data to be modeled jointly. Our cutpoint model allows for a more flexible mapping from the latent to observed data space. We also model the zero-inflation directly and are robust to outliers, as described above. SFA also estimates the underlying dimensionality.

The use of the Poisson by Slapin and Proksch (2008) notwithstanding, many analysts eschew both the Poisson and the Negative Binomial in their analysis of text. Examples include: the widely

---

8Note that most of the theoretical and algorithmic contributions can actually be found in Aldrich and McKelvey (1977).
used Wordscores model of Laver, Benoit, and Garry (2003) or the nonparametric content analysis of Hopkins and King (2010); the text based “slant” measures used by Gentzkow and Shapiro (2010); the LDA formulation of Blei, Ng, and Jordan (2003); Gerrish and Blei (2011) which converts term counts to proportions, thereby admitting a Dirichlet prior; and PCA or kernel PCA on the tf-idf matrix Spirling (2012). SFA’s use of a latent normal representation is, like these, not based on a Poisson model; see Murray et al. (2013, esp. 2.1) for a formulation close to SFA’s.

Mixed factor analysis models have been used to combine data of different types. In these models, mixed data, such as counts, binary, and continuous data, are placed on and analyzed on an common underlying latent scale. The Gaussian-, or $z$-, scale is a common choice. For example, Quinn (2004) converts observed continuous data to a $z$-scale, and then combines it with ordinal and categorical data on the same scale; for recent extensions, see Murray et al. (2013); Hoff (2007). Our model is closely related to that given in Murray et al. (2013, sec. 2.1). We differ in that we only have two types of data, text and votes, and hence have to estimate only one set of cutpoints, whereas Murray et al. (2013) consider the problem of generic types of data. Murray et al. (2013) work with the transformed ranks of all variables, eschewing cutpoints entirely. For that reason, our method is more powerful when modeling votes and text. SFA also introduces a cutpoint modeled tailored to several common aspects of text data, as we describe above.

Other methods have estimated dimensionality (Hahn, Carvalho, and Scott, 2012; Heckman and Snyder Jr. 1997). Aldrich, Montgomery, and Sparks (2014) show that sufficiently large cross-party variance can mask important within-party dimensions. We differ from these works in combing both vote (binary) and word (count) data.

To illustrate the use and efficacy of SFA, we analyze text and roll call votes from the contemporary US Senate.
3 Illustrative Application: The US Senate, 1997–2012

In this section, we apply SFA to recent US Senate data. The analysis proceeds in four steps. First, we describe the data and discuss the viability of SFA in this context. Second, we apply the method and present results. Third, we present several tests of internal validity. Fourth, we present a test of external validity through using the legislative model to scale newspaper editorials.

3.1 Data

We apply SFA to the eight recent sessions of the US Senate. We scale using both votes and words, returning both ideology estimates and our calibration of the underlying dimensionality. Our data come from two sources. Rollcall data come from VoteView. For the text data, we rely on floor speeches as gathered by the Sunlight Foundation. Following standard practice (e.g., Quinn et al., 2010; Grimmer and Stewart, 2013), we stem, eliminate stop words, and model unigrams and bigrams. Both vote and speech data are polled over the full session. We trim all terms that are not used by at least ten people at least ten times over the course of the session. A complete summary of the data can be found in the supplemental materials.

Before applying the method, we check that SFA is applicable in this case. First, voting is not always sincere in the US Senate, as there are always motions to recommit, etc. We note, though, that ideal point estimates from the US Congress have been used extensively in other studies and possess high face validity. To be particularly careful, if a bill is voted on several times due to different motions, we only include the final vote in our analysis.

Second, we consider the sincerity when speaking. Previous work has shown, albeit in the US

Figure 2: Posterior density over number of underlying dimensions for the joint word and vote model. We find a pronounced mode at two dimensions consistently across Senates. The average across all Senates appears in the top left corner.

House, that floor speeches are expressive rather than deliberative (Hill and Hurley 2002; Maltzman and Sigelman 1996). Many floor speeches aren’t even read verbally, but simply entered into the record, also suggesting that floor speeches are vehicles of expression rather than persuasion. For that reason, we feel more comfortable applying the method to floor speeches rather than, say, conference committee meetings.
3.2 Results

We present three sets of results. Each corresponds with the proportion of information in the votes coming from votes instead of words (\(\alpha \in \{0, 1/2, 1\}\)). We present results on the estimated number of dimensions and interpretation of each.

**Scaling results informed only by votes** (\(\alpha = 0\)). We begin with the model with information coming only from votes. This model places a posterior mass estimate of 100% on one dimension for each Senate. Posterior means of ideal point estimates correlate with DW-NOMINATE estimates ranging from 0.95 to 0.98 across the eight Senates analyzed here. See (Clinton, Jackman, and Rivers, 2004, Figure 1) for similar results.

**Scaling results informed by words and votes** (\(\alpha = 1/2\)). We next move on to the model that gives equal weight to words and votes. First, we consider the estimated number of dimensions, see Figure 2. The average density over the number of dimension parameters merging all Senates is in the top left corner, while the successive sessions are depicted from top to bottom and from left to right. A pronounced mode at two dimensions reappears consistently across Senates.

Not only is the finding of two dimensions consistent, but the two dimensions themselves are stable across sessions. The first closely coincides with the standard ideology dimension uncovered from scaling roll call votes. The second appears to be a leadership dimension, with party leaders at one end while a variegated mix of “rank and file” partisans and ideological moderates populate the other.

how SFA may recover dimensions attributable to non-positional preference. Our second dimension in the US Senate is anchored on one end by party leaders. These leaders have preferences over certain words in a manner largely uncorrelated with their vote choices. Under our model, votes
Distribution of Words Along Ideological Dimension when Scaled with Votes

108th Senate

112th Senate

Log Density

Scale

Log Density

Scale

administr
cut
health
need

expenditure
period
time
shall
apply

commit
foreign
procedure
immediate
consent
proceed
immediately
consider

consent
committee
author
meet
session
900
10

author
meet
session
consent
committee
conduct
hear
offic building

draken
offic
peal
entill
women
10

want
take
2010
detemin
option
spring

budget
stimulus
deal
trillion
spend

Figure 3: Log density of term weights, after scaling votes and terms together. Each local mode is labeled by the five terms closest to that mode. The left figure presents results from the Republican controlled 108th Senate, the right figure contains results from the Democratic led 112th Senate.

and terms map to a location in a latent space. Inferring this most preferred location from observed data is exactly a spatial representation of the most preferred outcome for each legislator.

Figure presents the log density of term weights, after scaling votes and terms together. The weights are oriented such that terms more likely to be spoken by Republicans are to the right. Each local mode is labeled by the five terms closest to that mode. The left figure contains results from the 108th Senate, a Republican-led session during President George W. Bush’s tenure. The right figure contains results from the 112th Senate, a Democratic-led session during President Barack Obama’s time as President.

We find a consistent pattern: for the majority party, the most extreme terms relate to parliamentary control words ("consent committee," “author meet,” “meet session”). For the minority party, the first dimension identifies ideologically relevant terms. For the Democrats during the 108th
Figure 4: Latent dimensions estimated by SFA, 112th Senate. Legislators’ preferred outcomes on the first dimension (x-axis) and the second (y-axis). The left plot labels party leaders, whips, and top chairmen. In the right plot cutting lines separate frequent from infrequent users, of the terms: “Boehner,” “student loan,” and “fiscal cliff.”

Senate, these terms included “adminstr,” as the Democrats soured on the current Presidential administration, and “health,” a centerpiece of the Democratic policy agenda. In the 112th Senate, with the Democrats in the majority, parliamentary control terms switched their ideological polarity, aligning with the Democrats (“author meet,” “meet session,” “consent committee”). The Republican end of this first dimension reflects that party’s programmatic concerns over fiscal balance (“budget,” “stimulus,” “debt,” “trillion”).

Next, we look at the preferred outcomes of legislators from the 112th. Points in Figure 4 are shaded in proportion to their first dimensional DW-NOMINATE score, showing the agreement between SFA and DW-NOMINATE on the first dimension ($\hat{\rho} \approx 0.95$). The left plot labels party leaders, whips, and top chairmen, showing the close relationship between locations on the second dimension and leadership. The first dimension captures the political battle lines, reflecting legisla-
tors left vs right policy differences, while the second, vertical, dimension reflects differences in the terms selected by leaders versus the rank and file members.

The right plot of Figure 4 contains cutting lines for three terms: “Boehner,” “student loan,” and “fiscal cliff.” The lines were constructed such that legislators on one side are expected to use the phrase above the median number of its raw usage, and on the other side legislators are expected to use the word below its median number of times. We find leaders are more likely to use the word “Boehner,” the House Speaker during this session. Republicans were more likely to use the term “fiscal cliff,” with leaders the most likely. Democrats were more likely to utter the phrase “student loan,” again with leaders the most likely to employ the term. SFA identifies a group of Republican moderates in the top “V”, here we label them by name. These moderates are not likely to use either “student loan” or “fiscal cliff.”

Scaling results informed by only words ($\alpha = 1$). We also apply SFA using only information from words. This is not our preferred model, as it ignores vote data, yet SFA still uncovers structure in the text data. The posterior density of estimated dimensionality for pooled floor speeches can be found in Figure 5. Results across all sessions are in the top left corner while the remaining sessions follow in order from top to bottom and from left to right. In contrast with the high concentration of probability on two dimensions in our preferred model, when we exclude the valuable information contained in votes and analyze oratory alone, we obtain a somewhat more diffuse density that accords a 75% probability to there being between five and eight dimensions, and a probability of over 95% that the underlying dimensionality is within the range $[4,11]$. Looking at individual sessions, we find a similar dimensionality, albeit with some year-to-year variation.

Figure 6 contains the top ten words at each of the first six dimensions of the 112th Senate.¹¹

¹¹ Results from the 105th–111th Senates are available with the supplemental materials.
Figure 5: Estimated underlying dimensionality for Senate floor speeches. Results across all sessions are in the top left corner and remaining sessions follow.

We note that the positive and negative level distinction along the $y$-axis is wholly arbitrary, as we only identify term levels up to a sign. Looking at the first column, we find that the first dimension starts with a set of non-controversial terms. These include parliamentary procedural terms (as opposed to parliamentary control terms) such as today wish, madam rise, and colleague support. Also on the non-controversial side are martial terms with universally positive affect during this Congress such as army, air forc, and deploy. On the other side are words that will be used in to
differentiate issues in other dimensions, such as *tax, vote*, and *peopl*. The other dimensions have at their extremes words connoting some underlying dimension of policy. For example, the second dimension ranges from judiciary and women’s issues at one end to fiscal concerns at the other; the fourth goes from a broad set of social welfare concerns to the consideration of judicial nominees. These lower dimensions adapt to the issues of the day. Tobacco, for example is present in the 105th Senate; Iraq comes and goes as an issue, and health care goes from dealing with seniors and Medicare in the 107th Senate to dealing with students and families in the 112th.

---

**Figure 6: Extreme Terms by Dimension, 112th Senate.** Extreme terms for the first six dimensions as estimated by SFA from the 112th Senate. The typesize of each term is proportional to the absolute value of the associated coefficient; terms earning positive coefficients appear in the upper part of the panel, those assigned negative coefficients are presented in the lower segment.
3.3 Internal Validity

A scaling method is only as persuasive as its demonstrable construct validity. We turn now to assessing SFA’s internal validity in the US Senate data.

Imputing estimates for legislators missing completely at random. First, we randomly discard the votes cast by ten legislators selected completely at random, coding all of their votes as “missing,” while we maintain all of their speech data. The left and right panels of Figure 7 plot
Figure 8: **Estimated Ideology when Only Leaders Votes are Informative.** The voting dimension estimates appear in the left panel, with the censored estimates measured on the vertical (y-axis) while the uncensored ones appear on the horizontal (x-axis). In the censored data the salience of the voting dimension drops, so that it becomes the second dimension. The righthand panel exhibits the leadership dimension, again the censored estimates correspond with the vertical (y-axis) and the uncensored ones coincide with the horizontal (x-axis).

the imputed versus fitted values (“X”) for the dropped legislators, for the first (left) and second (right) dimension. SFA recovers reliable first-dimension preferred outcomes well, except for some expected attenuation bias. The second dimension ideal points are recovered almost exactly. We remind the reader that the first SFA dimension coincides closely with the dimension that emerges from an analysis of the votes alone, and so we might expect it to be more affected by the loss of voting data, while the accuracy of our second dimension estimates, which are dominated by speech data, would be expected to suffer less from the censorship of the votes.

**Imputing estimates for members’ given only votes from leadership.** We next offer a more challenging test of internal validity. For this analysis, we coded all vote data except for the party
leaders and whips as missing, while maintaining all speech data. This left a vote record for less than 4% of the Senate. We then compared the SFA ideal point estimates to the SFA estimates using everyone’s speech, but only leaders’ votes. Essentially our exercise in censorship diminishes the importance of the dimension related to voting. When we estimate the censored data we again recover two dimensions, but their order is reversed, with the voting dimension becoming noisier, and falling into second place, while the leadership dimension, the evidence for which comes almost entirely through legislative speech, earns the higher dimension weight, see Figure 8. The left panel of the figure compares estimates for the voting dimension, which is the second dimension estimated with the heavily censored data (plotted along the vertical $y$-axis) while it corresponds with the first dimension of the uncensored estimates (graphed relative to the horizontal $x$-axis). Observations are labeled by party, and leaders’ locations are in bold and circled. As one would expect, with less than $1/25^{th}$ of the voting data, recovery of the first dimension is far from perfect, but remarkably the imputed scores correlate highly, at more than 0.85, with the estimates based on the full data set. The right hand panel compares estimates for the “leadership” dimension, which coincides with the first dimension based on the censored data, but with the second dimension based on the uncensored data set. In contrast with the voting dimension, the censored estimates correspond closely with their uncensored counterparts. Of course, the “leadership” dimension is driven mostly by words, and we did not censor those.

While this last exercise may seem a stunt, we note that in heavily whipped parliaments most legislators vote their parties, rather than their preferences (e.g., Kellerman, 2012), yet they still give speeches. In such settings we might use SFA to “bridge” between speeches actually given by members of a parliament to the votes that they would have cast had they not been “whipped”, anchoring the excercise by treating the votes of party principals as a genuine reflection of the leaders’
Figure 9: **Scaling newspaper editorials given only their text.** This figure presents the relative locations and differences between the ideal points for legislators and newspapers.

3.4 External Validity

Lastly, we consider assessing SFA’s ability to recover preference estimates with external validity. So far, we have used SFA to impute from legislative members to other members during the same session. We next turn from imputing ideal points for legislative members to imputing ideal points for non-legislative actors, namely newspaper editorials.
We apply SFA to word count data from unsigned editorials published during the two years that the 112th Congress was in session in the *New York Times*, the *Wall Street Journal*, and the *Washington Post*, using the same terms we employed in our analysis of the Senate. As above, we combine the word counts of these editorials with the Senate data, treating the editorials as legislators with a missing vote record.

As the term data come from different venues, the Senate floor versus the editorial page, the exercise is one of “out of sample prediction”. This leaves us with the question of whether the political meanings of the terms of discourse are the same in both venues. As a first approach to this issue, we treat the ideal points for both groups as coming from a mean-zero distribution. Results appear in Figure 9. We orient the dimension so that the Republicans have a positive value. The densities for the Republican and Democratic Senators are in the background, and the voting dimension legislator preferred outcomes are plotted as hatch marks along the $x$-axis. The results are largely as expected. If we treat the three sets of editorial boards as legislators who do not vote, we find the Wall Street Journal (WSJ) to the right of the Washington Post (Wash Post) and the New York Times (NYT) to its left. The distance between the Wall Street Journal and the Washington Post is about half the estimated distance between the New York Times and the Post.

4 Conclusion

We propose a method, Sparse Factor Analysis, for combining votes and text data in a single scaling procedure. The method models both word choice and vote choice in terms of the same ideal points. We furthermore develop a statistical framework that allows us to estimate both individuals’ most preferred outcomes and the underlying dimensionality of the joint word-vote space. The resulting methodology provides a close linkage between the choice-theoretic models of vote and word choice.
This tight connection allows the extension of SFA to more complex decision scenarios (Clinton and Meirowitz, 2003, e.g.). SFA allows the analyst to estimate the underlying number of latent dimensions, rather than having to impose dimensionality a priori.

Substantively, we analyze legislative speech and roll call voting from eight recent sessions of the US Senate. Combining both data sources reveals a consistent picture of a two dimensional Senate, with the first dimension coinciding with the voting dimension, while the second distinguishes leaders of both parties from the rank and file.

While SFA is designed to analyze individuals who both speak and cast votes, it allows us to impute policy preferences to non-voting political speakers, a potential we illustrated for the case of newspaper editorial boards. This may prove useful in confronting the perennial research problem of imputing the preferred policy outcomes of legislative candidates. While analysts can impute the ideology of victorious candidates from their subsequent congressional conduct, as they can infer the leanings of defeated incumbents from their previous voting records, measuring the preferences of defeated challengers has proven to be a more elusive goal. Yet every challenger spends time and energy generating political speech. SFA offers the possibility of imputing the most preferred policy such a candidate would have pursued had he been elected.

We hope the approach in this paper also finds purchase beyond the US Congress. For example, in strong party systems where votes are relatively uninformative, words may be used to help clarify the within-party variance in ideal points. We are currently exploring applications of the method in situations where voting is not perfectly reflective of underlying individual preference or where ideal points are allowed to evolve over time.
References


Aldrich, John, Jacob Montgomery, and David Sparks. 2014. “Polarization and Ideology: Partisan Sources of Low Dimensionality in Scaled Roll Call Analyses.” *Political Analysis*.


Lo, James, Sven-Oliver Proksch, and Jonathan B. Slapin. 2014. “Ideological Clarity in Multiparty Competition: A New Measure and Test Using Election Manifestos.”.


