Pivoted Text Scaling for Open-Ended Survey Responses

William Hobbs*

September 28, 2017

Abstract

Short texts such as open-ended survey responses and tweets contain valuable information about public opinions, but can consist of only a handful of words. This succinctness makes them hard to summarize, especially when the texts are based on common words and have little elaboration. This paper proposes a novel text scaling method to estimate low-dimensional word representations in these contexts. Intuitively, the method reduces noise from rare words and orients scaling output toward common words, so that we are able to find variation in common word use when text responses are not very sophisticated. It does this using a particular implementation of regularized canonical correlation analysis that connects word counts to word co-occurrence vectors using a sequence of activation functions. Usefully, the implementation identifies the common words on which its output is based and we can use these as keywords to interpret the dimensions of the text summaries. It is also able to bring in information from out-of-sample text data to better estimate the semantic locations of words in small data sets. We apply the method to a large public opinion survey on the Affordable Care Act (ACA) in the United States and evaluate whether the method produces compact, meaningful text dimensions. Unlike comparison unsupervised techniques, the top dimensions produced by this method are also the best predictors of issue attitudes, are well-distributed across respondents, and do not need much information from higher dimensions to make good predictions. Substantively, over time changes in the prevalence of the text dimensions help explain why efforts to repeal the ACA in 2017 were fragmented and unsuccessful.

Open-ended survey responses help researchers avoid inserting their own expectations and biases into their findings and allow for unexpected discoveries. Gleaning systematic information from

*The author appreciates comments and feedback from Adam Bonica, Nick Beauchamp, Chris Callison-Burch, James Fowler, Lisa Friedland, Dan Hopkins, Gary King, Kokil Jaida, Kenny Joseph, Ani Nenkova, Molly Roberts, Brandon Stewart, and Lyle Ungar. Special thanks to Dan Hopkins who is a co-author on a broader, substantive project on the Affordable Care Act and who graciously provided the data for this text method paper. This project was generously supported by the Russell Sage Foundation (grant 94-17-01)

unstructured open-ended responses, however, can be challenging. People write on their own terms and many write incomplete sentences using only a small number of loosely connected keywords. In the data we will use here, for example, the mean number of words in the responses is only 7 and 20% of the responses use 3 or fewer words not contained in a widely used stopword list.[1]

Bag-of-words approaches, including topic models (Blei, Ng and Jordan, 2003; Blei and Lafferty, 2007; Roberts et al., 2014) and scaling models (Deerwester et al., 1990; Slapin and Proksch, 2008), can work whether or not there is much grammatical structure. But standard methods are intended for analyses of general and sophisticated text corpora rather than short survey responses on a single issue. Because of difficulties inherent to studying general corpora, especially difficulties in accounting for common words that can span many topics (Wallach, Mimno and McCallum, 2009), they are designed in a way that does not take full advantage of information contained in common words. This reduces the ability to represent open-ended survey text in a small number of highly predictive and interpretable dimensions.

This paper proposes a method to better estimate the meaning of short and probably vague text on a focused issue, such as open-ended survey responses on a public policy or tweets about a protest movement. The method is similar to standard text scaling methods but reorients its output away from rare words and toward meanings in common words. To do this, its implementation uses a regularized canonical correlation analysis (CCA) between in-sample word co-occurrences and out-of-sample word embeddings (e.g. the average meaning of a word across all text on Wikipedia or Twitter) weighted to reflect in-sample word volumes. The implementation is closely related to text scaling methods based on latent semantic analysis (Deerwester et al., 1990), including methods widely used in political science such as WordFish (Slapin and Proksch, 2008) and correspondence analysis (Lowe, 2007, 2016).

The method, which we call canonical pivot analysis, uses few to no researcher defined hyperparameters in order remove the researcher from the measurement process.[2]

[1] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop (SMART)

[2] The hyperparameters are used only to induce a specific 'pivot' behavior that reorients output toward common

The specific approach resembles pivots used in domain adaptation (Blitzer, Foster and Kakade, 2011). These methods adapt general machine learning models to a different or more focused task. Typically, pivots are common words that do not have different meanings or functions across the two contexts, and they are the axes on which adaptation from one context to another is based.

We use common words in our text scaling method more or less how they are used in domain adaptation. We use them to adapt our scaling from rare words toward common words and to bring in information from out-of-sample data. Mechanically, our pivots are common words for which we are able to identify shared or symmetric representations across two contexts – in-sample word co-occurrences and out-of-sample word embeddings heavily weighted by our in-sample word counts. We find these symmetric representations when words exceed a soft threshold of frequency and specificity.

More intuitively, these pivots are moderately common to very common words that tend to appear with a certain set of words. That is, they are common and somewhat specific. Many people say these words and, when they say them, we can make a reasonable guess about what else they could have said – but often didn't say in only 7 words. Existing text scaling methods also implicitly optimize some form of this prediction.[3] Unlike existing methods, however, we have a relatively low bar for our guess, especially if a word is very common. Instead, we focus on getting a machine to identify the gist of a response that states, for example, only 'how are we going to *pay* for it' (emphasis added), associate common words that fall along a similar line of argument, and then order these word associations according to how common and coherent they are in the text. In focusing on the gist of a response, the pivot words are the axes on which we orient the output away from rare words and toward common words. [4]

Beyond the improved performance on short text, the method provides a few nice additions

words. We suggest reasonable ranges for these parameters. In our experience, changing the values of the hyperparameters at reasonable levels has very little effect on the lowest dimensions of the results.

[3] See, for example, Levy and Goldberg (2014).

[4] Another way to think of this is that we stretch distances for common words.

to standard text scaling that improve interpretation and stability. In particular, it provides a keyword metric (that is also the basis of the optimization) and a means of incorporating outside data. Keywords are very helpful for interpreting text summaries on multiple dimensions, but are not provided in the output of standard text scaling methods (they are important in topic models instead). Out-of-sample data, meanwhile, can help text scaling methods work better on small data sets.

We apply pivot analysis to a survey on attitudes toward the Affordable Care Act (ACA), and contrast the results with output from topic models and from text scaling techniques that also do not enforce categories on outputs. We find that pivot analysis is as good as standard factorizations at predicting issues attitudes in high dimensions and, critically for small social surveys, that it is much better at predicting responses in few dimensions.

Comparisons on additional survey responses show that the representations' top dimensions reflect cleavages between and within U.S. political parties. The different dimensions help provide explanations for changes in attitudes toward the ACA and relationships between dimensions of ACA attitudes and presidential candidate vote choices. The specific changes and the time frames over which they occurred provide clues to explain why repealing the ACA in 2017 was so difficult.

Uses for the Method

The method in this paper is designed to analyze short text data on a focused and potentially polarized topic. It is well-suited to many open-ended survey responses and to opinion statements on social media.

In particular, the method is tailor-made for open-ended survey responses on a specific issue, such as attitudes on abortion or immigration policy. The method will 'summarize' these texts even though they are very short and contain much less information than a document like a news article, press release, or speech.

It is also applicable to tweets and text from social media on a focused topic, such as tweets

containing a specific hashtag accompanied by a personal political statement. Well-known examples of these kinds of texts are tweets containing the text "#BlackLivesMatter" and "#YesAllWomen".[5] These texts are both public opinion statements and influential parts of political movements.

Specific Application and Motivation

Our specific motivation in developing this method is to summarize information contained in open-ended responses on attitudes toward the Affordable Care Act. This is part of a larger project on public and politician attitudes toward the law. The project will incorporate text responses to explain how people think about the ACA and how they justify their support or opposition to it.

Broadly, the effort aims to better understand dimensions of partisanship, the stability of attitudes toward the ACA over time, and why efforts to repeal and replace the ACA in 2017 were so fragmented, even though Republicans were unified in their dislike for the law. The text summaries will supplement analyses based on closed-ended surveys. Although we have a large amount of closed-ended data, we are limited in the number of questions we can ask, we do not always know what to ask ahead of time, and it is possible that our questions will create opinions on the ACA that respondents did not hold before we asked them.[6]

These summaries should be able to score even very short or seemingly vague responses, since respondents on political science surveys often hold strong attitudes without sophisticated or policy-based justifications for them.

Also, given our interest in both policy perceptions and within party conflict, these summaries

[5] The "#BlackLivesMatter" rose to prominence on Twitter after black teenager Michael Brown was killed by police in Ferguson, MO in August 2014. See more info here: https://www.nytimes.com/2016/08/23/us/how-blacklivesmatter-came-to-define-a-movement.html. The "#YesAllWomen" hashtag emerged after six people were killed near the campus of UC Santa Barbara in May 2014 by a man who blamed "the cruelness of women" for the attacks. See more info here: http://www.cnn.com/2014/05/27/living/california-killer-hashtag-yesallwomen/index.html. And here: http://time.com/114043/yesallwomen-hashtag-santa-barbara-shooting/.

[6] For example, our question wordings could make certain aspects of the ACA more salient than others, and do this in an unrealistic way. Our emphasis could then lead respondents to create opinions simply in response to our question (Zaller and Feldman, 1992).

should be able to discover multiple dimensions of attitudes and do this without supervision (i.e. without telling the method whether a person likes or dislikes the ACA or is a Republican or Democrat). Since ACA attitudes are correlated with partisanship at 0.65 in our data, supervised methods that project words onto a single dimension will recover that variable, whether or not the words tell us much about policy attitudes.

These motivations help decide what technique we use to analyze the data. Currently, there are two broad approaches to summarizing text data without supervision: topic modeling and scaling methods. Topic models, such latent Dirichlet allocation (Blei, Ng and Jordan, 2003), correlated topic models (Blei and Lafferty, 2007) and stuctural topic models (Roberts et al., 2014), are a form of source separation and split documents and sets of vocabulary onto distinct categories. This source separation works well on long and/or diverse corpora and it typically requires the researcher to specify the number of categories in the data a priori.

Scaling methods, on the other hand, compress variance in text usage onto a small number of continuous and potentially polarized variables (i.e. positive and negative variables). They work well on focused text corpora with sophisticated speakers. In political science, text scaling methods, including WordFish (Slapin and Proksch, 2008) and WordScores (Laver, Benoit and Garry, 2003; Lowe, 2007), are used as "ideal point" methods, with estimates similar to those from Poole and Rosenthal's NOMINATE on roll call votes (Poole and Rosenthal, 1985).[7] Scaling methods often do not require the user to specify the number of dimensions of the output, and the dimensions of the output have a natural ordering that is the amount of variance in the source data that an output dimension explains.

In analyzing our data on attitudes toward the ACA, we prefer a text scaling method over a topic model. All of our survey responses are about the same issue (i.e. the same topic), and so are hard to separate into distinct categories. Further, political conflict in the United States

[7] All of these text methods are well known (Lowe, 2016) to be closely related to latent semantic analysis, which uses singular value decomposition on a standardized term-document matrix.

is polarized and extremely low-dimensional, so a text scaling method that describes a polarized and low-dimensional semantic space will often be more useful than distinct but high dimensional topics.

Data and Challenges

We have a very large number of open-ended survey responses on the Affordable Care Act that we can use to study public attitudes on the law. Over 9,000 open-ended responses on the ACA were collected by the Kaiser Family Foundation and Pew Research Center between 2009 and 2016. These two data sets are publicly available and have been analyzed in prior work (Hopkins, 2017). We add to this data approximately 3,000 responses in 2016 from our own survey of political activists, people who are members of a political party and have high levels of political participation, along with 1,000 responses in 2016 from a national representative sample.

In the data, 11,000 or so respondents were asked two questions at the beginning of a longer survey on health care policy attitudes. The first two questions were: 1) "As you may know, a health reform bill was signed into law in 2010. Given what you know about the health reform law, do you have a generally favorable or generally unfavorable opinion of it?" 2) "Could you tell me in your own words what is the main reason you have a favorable/unfavorable opinion of the health reform law?". Around 2,000 thousand respondents were asked two similar questions before the ACA was signed into law.[8]

Although we had many responses, each response on its own appeared to contain very little information. The mean number of words in these responses was only 7 (median 6) and 20% of the responses used 3 or fewer words. Many respondents used the same words, for example: health (4,594), people (4,002), insurance (3,635), think (2,024), will (1,397), and government (1,305).

[8] Closed-ended: "As of right now, do you generally favor or generally oppose the health care proposals being discussed in Congress?". Open-ended: "What would you say is the main reason you favor or oppose the health care proposals being discussed in Congress."

Around 9 out of 13 thousand respondents used at least one of these words, and 4,500 people used only the top 100 words in the corpus plus one other word.

However, these common words were unevenly distributed across respondent types. For example, Republicans were significantly more likely to use the word "government" to justify their attitudes toward the ACA.

Ideally, we would have used an existing method to analyze variation in these ACA responses. We discovered, however, that scaling methods struggled to estimate the locations of common words. The existing scaling methods standardized word frequencies before estimation and this equalization effectively upweighted sophisticated words at the expense of common words.[9] In practice, this scored common words close to each other and spread them across many dimensions of the output.[10] Since most respondents only used common words, this limited our ability to use most of the responses in low dimensional and interpretable models, even as we observed clear partisan variation in common word use.

Due to this difficulty, we designed a method that was similar to standard text scaling, but performed well on short, keyword based responses on a focused and polarized topic. Because so many respondents used a small number of common words, we considered the possibility that these words were particularly important, and that they would provide clues to the overall structure of opinions. We tested this by orienting the overall word representations toward the most common words, so that common words were not erroneously scored close together and so that more precise terms mostly strengthened signals or disambiguated the common words.

We also added out-of-sample word embeddings to better estimate the *moderately* common words' representations. Moderately common words affect the document scores for many respondents but have substantially sparser in-sample co-occurrences than the most common words. This

[9] As well as, in some cases, words that regularly appeared as the only word in a sentence. This was a major problem with correspondence analysis compared to PCA on the standardized word co-occurrence matrix. The chi-squared distribution was a poor null model for the distribution of words.

[10] This is a generally accepted problem in text scaling methods and topic models.

adjustment helps our method perform well on even small numbers of open-ended survey responses.

Method

Our proposed method for scaling open ended survey responses is based on a decomposition of a particular covariance matrix. The decomposition it leverages, canonical correlation analysis (CCA), is fundamentally a linear regression with multiple dependent variables.

In a typical use case, a CCA on text works very much like standard text scaling such as latent semantic analysis (LSA) (Deerwester et al., 1990) on a term-document matrix, a singular value decomposition of a standardized co-occurrence matrix (Bond and Messing, 2015), or correspondence analysis (Lowe, 2016).[11] The primary difference between the CCA and these other methods is that a few adjustments to CCA and our input data will allow us to simultaneously 1) re-orient the factorization around common words; 2) add information from out-of-sample word embeddings; and 3) estimate keywords for each dimension.

Broadly, the 'pivoting' in this method is a way of weighting our scaling output toward common words without creating dimensions in our output that encode word frequencies and without weighting the output toward common words that are overly general. In practice, the output is similar to a tf-idf standardization, which assumes that very common words are not specific, but does not insert that functional form ex ante. Instead, the method relies on the structure of text data, especially an inverse relationship between word frequencies and the specificity of words' conditional word co-occurrence probabilities, to create the standardization. We call the behavior pivoting both because of a mechanical resemblance to pivots in other natural language processing methods and also because we pivot our output away from rare words and toward common words and, to a limited extent, toward words' semantic locations in out-of-sample data.

Importantly, our setup appears to be difficult for a researcher to manipulate. Further adjustment

[11] Note that canonical correlation analysis in Lowe (2016) is what we refer to as correspondence analysis. CCA here is a different matrix factorization, though it is very similar to a weighted correspondence analysis.

of the hyperparameters, within ranges that produce the desired 'pivot' behavior, have only limited effects on the lowest dimensions of the results, though they can be changed to bring in more or less smoothing from out-of-sample data.

We summarize our notation in Table 2 and the algorithm in Table 3. Step 3 in Table 3 is the central component of the method, the CCA. Other steps either feed into step 3 or apply output from it to the text documents we wish to analyze.

Note that the explanation for this method is somewhat involved, but the word score estimation itself is essentially one big moving part. Each step in the setup is tied to another.

The out-of-sample word embeddings are the exception to this single moving part, however. Pivot scores can be estimated without out-of-sample data and our application will produce almost the same as in-sample data only output using this method, given the hyperparameters we choose. We introduce the option here because it has the potential to be useful in cases where open-ended survey responses are less abundant.

Overview of Canonical Correlation Analysis

Before introducing pivot analysis, we first describe the more general canonical correlation analysis on which our method is based. Canonical correlation analysis uses a singular value decomposition (SVD) on a covariance matrix between two sets of variables. The SVD is an orthogonal transformation of data that compresses variance into as few variables as possible. After applying SVD, it is possible to truncate the output so that we are left with a small number of variables that still retain a large amount of information from the original data. This is useful when we have a large number of correlated variables from which we want to extract a small number of representative variables.

The SVD in a typical CCA is run on the covariance matrix between two sets of variables and their inverted covariance matrices. Like in a linear regression, the inverted covariance matrices adjust for different units across varying types of data. In its estimation, the SVD optimizes Pearson correlations, or cosine similarity between centered matrices:

$$\max_{\phi_x,\phi_y} \frac{\phi_x^\top C_{xy}\phi_y}{\sqrt{\phi_x^\top C_{xx}\phi_x}\sqrt{\phi_y^\top C_{yy}\phi_y}} \tag{1}$$

In this formula, C_{xy} is the covariance matrix of X and Y, where X is one set of input variables and Y is another input, while C_{xx} is the covariance matrix for X alone and C_{yy} for Y alone. ϕ_x is an eigenvector of $C_{xx}^{-1}C_{xy}C_{yy}^{-1}C_{yx}$ and ϕ_y is an eigenvector of $C_{yy}^{-1}C_{yx}C_{xx}^{-1}C_{xy}$, where $^{-1}$ indicates an inverted matrix.

ϕ_x and ϕ_y project the X and Y matrices onto a shared latent space that is a good representation of both data sets. These singular vectors are the coefficients from the model, like βs from a linear regression. Using a slightly simplied formula (Dhillon, Foster and Ungar, 2015), we multiply the singular vectors by either the left, X, or right, Y, input to the CCA to obtain the variables' locations in the shared space:

$$\phi_x^{proj} = C_{xx}^{-1/2}\phi_x \tag{2}$$

Canonical correlation analysis is typically used when there are two types of data that reflect the same underlying state, such as audio and video of an event or two translations of a speech. CCA maximizes correlation between two sets of data to estimate the shared underlying, or latent, state (e.g. the recorded event). In this alignment, attributes of one side of the data that do not appear in the other, or that do not help maximize correlation with the other side, are thrown out in the estimation of latent variables.

As an example of the use of CCA on text (and the primary inspiration for its use here), Dhillon et al. (2015) use CCA to take advantage of both the left (before) and right (after) contexts of a word in a sentence to train their embeddings to obtain two "views" of the data. This allows them to use more nuanced context around a word in a sentence. They find that the linear method performs as well as or better than existing non-linear methods for training word embeddings, the method works particularly well for rare words, and that adding in extra contextual information can help

disambiguate word meanings.

Overview of CCA in Pivot Analysis

Rather than use left and right contexts for words, we will scale our text based on in-sample word co occurrences and weighted out of sample word embeddings. This maps word co-occurrences and word counts to the same underlying space. The weights help us reduce the dimensionality of our text summaries and they are the primary workhorse, while the addition of out-of-sample word embeddings helps stabilize the output in small data sets.

For example, in our data, the word "government" is often accompanied by the words "intervention", "regulation", and "interference". We probably do not need to estimate that these words have subtly different meanings and trying to do so would rely on very noisy data. But we do care that a large cluster of people uses the word government, along with other words that reiterate its broad meaning. Our method focuses on scaling the word government and drags its accompanying words along with the scaling. Table 1 highlights this emphasis.

Standard text scaling	Pivot analysis
government **intervention**	**government** intervention
government **interference**	**government** interference
government **regulation**	**government** regulation

Table 1: Pivot analysis upweights common words relative to more rare words. It does this in a way that allows us to simultaneously estimate semantic locations for common and rare words, as well as bring in small amounts of data from out-of-sample sources. Its focus on common words should help us distill more low-dimensional and representative summaries from the open-ended survey data. If we consider variation in the rare words, they can account for a lot of variation in the data when we add their variance together and this complicates the compression of word usage onto a small number of dimensions.

The approach is similar to methods like the ridge regression (Hoerl and Kennard, 1970) and Lasso (Tibshirani, 1996). These methods reduce over-fitting by shrinking coefficients in linear regressions closer to 0, and perform well when there are a large number of correlated variables that measure the same underlying information. The amount of shrinkage over the variables is closely

related to their variance contribution in an orthogonal transformation of the data (Hastie, Tibshirani and Friedman, 2001). Variables that account for more variation in the data have coefficients that are shrunk less than ones accounting for little variation.

In our CCA, we are shrinking how much rare words contribute to the text scaling, in addition to a regularization like one in a ridge regression.[12] Beyond our specific interest in unsophisticated speakers, this reduction matters because rare words can introduce noise to our compression – similar to increasing R squared in a linear regression by introducing a large number of random variables. Unlike Lasso and ridge, however, the CCA still assigns coefficients to all words without shrinkage because it estimates two sets of coefficients: one set with shrinkage, which we use as a keyword metric, and one set without, which we use to score documents.

Although we weight output toward common words, our specific setup for the CCA and the structure of text data limit how much very common words contribute to our scaling, in a way similar to tf-idf standardization.[13] The CCA throws out data that does not maximize correlation between two views of the data, especially after truncation, and there is an inverse relationship between a word's frequency and how exclusively a word occurs with other words. When co-occurrence information is spread among a variety of words (i.e. it is not exclusive to a cluster), the CCA struggles to maximize correlation between orthogonal co-occurrence vectors and frequencies.

To put this another way, we are able to find shared representations for the word government across our two views of the data when we can drag its accompanying words along orthogonally. There is enough uniqueness in the conditional word co-occurrence probabilities for the word government to separate those probabilities onto a polarized dimension that describes the variation in our data set – and we can do this to the extent that we recreate the word frequencies of the word

[12] This regularization only forces the CCA to behave like existing text scaling methods (i.e. PCA and related approaches). The weighting is the key shrinkage in pivot analysis.

[13] tf-idf is a commonly used standardization in text analysis. It is word frequency multipled by inverse document frequency. Word frequency is often just the number of times a word appears in a document. Inverse document frequency (IDF) (Spärck Jones, 1972) quantifies how specific a word is in an entire corpus and it penalizes words that appear in many documents.

government with a unique and separable set of its co-occurrences. Other common words, such as the word 'time', are associated with too many different words to place them on a unique top dimension, so we do not pivot our low-dimensional scaling toward them.

Input

INPUT DATA	
M	Term-document matrix (in-sample data)
W	Word embedding matrix (out-of-sample data)
k	Regularization scalar - for *l2* norm
b	Tuning scalar - element-wise power, upweights common words
a	Tuning scalar - element-wise power, upweights word embeddings
I	Identity matrix
DERIVED DATA	
G	Word co-occurrence matrix - $M^\top M$
D_g	Diagonal of G matrix
D_g^{-1}	One divided by elements of D_g - this will divide the rows or columns of a matrix by elements of D_g
X	Row standardized word co-occurrence matrix - $D_g^{-1}G$ - left input of CCA - in-sample data
$\bar{Y}$	Word embedding matrix with weights - $\bar{Y} = D_g^b W^{\circ a}$ - right input of CCA - out-of-sample data b is a power for the vector D_g $\circ a$ is an element-wise/Hadamard power for the matrix W
σ	Leading eigenvalue of $X^\top X$ - for *l2* regularization
P_j	Column means of X - for evaluating tuning only
P_i	Row means of X - for evaluating tuning only
c	The soft, scalar cutoff for the keywords - for evaluating tuning only
C	Covariance matrix - C_{xy} is the covariance matrix of X and Y
COEFFICIENT AND OUTPUT DATA	
ϕ	Singular vector - ϕ_x is a left singular vector and ϕ_y is a right singular vector
ϕ^{proj}	Projection - ϕ_x^{proj} projection from X to shared space with Y, ϕ_y^{proj} projection from Y
ϕ_x^{fin}	Word scores - projections/coefficients with correction
ϕ_y^{proj}	Pivot scores - basis of keyword metrics using the Euclidean norm of the scores $\|\phi_y^{proj}\|$
$M\phi^{fin}$	Document scores

Table 2: This is a reference table for the notation used below.

In our CCA, one side of the input will be our in-sample data, X, that is the word co-occurrence matrix row divided by its diagonal:

$$X = D_g^{-1} G \tag{3}$$

where G is the word co-occurrence matrix and D_g^{-1} is 1 divided by Gs diagonal. For clarity, $G = M^\top M$, where M is the term document matrix. The term-document matrix M is a matrix with rows for each document and columns for each word. The value in each element is the number of times a word occurs in a specific document.

1. Standardize word co-occurrences G with diagonal D_g:	$X = D_g^{-1}G; \quad G = M^\top M$
2. Weight out-of-sample data W by word counts:	$Y = D_g^b W^{\circ a}$
2b. (optional) Predict usage with knowledge embeddings: (recommended) Whiten embeddings	$W = W_{Wik} CCA(W_{Wik}, W_{Twi})_{left}$
3. Run CCA between X and Y with regularization k:	$\max_{\phi_x,\phi_y} \frac{\phi_x^\top C_{xy} \phi_y}{\sqrt{\phi_x^\top (C_{xx}+k\sigma I)\phi_x}\sqrt{\phi_y^\top C_{yy}\phi_y}}$.
3b. Induce pivots with b such that:	$\frac{1}{e^{-\lambda}+1} \propto \|\phi_y^{proj}\|; \quad \lambda = 2b\left(\ln\left(\frac{P_j}{P_i}\right) - c\right)$ if $\ln\left(\frac{P_j}{P_i}\right) < 0$ then $\frac{1}{e^{-\lambda}+1} \to 0$ $\max\left(\left\|\phi_{y_n}^{proj}\right\|\right) \propto \ln\left(\frac{P_j}{P_i}^b + 1\right) \to$ rectifier
4. Correct for pivots ϕ_x^{fin}:	$\frac{\phi_x^{proj}}{\|\phi_y^{proj}\|+1} = \phi_x^{fin}$
5. Apply projections to term-document matrix M:	$M\phi_x^{fin}$

Table 3: *Summary of pivot analysis.* Notation for this table is introduced in Table 2. Projections are estimated using singular value decomposition. Larger bs induce the desired "pivot" behavior (i.e. upweight common words) and larger (odd) a increases the effect of out-of-sample data (i.e. upweight word embeddings). We standardize the final document scores based on the number of words in a document.

This matrix is the starting point of our scaling. A principal component analysis of this matrix would return results similar to previous methods. For example, $D_g^{-1}G$ is closely related to the factorized matrices in topic models (Roberts, Stewart and Tingley, 2016) and existing text scaling methods, including LSA (Deerwester et al., 1990) and correspondence analysis (Lowe, 2007; Bonica, 2014).

This particular matrix has worked well on sparse and heavily skewed data (Bond and Messing, 2015). It is especially useful because it provides conditional word co-occurrence probabilities. In our scaling, we want to optimize a prediction about what sets of words tend to go together and these probabilities provide the necessary information for that optimization. These probabilities also retain frequency information that we can use to pivot our output toward moderately common to very common words that tend to appear within a limited set of arguments (i.e. a clustered set

of accompanying words). Although it is possible to weight the chi-square statistic matrix used in correspondence analysis, that matrix is not correlated with word counts in a way that can be used for pivoting, since the co-occurrences and counts are explicitly decorrelated.

Prior to calculating the word co-occurrence matrix, we only remove words that appear in the SMART stopword list[14] or that appear only once in the corpus. In this pre-processing, we rely on defaults in the "stm" R package (Roberts, Stewart and Tingley, 2016), the most commonly used software for text analysis in political science. We do not stem the text, however, because our word embedding data is not stemmed.

For our other input to the CCA, the out-of-sample data, we use a pre-trained word embedding matrix provided online by Pennington et al. (Pennington, Socher and Manning, 2014).[15] This word embedding matrix is essentially output from text scaling run on a massive amount of data from Wikipedia and/or Twitter. It contains the semantic location of a word in the entire English language across 200 to 300 numeric columns in each row of the matrix. We will denote the word embeddings using W. We use these embeddings because they are easy to access and are trained on much more data than we have in the open-ended survey responses. The out-of-sample word embeddings simply give us more data to work with as we estimate locations of words. At the same time, our method is ultimately very closely tied to the in-sample data, so this added data mostly smooths our final estimates (unless we tune its hyperparameter to very high levels).[16]

[14] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

[15] We run an additional CCA between two versions of the GloVe embeddings, Twitter and Wikipedia, to remove context specific idiosyncracies in the data sets. This steps whitens our input data.

[16] Smoothing here means that we bring in very little information from the out-of-sample embeddings, but that we can infer a relatively uncommon word's meaning based on a combination of its location in the word embeddings and its location relative to other words in our own corpus. Very high levels of our tuning parameters for this behavior will bring the in-sample data closer to the out-of-sample data, as we will discuss later in this paper. The appropriate amount of this tuning is currently subjective, however, so we leave evaluation of high levels of the tuning parameter to future work. We will be able to provide an objective measure of its effect.

Inducing Pivots

We require a few adjustments to the ordinary CCA and its input data to produce extremely low dimensional behavior.

First, CCA is scale invariant, but we want it to respect the variance structure of our in-sample word co-occurrences. Because of the inverted covariance matrix for X, C_{xx}, CCA does not penalize the use of low variance dimensions when predicting word counts. To keep some or most of the same structure, we add a regularization to C_{xx}, k, using multiples of the leading eigenvalue of that matrix, σ.

$$\max_{\phi_x,\phi_y} \frac{\phi_x^\top C_{xy} \phi_y}{\sqrt{\phi_x^\top \left(C_{xx} + k\sigma I\right) \phi_x} \sqrt{\phi_y^\top C_{yy} \phi_y}} \tag{4}$$

Put simply, this keeps our output close existing scaling methods. It is perhaps helpful here to think of X as the components of a principal component analysis. This regularization forces the CCA to prefer the top dimensions of the principal components over lower dimensions. To fully respect the variance structure of the original data, we can simply replace the inverted covariance matrix with an identity matrix. In our data, the leading eigenvalue scales the pivots' output to unit vectors.

A smaller regularization than the identity matrix is sometimes useful because it identifies tightly clustered phrases. In our case, this is useful because tightly clustered phrases suggest coordination on a politician's talking point. For example, clustered phrases in our data include "prefer single payer" and "takes freedom away".[17]

Next, the CCA does not weight common words more than rare ones when optimizing correlations from our in-sample data to the word embeddings. Without this, we have no pivots (i.e. no sparse, shared representations for common words across in-sample co-occurrences and out-of-

[17] This behavior is not always desirable. For example, in social media platforms like Twitter, people can copy each others' language directly. With artificially low overlap between retweets and other related language (i.e. limited semantic context), the distance between copied language and the rest of the corpus will be exaggerated.

sample data). To add this behavior, we multiply the word embeddings by the word counts. We also add an element-wise power (i.e. Hadamard power) to allow us to adjust the effect of the out-of-sample data on our output:

$$Y = D_g^b W^{\circ a} \tag{5}$$

where b sets the weighting level and a, an odd integer, controls the amount of smoothing inserted from out-of-sample data. $a = 1$ provides very little out-of-sample information and is the only value for this parameter we will consider in depth here. To explain the role of out-of-sample data more intuitively here, our weights wash out the effects of rare words and the tuning parameter a adds information for moderately common/not too rare words back in based on the out-of-sample word embeddings.

Tuning Pivots

The above formulas are sufficient to implement the CCA in pivot analysis. From here, we explain how to tune the input parameters, as well as how to recover keywords and document scores from the output.

Given the exponential, or inverse-rank frequency, distribution of word counts, we induce an activation function for weighting common words when $b > 0$. To induce *pivots*, we set b to a level high enough to scale only the common words. With a sufficiently large b, we hope to recover a 1 to 1 relationship between the two views of our data for only our common words and an overall representation that has been reoriented toward common words. b that is not sufficiently large will produce a sigmoid relationship for scores of common words' between the two views of our data.

Simply raising b until the singular value decomposition can no longer be estimated works in practice. It is potentially helpful to describe the activation function our weighting produces, however.

The weighting and activation on a *single* dimension is a softplus function, with full activation approximately $\ln\left(\frac{P_j}{P_i}^b + 1\right)$, where P_i is the row mean of the symmetric matrix $D_g^{-1}G$ and P_j is the column mean of $D_g^{-1}G$ (i.e. the input matrix X). Because of this, large b leads to a smooth approximation to a rectifier, and words with $\ln\left(\frac{P_j}{P_i}\right) < 0$ have near 0 weight as pivots.[18] Whether a word is activated in a single dimension is then driven by:

$$\ln\left(\frac{P_j}{P_i}\right) >> 0 \tag{6}$$

As an example, the word "government" has a column mean in our data of 0.15 and a row mean of 0.003. Roughly, this means that if a person says any random word, then the chance of them also saying the word "government" is 15%. Similarly, if a person says "government", their chance of saying a given random word is 0.3%. When the ratio of these probabilities is large that word is a pivot word.

Words that exceed this threshold have more polarized word scores if they tend to occur with a highly specific set of terms on a dimension. Most often these highly specific, common words are parts of very tightly clustered phrases, such as 'universal access' or 'children stay on parents insurance'. Words that exceed the threshold but are less specific can still be activated on a dimension to a more limited extent if they are very common, especially given our regularization k.

At the same time, we observe that activation over *all* dimensions (in text data) is approximately the logistic function for the Euclidean norm:[19]

$$\frac{1}{e^{-\lambda}+1} \propto \|\phi_y^{proj}\| \tag{7}$$

where λ equals $2b\left(\ln\left(\frac{P_j}{P_i}\right) - c\right)$. c is a feature of the data. In our data, c is approximately 0.9

[18] The pivot scores are related to the hyperbolic functions. Large b induces semantic dilation around common words.

[19] The word embeddings will affect this functional form over all dimensions, even though they do not affect word scores in low dimensions. Having pivot scores equal to 0 for rare words is more important than the precise functional form.

and around 8% of words exceed that threshold.[20] The form of this logistic function in a given data set is affected by the specific inverse relationship between term frequency and specificity, and the function is not clearly logistic when the inverse relationship does not exist (e.g. in non text data such as campaign contributions).

Close approximation to the above logistic function gives us the appropriate tuning for the pivot analysis method. We show convergence to that functional form around the constant c in the appendix Figure 10.

To provide somewhat more intuition for that tuning in words, our hyperparameters alter the weighting function in the following ways. Raising the power of D_g^b in the word embedding matrix multiplication $D_g^b W^{\circ a}$ produces steeper separation at c, while greater (odd) a will produce noisier separation at c – where "noise" is the added information from out-of-sample word embeddings.[21]

Steeper separation at c is a sharper separation between pivot words and the rest of the data. Without this separation and a 1 to 1 relationship between pivot scores and overall scores, we no longer have our keyword metric. Greater odd a allows us to add in some information for moderately common words based on out-of-sample data. Very common words and very rare words are largely unaffected by it, except when a is tuned to very high levels.

We visualize the effects of tuning b in Figure 10 in the appendix and visualize a to increase the effects of word embeddings in Figure 11. Tuning higher a smooths the pivot transition for $\ln\left(\frac{P_j}{P_i}\right) >> 0$ and this can be visualized over all dimensions at a transition $\ln\left(\frac{P_j}{P_i}\right) = c$.

Keywords and coefficient adjustment

Once we induce pivot behavior with large b, we will achieve high correlations between the two sets of data – but only for common words. Because of this, the ϕ_x^{proj} scores provide the rescaled word scores that we multiply by the term-document matrix to produce document scores, while the

[20] c's location affects high dimensions of the output, but has little effect on low dimensions.

[21] Note that we will not achieve a balanced looking sigmoid function for extraordinarily skewed text data.

ϕ_Y^{proj} scores show pivot scores that anchored the overall representations and that we can use as a keyword metric.

We multiply ϕ_y^{proj} by the corresponding canonical correlation (i.e. the corresponding eigenvalue) to place the pivot scores on the same scale as the overall word scores. ϕ_x^{proj} and ϕ_Y^{proj} will then be similar to equivalent for the pivot words, while relatively rare words in ϕ_Y^{proj} will remain close to zero.

Before applying the word scores back to the documents, we adjust the overall word score projections according to:

$$\frac{\phi_x^{proj}}{\|\phi_y^{proj}\|+1} = \phi_x^{fin} \tag{8}$$

where $\|\phi_y^{proj}\|$ is the Euclidean norm of the pivot scores, and measures the degree to which a word is a pivot word. The value is standardized so that the largest value is 1. This halves the size of the word scores for pivot words only and corrects for the specific non-linearity that our weighting produces. We visualize this adjustment in Figure 1 and Figure 7.

To explain this more intuitively, our weighting lets us find dimensions based on common words, but the weighting then scores common words too far away from the center once we've defined our dimensions around them. This adjustment moves the common words back toward the center so that we don't score documents very strongly on one dimension if they simply use the words 'health care'. We require that the documents have repeated and consistent or highly specific word usage to score highly on a dimension.

Our last step is to return document scores based on our word location estimates. To do this, we simply multiply the projection, ϕ_x^{fin}, (i.e. the coefficients) by the original term document matrix M, then adjust these document scores for the total number of words used in a document.[22]

[22] We divide the scores by the number of words in a document to a power between 0.5 (more words add more information at a rate of square root of n) and 1 (more words do not add more information). In our data, longer responses typically use more complete sentences without adding many more substantive words. A value less than 1 accounts for the more grammatical responses. We use 0.75 and recommend this value in general. The choice has little

Related Work

Both our common word estimation and domain adaptation is accomplished in a way similar to structural correspondence learning (Blitzer, Foster and Kakade, 2011). Blitzer et al. identify words that are common and have the same usage in two contexts, and use these words as "pivots" to adapt pre-trained data to a new corpus.

We also use pivots, but we only use word counts to identify keywords, rather than using supervision on labeled data. This assumes that very common words are unlikely to be jargon. The method also differs because it is very strongly tied to in-sample data and focuses on orienting the representations toward word counts. The out-of-sample data almost exclusively smooths the final estimates, and tuning the method to produce estimates closer to the out-of-sample data provides only small predictive improvements.

Our focus on keywords means that we prioritize estimating locations for a small proportion of words, rather than many rare words. Matrix factorization techniques used in computer science tend to do the opposite of this. For example, word2vec (Mikolov et al., 2013), SVD with PPMI standardization (Levy and Goldberg, 2014), and GloVe (Pennington, Socher and Manning, 2014) discriminate between common and rare words to obtain precise estimates for a full vocabulary. Otherwise, these models are closely related to pivot analysis.

Of course, orienting around common words probably ignores subtleties and idiosyncracies in sophisticated text. However, this relative ignorance allows us, we hope, to produce interpretable representations. Prior work has found a trade-off between predictive accuracy and interpretability (Chang et al., 2009). Further, in our case, we should be able to achieve interpretable dimensions without much loss in accuracy. Our outcome of interest is a single dimension of favorability toward a public policy and most of the justifications on it are short and simple.

effect on the results, however.

Application to Open-Ended Surveys on the ACA

We now apply our method to the data on the Affordable Care Act. We leave the hyperparameter a at 1 so that the word embeddings, out-of-sample data, only provide a small amount of smoothing to the estimates. We also leave the regularization k at 1, the leading eigenvalue of the in-sample word co-occurrences, so that clusters of speech have somewhat greater weight.

Next, tuning b to 2 is sufficient to induce pivoting. As a reminder, the pivots are words that are moderately to very common and that are also somewhat specific. We use them as axes on which to pivot our output away from rare words and toward common words.

In inducing 'pivots', a sufficiently large b minimizes the effects of rare words to the point that words with co-occurrence probabilities $\ln\left(\frac{P_j}{P_i}\right)$ less than 0 receive little to no weight in our reorientation toward common words, as measured by the right singular vectors of our decomposition (our pivot scores). The specific functional form of this tuning is a linear/symmetric relationship between the left (overall scores) and right (pivot scores) singular vectors of our decomposition for words with co-occurrence probabilities $\ln\left(\frac{P_j}{P_i}\right)$ much larger than 0.[23]

Beyond orienting our scaling output toward common words, this tuning gives us keyword scores that accurately reflect the polarization in words' scores in our overall estimates. We visualize the various adjustments to these hyperparameters in Figures 10, 11, and 12 in the appendix.[24]

We first show the keywords from the top 2 dimensions of our output in Table 9. The keywords here are a word's ϕ_y^{proj} on a dimension multiplied by its total activation (unit standardized $\|\phi_y^{proj}\|$). We named the dimensions ourselves.

These keywords appear to be highly informative. They pick up both specific components of ACA policy and broad opinions on it.

[23] In practice, it is fine to simply tune b with increasing positive integers until the matrix is computational singular, then subtract one from the computationally singular b.

[24] We also pre-process the word embeddings, step 2b in the Table 9, using no regularization because using only the Wikipedia embeddings prevents the Euclidean norm of the pivot scores from converging around c, as happens using only in-sample data. This affects visualization of the Euclidean norm, but does not affect the low dimensional representations.

In Table 5, we show example responses that scored highly and uniquely on one of the top two dimensions. Although not perfect, of course, these examples suggest that the method works well on the document level, even though we do not incorporate document information into our CCA.

Keywords			
Dimension 1		*Dimension 2*	
"patient protection"	"role of government"	"universal access"	"personal cost"
Pro	Anti	Pro	Anti
preexisting	government	universal	premiums
conditions	unconstitutional	health	deductibles
parents	run	access	money
children	involved	everyone	working
stay	us	care	pay
pre	economy	step	even
young	direction	direction	income
age	peoples	needs	middle
opportunity	business	americans	high
coverage	much	affordable	medicare
helps	control	provide	going
year	dont	right	low
now	medicine	preexisting	less
able	president	single	fine
insurance	end	every	taxes
can	everything	conditions	elderly
uninsured	taxes	country	higher
allows	socialized	provides	everything
condition	socialism	issues	doctors
still	every	system	put

Table 4: *Output summary.* This table shows the keywords in the top two dimensions. We identify keywords by multiplying ϕ_y^{proj} by its unit vector Euclidean norm. The dimensions and keyword identification are unsupervised. We added the pro and con labels after linking the scores to the preceding closed-ended survey response.

Right - Role of Government - Dimension 1 (Anti)
because i think that the government has no business doing that. it should be left to private companies. a solution should be made so that the government doesn't take it over.
its because the government is going to take over
i think was well conceived but poorly executed
too much control is given to the government
its not up the government to tell us what to buy, it think they should throw the whole thing in the garbage.
too much politics involved in what should be private
Left - Patient Protection - Dimension 1 (Pro)
the coordination and the ability for more people to get coverage and the rules
gives students the ability maintain insurance on parents coverage
coverage for those people who do not have insurance
this will hopefully help my children receive insurance and go to the doctor more frequently.
my daughter is able to get insurance that she was not able to get before for her and her family
21 year olds are now covered also
Right - Personal Cost - Dimension 2 (Anti)
isn't going to benefit anybody and he doesn't listen to the people
going to cost the middle class people a lot more money, which is not fair
friend got it to avoid fine & had to use it for gallbladder surgery & had a bill balance to pay over $6,000
some of the insurances are high for the middle class.
i think there are areas of patients that will be covered that they don't do anything to earn it; they have no value in it, they should have a co-pay and do something to earn it–something that they have to pay for coverage
how are we going to pay for it
Left - Universal Access - Dimension 2 (Pro)
something has to be done to improve our health care in america - it is not perfect but it is a start.
everyone should have access to healthcare.
because everybody deserves to have good health care and good treatments they deserve to get the right medicine and not the generic kind and because nobody deserves to struggle to keep them self healthy
feels the government should not be a health care provider
i think they're trying to get more health care for the underprivileged
it going to be good for everybody; we're going to have better health insurance and better health care in general

Table 5: *Examples of open-ended responses.* This table shows a random sample of responses that score relatively highly (greater than 1 standard deviation) on one of the top 2 dimensions and low (lower than 0.25 standard deviation) on the other. These are responses that we score as relatively unambiguous. A total of 2,000 responses fit these criteria.

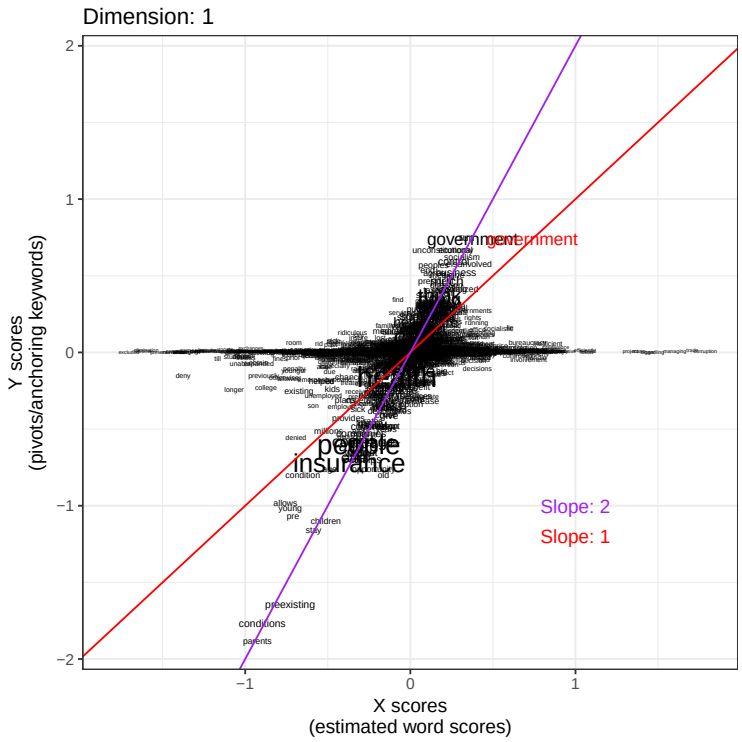


Figure 1: *Relationship between pivot scores and overall scores.* This figure shows the relationship between the pivot scores and the overall scores on a single dimension (dimension 1). For large b, all words receive overall scores (x axis) and only common words receive pivot scores (y axis). This leads to a one to one relationship between common words' pivot scores and overall scores. We adjust this to a two to one relationship to correct for non-linearity – shown in Figure 7 in the appendix.

Visualizing Pivot Scores

Because pivots are central to our method, we next visualize the distributions of pivot scores in our data. These are illustrations of the equations we introduced in the methods section.

Figure 1 visualizes the relationships between pivot words and all words for the top two dimensions of the output. The x axis is the word's score and the y axis is the word's pivot score. The words with scores away from 0 on the y axis are keywords in the dimension, while scores away from 0 on the x axis are accompanying words in a dimension. We described this output of the method in the 'keyword and coefficient adjustment' section above.

In this figure, the red line has a slope of 1 after we multiply the keyword CCA projections by

the canonical correlation. The purple line has a slope of 2 after we apply the adjustment $\frac{\phi_x^{proj}}{\|\phi_y^{proj}\|+1} = \phi_x^{fin}$). CCA maximizes the total weight along the red line, and we map the pivots to their word scores based on the purple line. Achieving linear slopes here after tuning b to a high level gives us the keyword scores that accurately reflect a word's polarization in our final document scores.

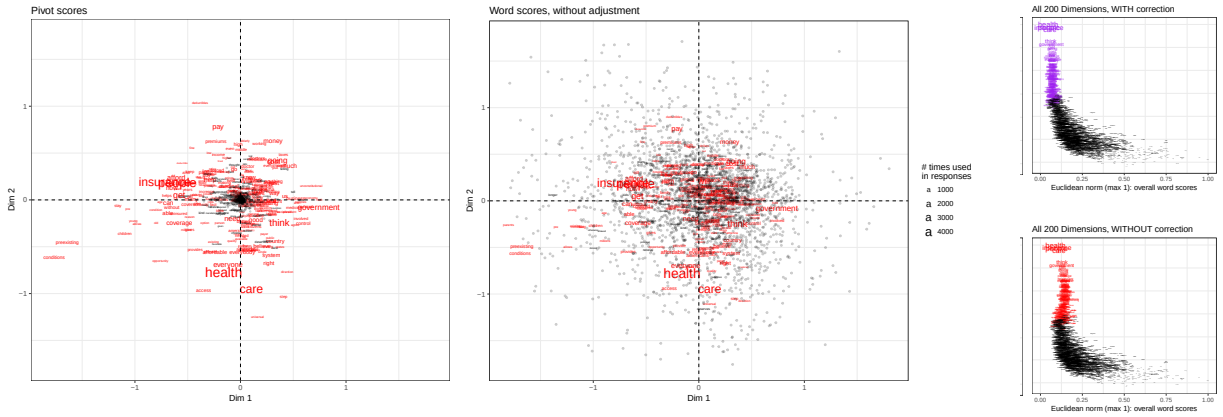


Figure 2: *Additional explanation of Figure 1.* This figure shows how the pivot scores and the overall word scores are related, along with the effect of the non-linearity adjustment on the overall scores. It shows the same information as Figure 1. In the figures, the color red indicates a pivot word without a linearity adjustment and purple shows the score with an adjustment. The left two panels show the pivot scores (far left) and word scores (middle left). The x axis for both panels is the score on dimension 1 and the y axis is the score on dimension 2. The pivot words have close to the same coefficients in both views of the output. The far right panels show how the adjustment $\frac{\phi_x^{proj}}{\|\phi_y^{proj}\|+1} = \phi_x^{fin}$ affects the Euclidean norm of the overall scores. The x axis is the Euclidean norm and the y axis is $\ln\left(\frac{P_j}{P_i}\right)$. We show this adjustment for the top two dimensions of the output in Figures 7 and 8 in the appendix.

The left two panels of Figure 2 also shows that the pivot words have the same scores in both views (pivot and overall) of the data, and that the rest of the words appear in only the overall scaling (i.e. they are oriented around common words). This shows the same linear relationship for pivots and overall scores that we show in Figure 1. The two vertical panels on the right side of this figure again show the effect of the linearity adjustment $\frac{\phi_x^{proj}}{\|\phi_y^{proj}\|+1} = \phi_x^{fin}$ on the word scores.

Figure 3 shows the pivot scores for the 1st dimension, the 2nd dimension, and the Euclidean norm on all dimensions. The Euclidean norm in the far right panel is standardized to a maximum

of 1. The x axis in these panels is the pivot score and the y axis is $\ln\left(\frac{P_j}{P_i}\right)$. For the Euclidean norm, the function is logistic, while for the individual dimensions it is bounded by a softplus function. On the x axis, a value close to one means that a word is a pivot word, while a value close to 0 means that a word is not a pivot.

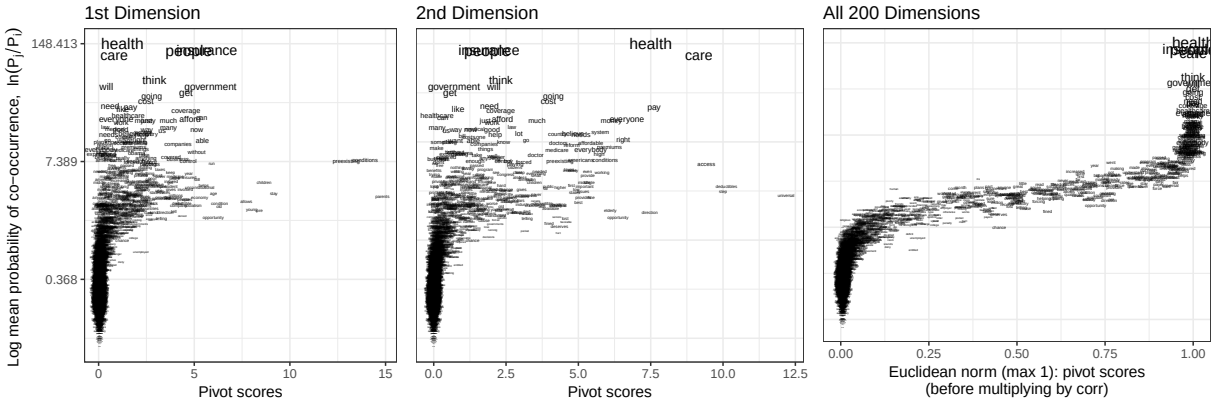


Figure 3: *Keyword weights.* These figures show absolute values for pivot scores. The x axis is a word's total weight in a given number of dimensions (noted in the figure title) and the y axis is $\ln\left(\frac{P_j}{P_i}\right)$. The x axis in the far right panel shows the values for $\|\phi_y^{proj}\|$. The softmax here is induced by the structure of the data, and the exponent in the word count times word embedding portion of the formula controls its form.

This figure shows the performance of the method that resembles tf-idf standardization. A word that lags above and to the left of the final line is typically a common but ambiguous word. A person who starts with that word can follow up in many different semantic directions. In the ACA data, "health care" is a common but ambiguous starting point in an open ended survey response.

Predictive Evaluation

We evaluate the dimensions of the pivot analysis output by predicting responses to the closed ended question in the surveys on favorability toward the Affordable Care Act. This is a straight forward mapping to the closed-ended response. If the predictions are sufficiently low dimensional, we will

be able to visualize dimensions of attitudes toward the Affordable Care Act in a biplot.[25]

In the evaluation, we predict the closed ended response on favorability toward the Affordable Care Act using a Lasso (Tibshirani, 1996). This is a penalized regression that selects the best independent predictors of ACA attitudes from our text dimensions.[26] To compare the coefficients, we first scale each dimension so that each variable in the regression has the same variance. The coefficients from this model are then the additive dimensions of attitudes toward the Affordable Care Act, and the size of the coefficient reflects a dimension's importance in prediction.

To help visualize these results, we run k means clustering on the representations – with the dimensions weighted by their Lasso coefficient. If we are *not* able to visualize the dimensions in two dimensions, then colors from the k means clustering will appear randomly distributed in the biplot.

Figure 4 shows the biplot for the top two dimensions of word scores from our method. It shows the unadjusted scores for all words (ϕ_x^{proj}) so that common words remain easy to see, since the adjustment brings common words closer to the center. The x axis is the first dimension of the scores and the y axis is the second dimension. The size of the words is the number of occurrences in the data set. To orient the reader, an individual who says "government" in the responses usually does not like the government.

The figure shows clear separation in colors based on the prediction weighted clustering, meaning that we can adequately visualize dimensions of ACA attitudes in only two dimensions. Higher dimensions do not substantively affect the clusters.

[25] The prediction based approach allows us to get around complicated survey based evaluations of our scores. In evaluations that test whether people are able to group words in the same way as a model, for example, we would likely perform better if we estimate a high number of dimensions on which clusters of words are tightly packed. These tightly packed clusters would be easy for non-experts to associate. However, the large number of clusters would no longer be in line with our goal of producing latent and low dimensional representations of attitudes. The low dimensional representations should link loosely associated arguments.

[26] Our Lasso uses the defaults in the 'glmnet' package (Friedman, Hastie and Tibshirani, 2010). The glmnet Lasso function selects the regularization level using smallest cross-validated error.

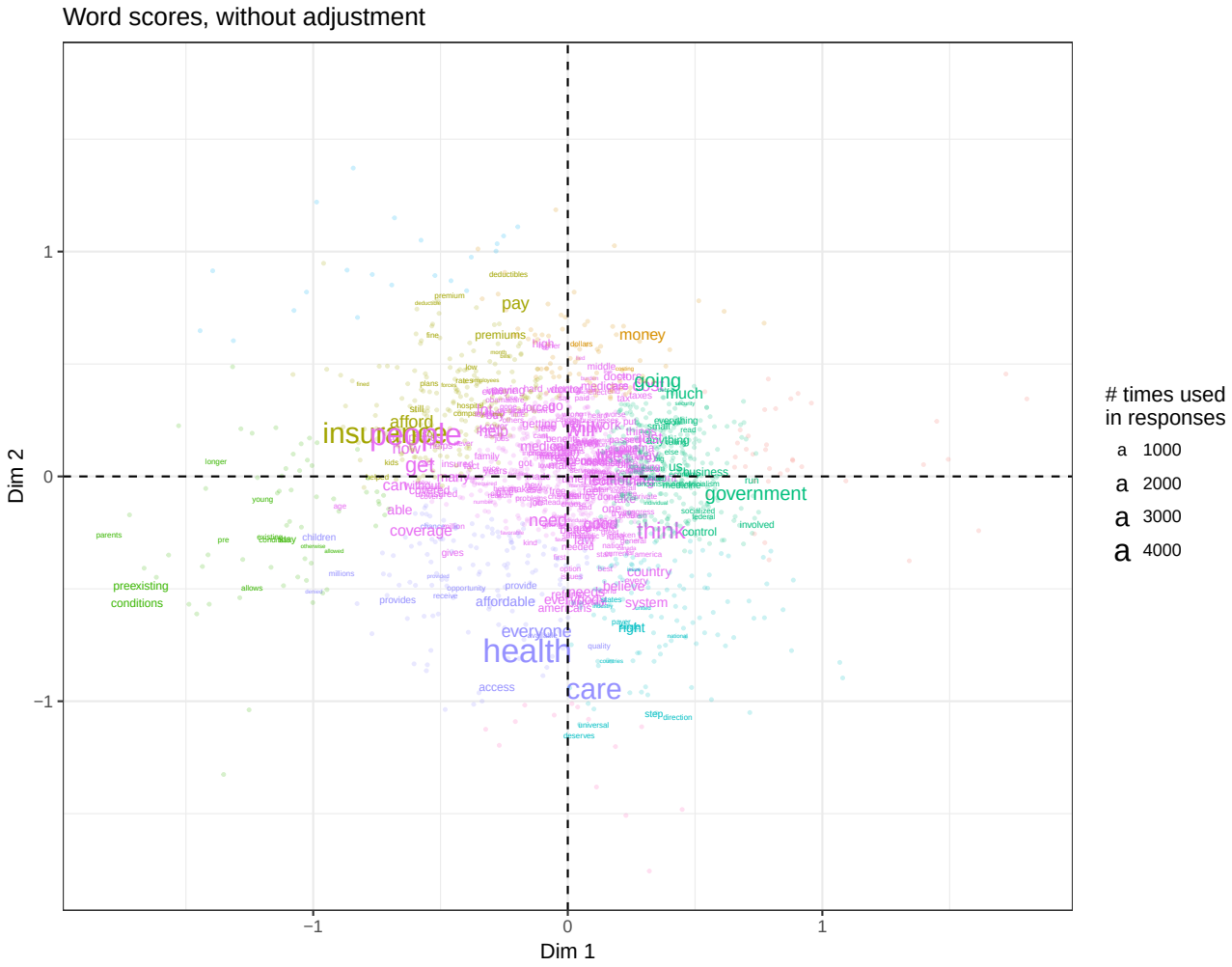


Figure 4: *Word scores.* This panel shows the word scores for the first two dimensions. We show text for words with frequency greater than 50 and points for words stated 5 to 50 times. Note that the final word scores have an adjustment that moves the pivot words closer to 0 (see Figures 1 and 8). This adjustment moves the pivot words close to the inner edge of their associated word clusters and corrects for non-linearity in the pivot scores.

Comparisons to other continuous methods

In Figure 5, we compare the Lasso on our data to other methods. In the comparison in the left panel, we again rescaled the word scores from each method to equal variance vectors and then estimated the Lasso 100 times for each method. We trained the Lassos on a training set (4/5 of the data) and evaluated the predictions on a test set (1/5 of the data). The test and training set were the

same for all methods and iterations.[27]

For all but GloVe and the topic models, the x axis denotes a regression on the first n dimensions of a method's output. For GloVe and the topic models, we asked the method to return the number of dimensions then predicted using that output. The topic model results are shown at the number of topics minus 1. We show the area under the ROC curve for the first 10 regression models from each method.

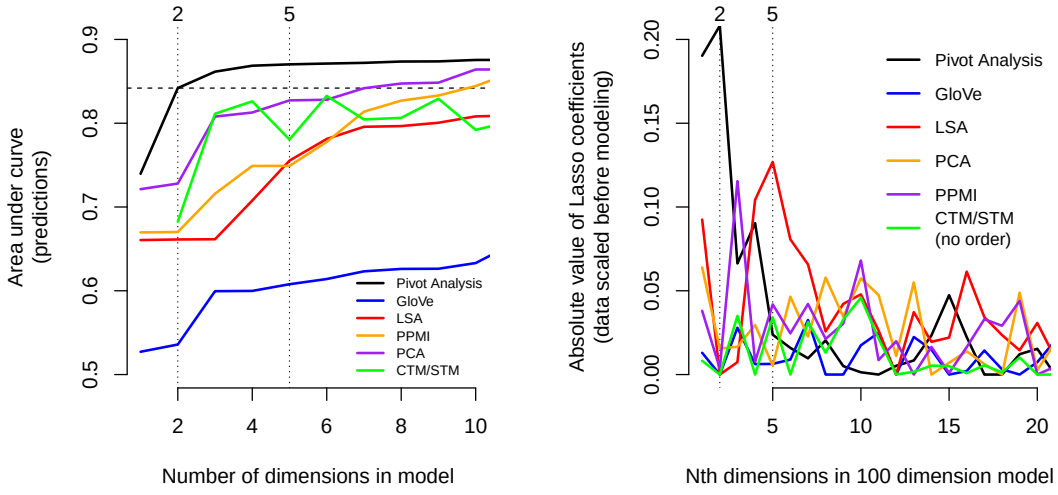


Figure 5: *Dimensionality of other methods.* The left panel of this table shows area under the ROC curve for 100 penalized regressions predicting ACA favorability from each method's output. It shows that pivot analysis achieves high AUC in low dimensions, while other methods converge to high accuracy more slowly. The right panel shows the absolute value of coefficients from a 100 dimensional penalized regression predicting ACA favorability from each method's output. Predictive dimensions are concentrated in the first dimensions of pivot analysis, while predictive dimensions are spread across the output from other methods. The dimensionality for topic models here is the number of topics minus 1.

Figure 5 shows that the Lasso is using substantially lower dimensional information in pivot analysis than in all other comparison methods. Pivot analysis is capable of condensing much of the

[27] The comparisons for 'LSA' and 'PCA' are close matches to methods used in political science, such as Wordfish and correspondence analysis (Lowe, 2016). These methods happen to perform worse on our data than their typical performance, however, possibly due to some common words appearing with no other words. This lack in co-occurrences is washed out in our method, but is picked up as important variation in these other methods.

information in the text responses into a small number of variables. Given this low dimensionality, a researcher will not need to select one out of many possible variables for later analyses.

In the right panel, we show the absolute value of coefficients from a 100 dimensional Lasso predicting ACA favorability in our data. This shows that a Lasso chooses the first dimensions of pivot analysis's output, but chooses higher dimensions from other methods. A researcher that uses theses methods would need to justify their high dimensional choice in later analyses, while pivot analysis provides a hands off and useful ordering.

This substantial improvement in performance in low dimensions comes with very small cost. Pivot analysis performs only marginally worse than PCA and PPMI in 200 dimensions (pivot AUC: 0.914; PCA AUC: 0.916; PPMI AUC: 0.919).

Comparison to topic models

Topic models can recover similar dimensions, but perform worse on both low and high dimensional prediction in our data.

As shown in Figure 5, selecting the number of topics in a correlated topic model between 4 and 20 will all return dimensions that predict ACA attitudes at the same levels, and at lower predictive accuracy than the first 2 of our dimensions. Selecting the number of topics automatically using information criteria will give around 50 topics.

We show similar dimensions from a correlated topic model in Table 6, where we have run several models and chosen 4 topics so that the results look somewhat like our output.[28]

Of particular interest to our model, the table shows that pivot scores' keywords may be closer to the frequency and exclusivity metric (FREX) for keywords in a topic model (Bischof and Airoldi, 2012) than the highest probability metric for keywords. Since pivot analysis optimizes the keyword metric ϕ_y^{proj}, this table suggests that pivot analysis orients dimensions based on a combination of

[28] One option for using our method would be to first estimate the pivoted representations, then choose a number of topics that resembles that output. Explicitly linking topic models and discretization to our approach is beyond the scope of this paper, however.

*Keywords (**highest probability**) - TOPIC MODEL*				*Keywords (**FREX**) - TOPIC MODEL*			
Topic 1	Topic 2	Topic 3	Topic 4	Topic 1	Topic 2	Topic 3	Topic 4
"health insurance / cost"	"government / cost"	"benefit / cost"	"health care"	"health insurance / cost"	"government / cost"	"benefit / cost"	"health care"
No relationship	Anti	Anti	Pro	No relationship	Anti	Anti	Pro
people	government	think	health	afford	government	small	care
insurance	going	healthcare	care	coverage	going	fair	everyone
get	cost	work	will	now	much	hard	needs
afford	like	country	everyone	many	money	went	available
can	much	something	need	conditions	things	nation	sure
coverage	money	system	help	without	take	basically	elderly
now	just	go	lot	preexisting	medicare	rich	chance
pay	us	know	good	covered	anything	living	basic
many	right	expensive	believe	high	control	major	heath
law	want	buy	everybody	uninsured	away	real	accessible
medical	way	business	needs	insured	run	healthcare	harder
companies	costs	getting	affordable	class	nothing	hours	answer
conditions	one	doctor	make	cover	taxes	drug	effort
without	things	dont	reform	children	taking	started	need
preexisting	doctors	free	able	middle	choice	work	fortunate
premiums	take	needed	better	stay	long	self	hoping
access	medicare	benefits	americans	parents	step	benefits	improvement
covered	anything	keep	bill	existing	single	reach	provide
really	pay	making	time	age	else	seems	ensure
paying	well	made	get	allows	payer	supposed	senior

Table 6: *Keyword metrics from 4 topic correlated topic model.* This table suggests that our pivots resemble the frequency and exclusivity metric for keywords in topic models.

frequency and exclusivity.

Effects of Out-of-Sample Word Embeddings

This paper focuses on scaling text using very little out-of-sample data. It is possible to use more out-of-sample data, however, and tuning a to increase the effects of the out-of-sample word embeddings has small effects on our results.

For example, with $a = 3$ and all other parameters the same, we find that the top 2 dimensions better fit our category labels (Table 9 in the appendix). This suggests that the out-of-sample word embeddings make the results more general – that is, less specific to the Affordable Care Act. On the other hand, tuning the hyperparameter a to higher levels without also increasing b reduces our ability to predict stances on the Affordable Care Act. The dimensions begin to be too general to be useful.

In smaller data sets on political opinions, such as abortion, we have observed larger increases in low dimensional predictive accuracy when using higher a values.

Correlates with Other Survey Responses

In addition to the open-ended survey responses collected by the Kaiser Family Foundation and Pew Research Center between 2009 and 2016, we also have a smaller collection of responses to the same question in our own surveys. This survey has a larger variety of closed-ended responses and has these responses going back several years.

We applied the word representations to our nationally representative sample to see whether the top dimensions were correlated with vote choice and change in ACA attitude after controlling for partisanship. Partisanship is by far the best predictor of both vote choice and ACA attitudes – prior to the 2016 election, it was correlated with ACA attitudes at about 0.65. Finding that a variable is significantly correlated with a political outcome after including partisanship as a control is a high standard.

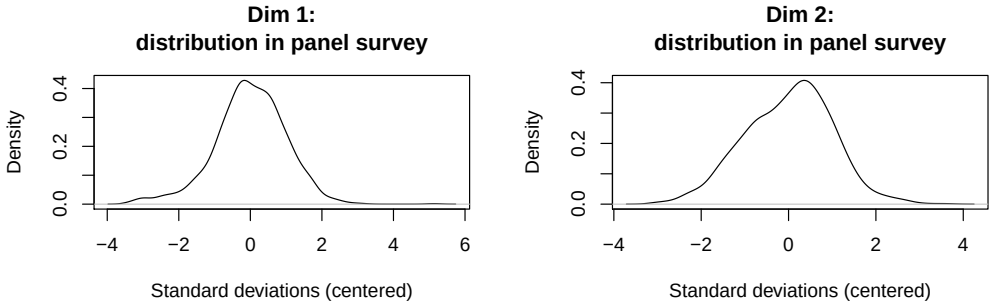
Table 7 shows that the top two dimensions of our output are correlated with 2016 vote choice, controlling for 2016 partisanship (model 2). The 2nd dimension but not the first is also associated with whether a Republican voted for Donald Trump instead of Republican establishment candidates (3). This suggests that pivot analysis picks up political cleavages both across and within American political parties.

The 2nd dimension is further associated with changes in ACA attitudes from 2012 to 2016 (1). People who talked about personal costs, rather than universal access, were more opposed to the law. This is an important change because major components of the ACA were only implemented after 2012. Text dimensions that were correlated with 2016 but not 2012 attitudes were more likely to be due to personal experiences with the law, or due to copying politicians' new, post implementation talking points.

In these models, we include as controls: education, income, partisanship in 2012 (1 only), partisanship in 2016 (2 and 3 only), Barack Obama feeling thermometer, gender, age and age squared, race/ethnicity. The figures above the table show the distribution of the document scores in our nationally representative panel survey of around 1 thousand people, training on all text

responses.

Of these variables, only a relative dislike of Barack Obama (not high feeling thermometer) and race/ethnicity (white) were associated with whether a Democratic respondent voted for Bernie Sanders instead of Hillary Clinton. Higher dimensions of the text responses were also unrelated to the Democratic primary votes.



	Dependent variable:		
ACA attitude: positive is more opposed	ACA 2016 vs ACA 2012	Donald Trump vs Hillary Clinton	Donald Trump vs Republican establishment
	(1)	(2)	(3)
Dim 1 (+ role of government vs − patient protection)	−0.05 t = −0.81	0.05*** t = 4.33	−0.01 t = −0.33
Dim 2 (+ personal cost vs − universal access)	0.15* t = 2.42	0.05*** t = 4.64	0.14*** t = 3.49
Observations	768	890	387
R^2	0.08	0.65	0.09
Note:			*p<0.05; **p<0.01; ***p<0.001

Table 7: *Between and within party correlates in small panel survey.* All independent variables are scaled so that a unit change correspondents to a one standard deviation change in the panel corpus. The dependent variables in the vote choice models are coded 1 if the respondent voted for the first candidate (Donald Trump) and coded -1 if the respondent voted for the second candidate(s). Republican "establishment" candidates are: Jeb Bush, Chris Christie, John Kasich, and Marco Rubio.

ACA Attitudes Over Time

In Figure 6, we show changes in the mean of each of the top two dimensions over time based on the Kaiser Family Foundation and Pew Research data 2009 through 2015.[29] We plot separate time series for respondents who selected favorable or unfavorable in the preceding closed-ended response. The error bars are bootstrapped 95% confidence intervals.

These changes over time align with 1) the ACA being signed into law in 2010 and 2) the implementation of major components of the law. In the top left panel, for example, we see that respondents who felt favorably about the law had not yet started to discuss specific policies, including changes benefiting people with pre-existing conditions and allowing young people to stay on their parents' health insurance.

In the bottom right panel, there is a small uptick in discussion of personal costs in late 2013 that continues into 2015. The coarse dotted line is the average of unfavorable responses for 2009 through early 2013. This late 2013 change corresponds to announcements on changes to individuals' health insurance policies, including changes in out-of-pocket costs. For example, in late October to November 2013, many Americans received notices that their insurance plans would no longer be offered and that they would need to purchase new plans.[30] This shift was accompanied by a small increase in unfavorable attitudes toward the ACA in the Kaiser Health Tracking Poll.[31]

The increase in discussion of personal cost by people unfavorable to the law comes at the same time as a decrease in the more abstract role of government response (top right panel). This decline in role of government responses continues a long slow decline since 2009 among individuals unfavorable to the law. Overall, the decline suggests that attitudes toward the ACA became more concrete over time and with the law's actual implementation. This pattern raises the prospect that abstract rhetoric by Republicans in Congress – focusing on repeal of the ACA – had less traction

[29] Our data in 2016 comes from the surveys of activists.

[30] https://www.washingtonpost.com/news/wonk/wp/2013/10/29/this-is-why-obamacare-is-cancelling-some-peoples-insurance-plans/

[31] http://www.kff.org/interactive/kaiser-health-tracking-poll-the-publics-views-on-the-aca/

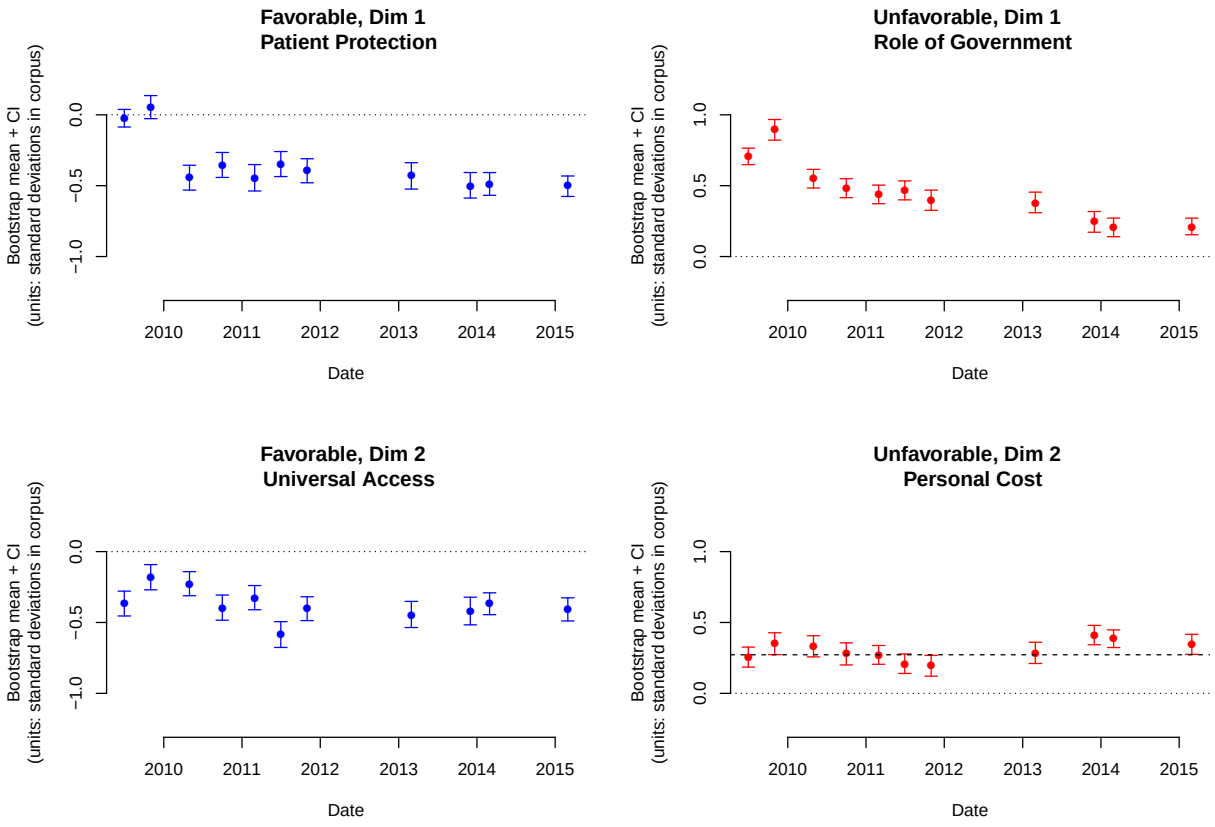


Figure 6: *Changes in ACA justifications over time.* This figure shows the 2009 through 2015 means of the 1st and 2nd dimensions of our text scaling separated by respondents who felt favorably or unfavorably about the ACA. The error bars are 95% confidence intervals based on bootstrapped standard errors. This data is subset to include only respondents in the Kaiser Family Foundation and Pew Research surveys and excludes our nationally representative sample in 2016, along with the surveys of activists in 2016.

with the public after the ACA's implementation.

An implication of this analysis, then, is that the decreased focus on the role of government among the public could have contributed to the Republicans' initial failure to repeal the ACA. For the ACA, this policy feedback (Campbell, 2012) may have had a specific form; one that mostly affects within party politics. Republicans were unified in their dislike for the ACA, but disagreed on what should be done to change it. This disagreement increased in severity after the ACA's implementation, as abstract role of government concerns became less dominant and concrete concerns

about personal costs increased.

During debate over policy proposals to repeal and replace the ACA, commentators highlighted divides between moderate and conservative Republicans. Here, we go further to identify divisions in public opinion that did not map cleanly onto a single dimension of political conflict, implying that the Republican division was more complex than competing degrees of conservatism would have been. Instead, the moderate to conservative spectrum on the right was composed of ultimately competing preferences about personal costs and the role of government. After implementation but not before, Republicans either satisfied both of these criticisms or split their party on these lines.

Discussion

Pivot analysis provides ordered and interpretable representations of short text data, along with keywords to help evaluate results. Its output substantially outperforms existing techniques on low-dimensional predictions. The top dimensions from pivot analysis further correspond to intuitive explanations for individuals' changing justifications for supporting or opposing the Affordable Care Act – pre and post implementation of the law– as well as known political cleavages within and across American political parties. Adding information from word embeddings pre-trained on general text makes the representations more general, but relying too much on that information comes at the expense of domain-specific meanings.

The agreement on the topic in an open-ended survey thus seems to allow respondents to repeat a small set of vocabulary in meaningful patterns. Our departure from prior work is due to our specific interest in representing short and focused texts in an interpretable way. This method, then, will not be well-suited to all texts and purposes. In particular, latent Dirichlet allocation (Blei, Ng and Jordan, 2003) and correlated topic models (Blei and Lafferty, 2007; Roberts et al., 2014) will likely outperform this method when text is particularly diverse, and for which very common words will not necessarily be useful starting points for scaling the text.

Existing unsupervised methods for long form political text (Slapin and Proksch, 2008) may also outperform this method when there is a large variety of word usage by sophisticated speakers, as these provide confidence intervals for individuals. Other methods in political science, such as semi-supervised methods that use hand labels (Benoit et al., 2016), methods that include other indicators of political preferences (Kim, Londregan and Ratkovic, 2016), and supervised methods (Laver, Benoit and Garry, 2003; Lowe, 2007; Beauchamp, 2012), will also be well-suited for tasks like measuring ideology in legislatures.

References

Beauchamp, Nicholas. 2012. "Using Text to Scale Legislatures with Uninformative Voting." pp. 1–44. 39

Benoit, Kenneth, Drew Conway, Benjamin E Lauderdale, Michael Laver and Slava Mikhaylov. 2016. "Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data." *The American Political Science Review* 110(2):278–295. 39

Bischof, Jonathon and Edoardo M Airoldi. 2012. Summarizing topical content with word frequency and exclusivity. In *ICML*. 32

Blei, David M, Andrew Y Ng and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3(Jan):993–1022. 2, 6, 38

Blei, David M and John D Lafferty. 2007. "A correlated topic model of science." *The Annals of Applied Statistics* 1(1):17–35. 2, 6, 38

Blitzer, John, Dean Foster and Sham Kakade. 2011. Domain adaptation with coupled subspaces. In *AISTATS*. 3, 22

Bond, Robert and Solomon Messing. 2015. "Quantifying Social Media's Political Space: Estimating Ideology from Publicly Revealed Preferences on Facebook." *American Political Science Review* 109(01):62–78. 9, 15

Bonica, Adam. 2014. "Mapping the Ideological Marketplace." *American Journal of Political Science* 58(2):367–386. 15

Campbell, Andrea. 2012. "Policy Makes Mass Politics." *Annual Review of Political Science* 15(1):333–351. 37

Chang, Jonathan, Jordan Boyd-Graber, Sean Gerrish, Chong Wang and David M Blei. 2009. "Reading tea leaves: How humans interpret topic models." *NIPS* . 22

Deerwester, Scott, Susan T Dumais, George W Furnas, Thomas K Landauer and Richard Harshman. 1990. "Indexing by Latent Semantic Analysis." *Journal of the American Society for Information Science* 41(6):391–407. 2, 9, 15

Dhillon, Paramveer S, Dean P Foster and Lyle H Ungar. 2015. "Eigenwords: Spectral word embeddings." *Journal of Machine Learning Research* 16:3035–3078. 11

Friedman, Jerome H, Trevor Hastie and Rob Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33(1). 29

Hastie, Trevor, Robert Tibshirani and J Jerome H Friedman. 2001. *The elements of statistical learning*. New York: Springer. 13

Hoerl, Arthur E and Robert W Kennard. 1970. "Ridge regression: Biased estimation for nonorthogonal problems." *Technometrics* 12(1):55–67. 12

Hopkins, Daniel J. 2017. "The Exaggerated Life of Death Panels: The Limited Influence of Elite Rhetoric in the 2009-2012 Health Care Debate." *Political Behavior* . 7

Kim, In Song, John Londregan and Marc Ratkovic. 2016. "Estimating Spatial Preferences from Votes and Text." *Political Analysis* pp. 1–50. 39

Laver, Michael, Kenneth Benoit and John Garry. 2003. "Extracting policy positions from political texts using words as data." *American Political Science Review* 97(2):311–331. 6, 39

Levy, Omer and Yoav Goldberg. 2014. "Neural word embedding as implicit matrix factorization." *NIPS* . 3, 22

Lowe, Will. 2007. "Understanding Wordscores." *Political Analysis* 16(04):356–371. 2, 6, 15, 39

Lowe, Will. 2016. "Scaling Things We Can Count.". 2, 6, 9, 31

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado and Jeffrey Dean. 2013. "Distributed representations of words and phrases and their compositionality." *NIPS* . 22

Pennington, Jeffrey, Richard Socher and Christopher D Manning. 2014. "Glove: Global Vectors for Word Representation." *EMNLP* 14:1532–1543. 16, 22

Poole, Keith T and Howard Rosenthal. 1985. "A spatial model for legislative roll call analysis." *American Journal of Political Science* pp. 357–384. 6

Roberts, Margaret, Brandon Stewart and Dustin Tingley. 2016. "stm: R Package for Structural Topic Models." *Journal of Statistical Software* . 15, 16

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58(4):1064–1082. 2, 6, 38

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(3):705–722. 2, 6, 39

Spärck Jones, Karen. 1972. "A statistical interpretation of term specificity and its application in retrieval." *Journal of Documentation* 28(1):11–21. 13

Tibshirani, Robert. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society Series B (Methodological)* 58(1):267–288. 12, 29

Wallach, Hanna M, David M Mimno and Andrew McCallum. 2009. "Rethinking LDA: Why Priors Matter." *NIPS* pp. 1973–1981. 2

Zaller, John and Stanley Feldman. 1992. "A simple theory of the survey response: Answering questions versus revealing preferences." *American Journal of Political Science* pp. 579–616. 5

Appendix

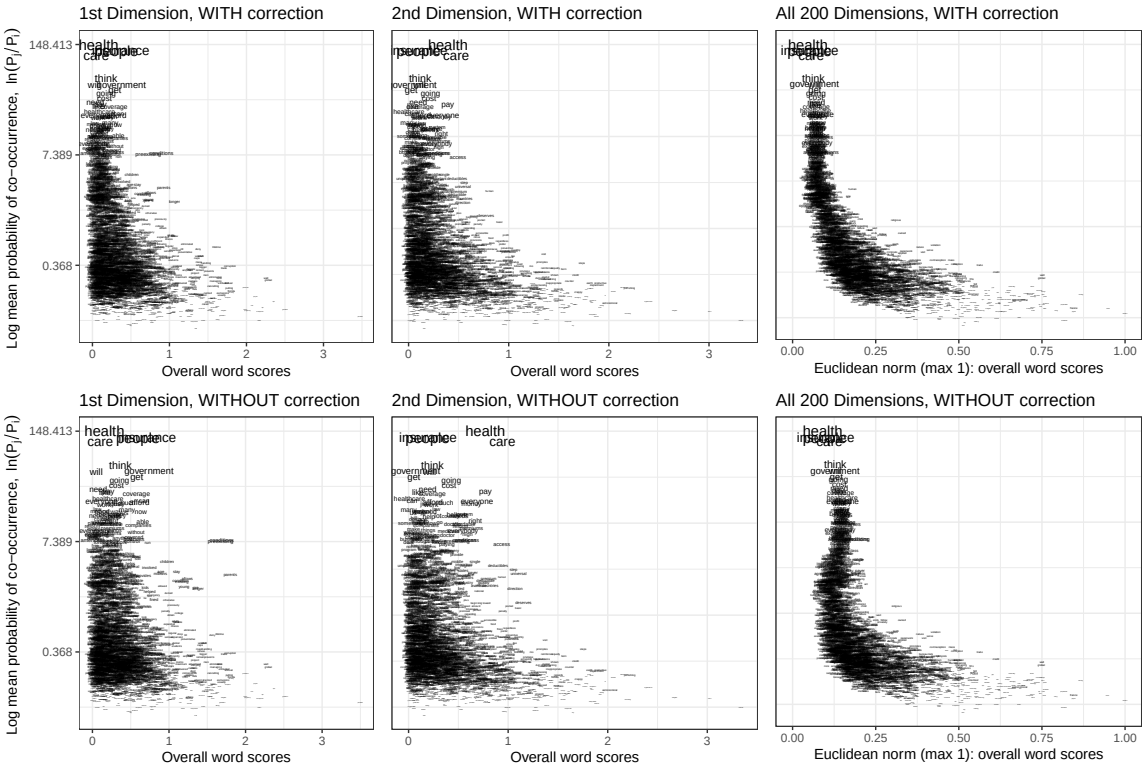


Figure 7: *Word scores, adjusted.* The top 3 panels show the adjust coefficients, while the bottom 3 panels show unadjusted coefficients.

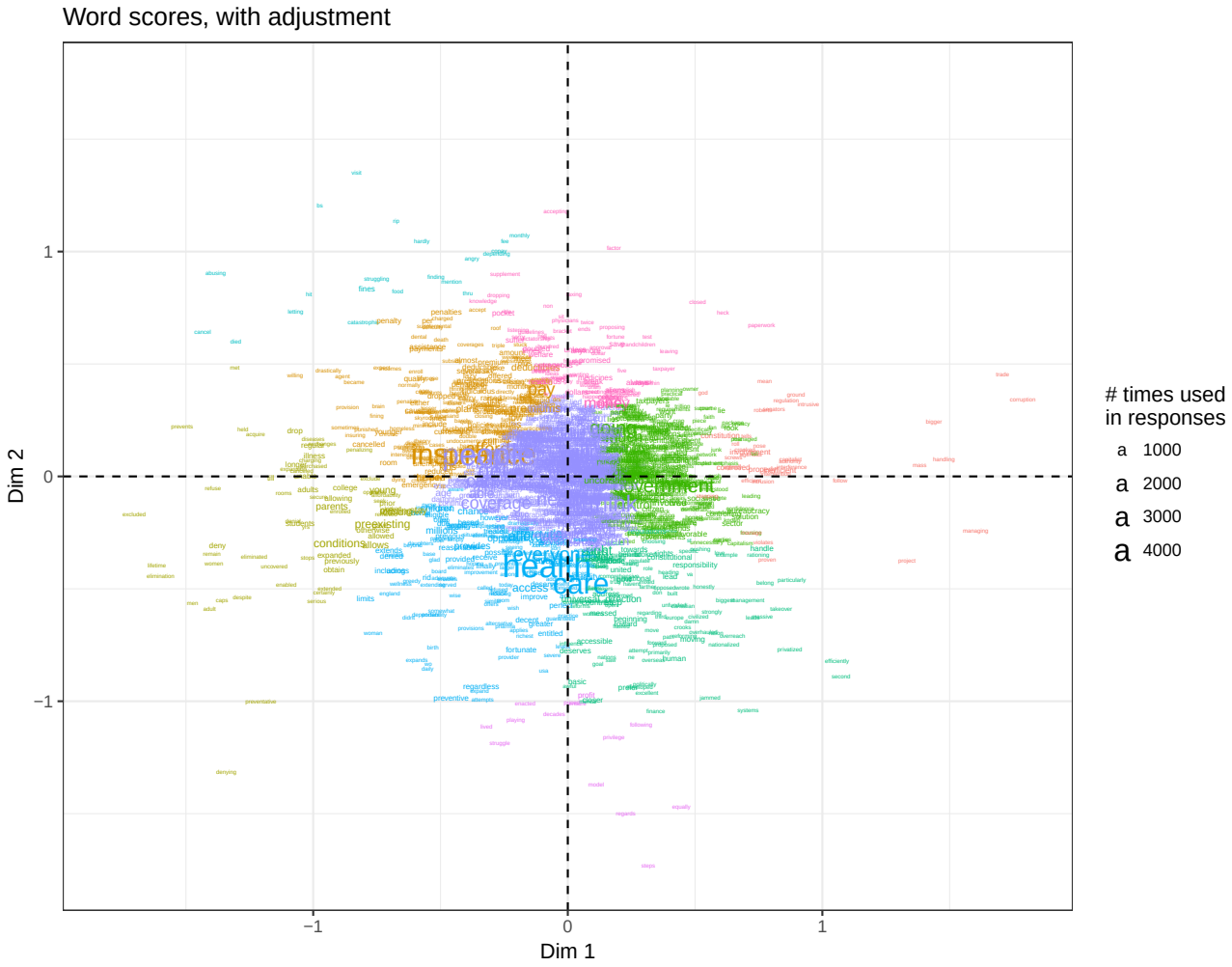


Figure 8: *Word scores, adjusted.*

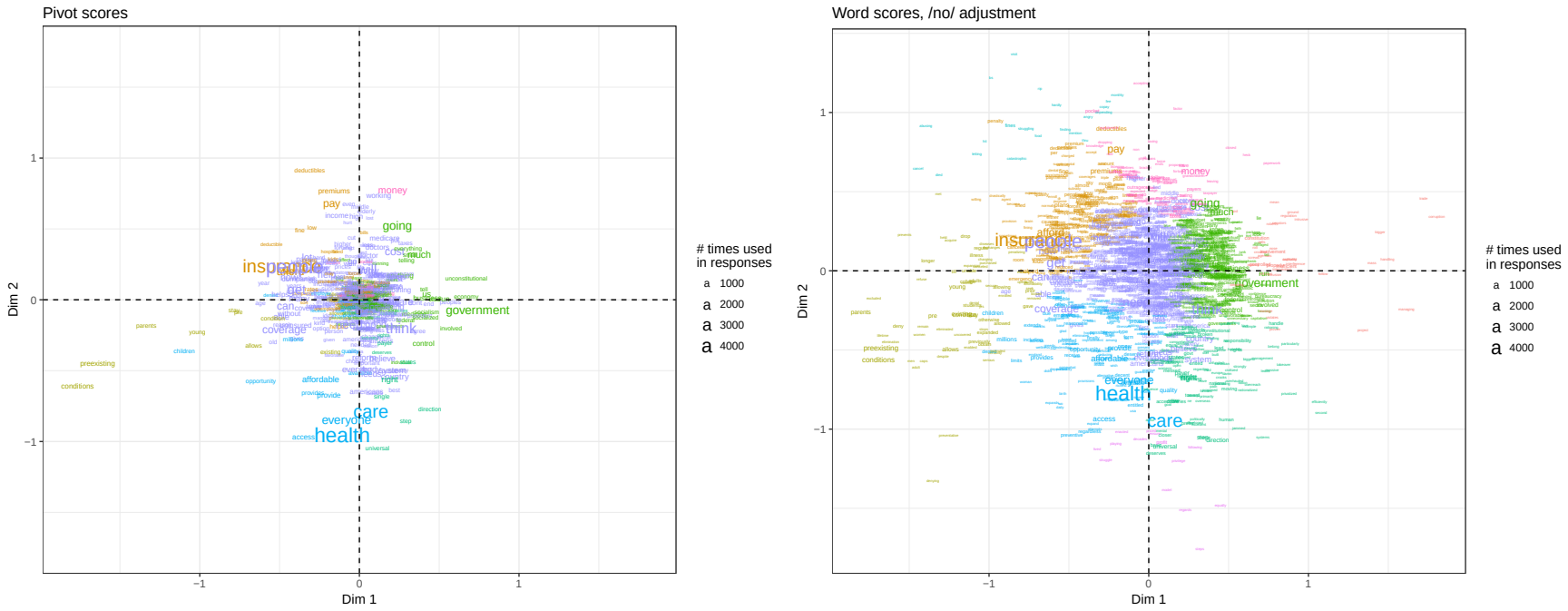


Figure 9: *Word scores.* The left panel shows the word scores for the pivot words. The right panel shows the word scores for all words. Note that the final word scores have an adjustment that moves the pivot words closer to 0 (see Figures 1 and 8). This adjustment moves the pivot words close to the inner edge of their associated word clusters and corrects for non-linearity in the pivot scores.

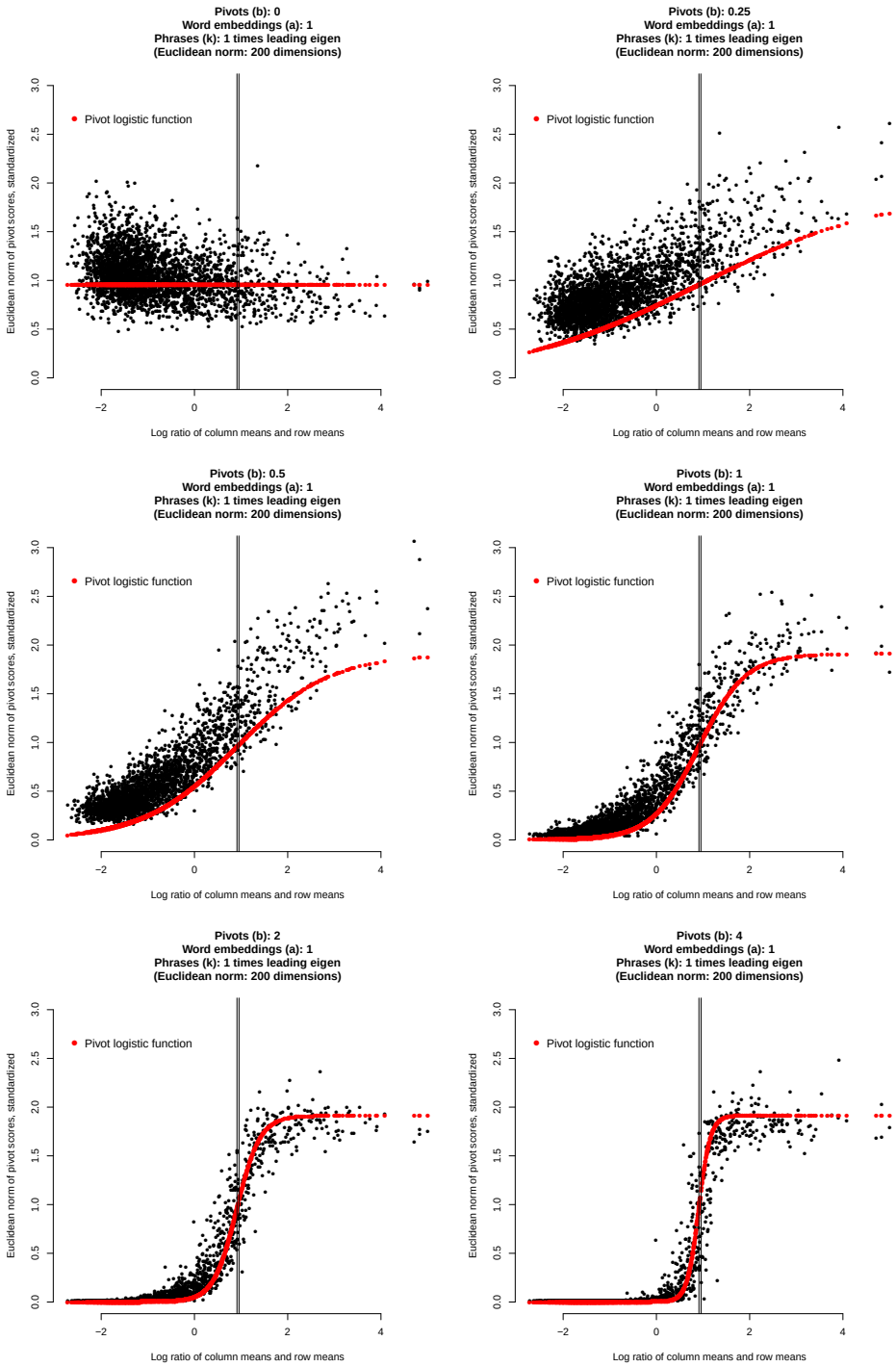


Figure 10: *Tuning* ***b*** *to induce pivots.* We use the *b* value in the bottom left panel.

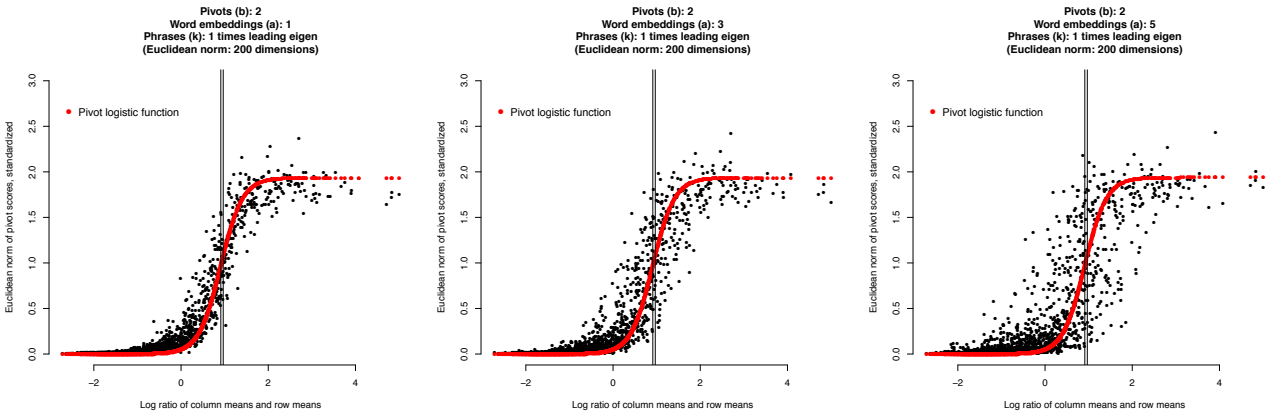


Figure 11: *Tuning* ***a*** *to increase effect of word embeddings and smooth the pivot transition.* This is a smoothing parameter. We use the *a* value in the left panel.

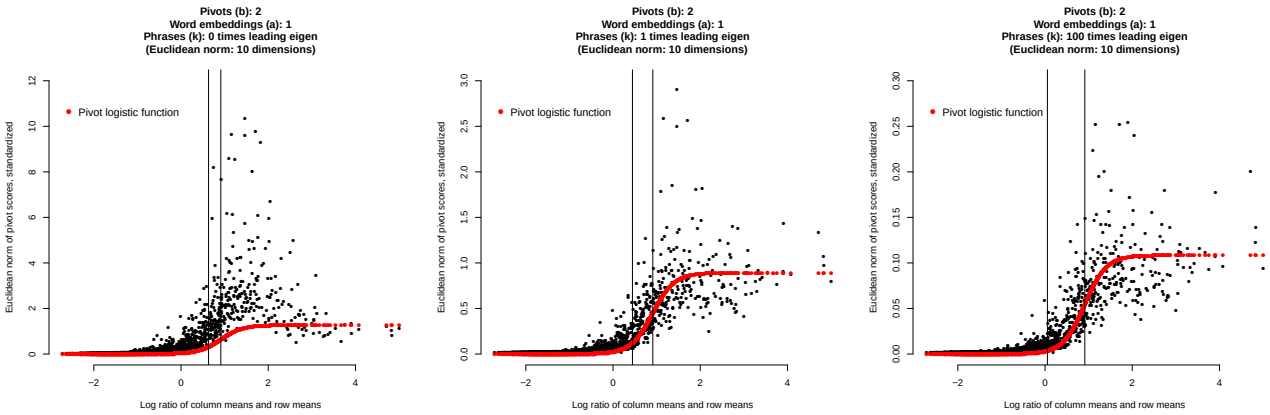
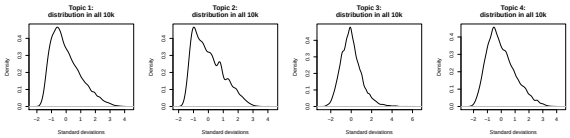


Figure 12: *Tuning* ***k*** *to remove CCA scale invariance.* We use the k value in the middle panel. This figure shows the Euclidean norm on only the first 10 dimensions. These function approximate the logistic function in red in high dimensions. For some values of the hyperparameters, the very common words will not converge to the logistic curve, and the function will resemble tf-idf standardization.



Keywords (highest probability) - TOPIC MODEL			
Topic 1	Topic 2	Topic 3	Topic 4
"health insurance / cost"	"government / cost"	"benefit / cost"	"health care"
No relationship	Anti	Anti	Pro
people	government	think	health
insurance	going	healthcare	care
get	cost	work	will
afford	like	country	everyone
can	much	something	need
coverage	money	system	help
now	just	go	lot
pay	us	know	good
many	right	expensive	believe
law	want	buy	everybody
medical	way	business	needs
companies	costs	getting	affordable
conditions	one	doctor	make
without	things	dont	reform
preexisting	doctors	free	able
premiums	take	needed	better
access	medicare	benefits	americans
covered	anything	keep	bill
really	pay	making	time
paying	well	made	get

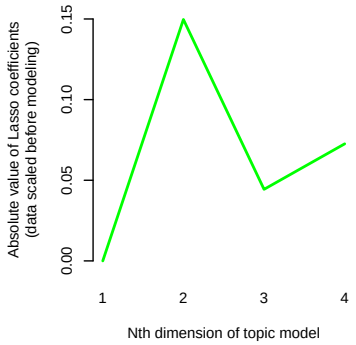


Table 8: *Topic model summary (4 topics).*

Keywords (FREX) - TOPIC MODEL			
Topic 1 "health insurance / cost"	Topic 2 "government / cost"	Topic 3 "benefit / cost"	Topic 4 "health care"
No relationship	Anti	Anti	Pro
afford	government	small	care
coverage	going	fair	everyone
now	much	hard	needs
many	money	went	available
conditions	things	nation	sure
without	take	basically	elderly
preexisting	medicare	rich	chance
covered	anything	living	basic
high	control	major	heath
uninsured	away	real	accessible
insured	run	healthcare	harder
class	nothing	hours	answer
cover	taxes	drug	effort
children	taking	started	need
middle	choice	work	fortunate
stay	long	self	hoping
parents	step	benefits	improvement
existing	single	reach	provide
age	else	seems	ensure
allows	payer	supposed	senior

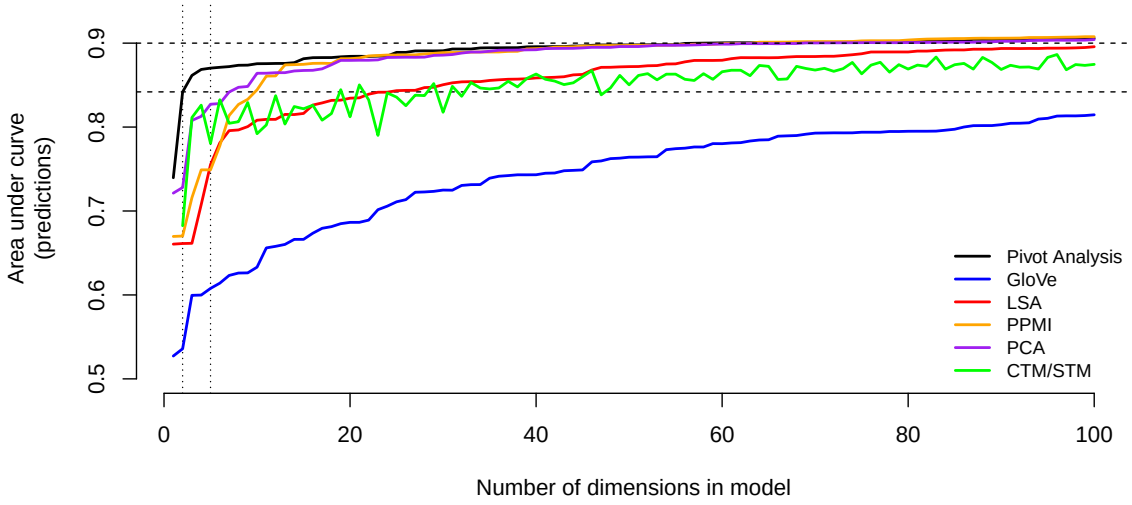


Figure 13: *Dimensionality of other methods (all 100 dimensions).*

Keywords			
Dimension 1		*Dimension 2*	
"patient protection"	"role of government"	"universal access"	"personal cost"
Pro	Anti	Pro	Anti
preexisting	direction	universal	premiums
conditions	run	care	tax
children	government	access	deductibles
age	socialized	health	money
fine	every	direction	pay
parents	much	step	high
able	business	provide	middle
can	dont	single	deductible
afford	small	everyone	fined
benefit	senior	americans	rates
insurance	congress	affordable	low
coverage	us	needs	much
now	federal	right	income
without	involved	old	lied
people	control	available	doctor
covered	everything	national	bills
companies	medicine	every	also
get	done	system	keep
stay	choice	country	dollars
helps	general	important	doctors

Table 9: *Output summary (with word embeddings).* Including more information from word embeddings makes the representations slightly less domain-specific.