Scaling Data from Multiple Sources

Ted Enamorado* Gabriel López-Moctezuma† Marc Ratkovic‡

April 18, 2019

Abstract

We introduce a method for scaling two data sets from different sources. The proposed method estimates a latent factor common to both datasets as well as an idiosyncratic factor unique to each. In addition, it offers a flexible modeling strategy which permits the scaled locations to be a function of covariates, and efficient implementation allows for inference through resampling. A simulation study shows that our proposed method improves over existing alternatives in capturing the variation common to both datasets, as well as the latent factors specific to each. We apply our proposed method to vote and speech data from the 112th U.S. Senate. We recover a shared subspace that aligns with a standard ideological dimension running from liberals to conservatives, while recovering the words most associated with each senator's location. In addition, we estimate a word-specific subspace that ranges from national security to budget concerns, and a vote-specific subspace with Tea Party senators on one extreme and senior committee leaders on the other.

*Ph.D. Candidate, Department of Politics, Princeton University, Princeton NJ 08544. Email: tede@princeton.edu, URL: http://www.tedenamorado.com

†Assistant Professor, Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena CA 91125. Email: glmoctezuma@caltech.edu, URL: http://glmoctezuma.com

‡Assistant Professor, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 608-658-9665, Email: ratkovic@princeton.edu, URL: http://www.princeton.edu/∼ratkovic

1 Introduction

Increasingly, political scientists confront not just large amounts of data but different *types* of data. As examples, political actors will often generate text data and vote data (e.g. Lauderdale and Clark, 2014); countries may have sets of qualitatively distinct attributes, such as political, social, and economic indicators (e.g. Coppedge et al., 2015); the same survey questions may be given to different groups of actors (e.g. Shor and McCarty, 2011); campaign contributions may flow from the same actors to both state and federal candidates (Bonica, 2014). In each case, the researcher must analyze data on different attributes for the same actors (say, tweets and votes from legislators, Barbera 2016), or the same attributes but on different actors (say, surveys given to both legislators and the mass public, Bafumi and Herron 2010).

As a first pass, the data from different sources may simply be pooled and scaled (Quinn, 2004; Hoff, 2007; Jackman and Trier, 2008; Murray et al., 2013). Pooling suffers, though, when one data set has much more information, swamping the information from the other set. Combining data from different sources creates even more subtle theoretical and empirical issues. Jessee (2016) illustrated the underlying problem rather elegantly. Using survey data for citizens and legislators, he showed that scaled locations can vary as the relative numbers of individuals from two samples are pooled and used to estimate ideological positions. The problem arises because the different groups give different weights to each question, and it generalizes to the problem of how to weight data coming from two different sources.

Existing approaches have addressed, but not quite solved, the issue of how to weight different types of data. For example, Kim, Londregan and Ratkovic (2018) develop a choice-theoretic model for combining words and votes, but a tuning parameter that balances the proportion of information coming from each source is not estimated within the model. Hobbs (2017) combines information from multiple text sources using a version of canonical correlation analysis (e.g. Hastie, Tibshirani and Friedman., 2013, Sec. 3.7), a

method closely related to ours. The method advanced by Hobbs (2017), though, is tailored to short bursts of speech and does not offer means of inference. Similarly, Weighted Multidimensional Scaling (WMDS) of Borg et al. (2013); Borg and Groenen (2005) combines multiple dissimilarity matrices to recover a single underlying dimension (see also Jacoby, 1986, 2009). WMDS, though, returns only locations for the observations and not for the outcomes, i.e. votes or words, on which the observations are measured. This issue also plagues methods that must pre-select, rather than estimate, ideologically charged words (Groseclose and Milyo, 2005; Gentzkow and Shapiro, 2010; Martin and Yurukoglu, 2017).

We develop a general framework for combining data from multiple sources. The method, *Multi-Dataset Multidimensional Scaling* (MD2S) simultaneously scales two datasets, decomposing the data into three separate factors: one spanning a latent space common to both datasets, and two idiosyncratic subspaces – one per dataset. For example, combining votes and words on the same actors, MD2S estimates three latent scales. The first is a joint scale informed by both words and votes. The second is informed by words, but contains no information from votes. Likewise, the third is informed by votes, but not words.

We build off work in statistics and education focusing on recovering the correlation and shared factors across multiple surveys or exams (Tucker, 1958; Browne, 1979; Anderson, 1989; Klami, Virtanen and Kaski, 2013; Bach and Jordan, 2005; Gupta et al., 2011; Tipping and Bishop, 1999). This model, "Inter-Battery Factor Analysis," is precisely the model described above. We offer a likelihood-based method for optimally weighting the information coming from the two sources, allow the user to include covariates in estimating the scaled locations, and derive and implement an efficient algorithm for estimation.

The advantages of our proposed method are fourfold. First, we recover scaled locations for both observations (say, legislators) and features (say, text and votes). Estimating, for example, which words anchor a dimension's extremes greatly facilitates interpretation. Second, we allow for inference on the number of latent dimensions. Distinguishing a dimension that is signal from one that is noise is a perennial

problem, often unaddressed, in the scaling literature. To this end, we implement a permutation test to distinguish a given dimension from noise. Our third advance is in terms of estimation. Building on insights first advanced in Aldrich and McKelvey (1977), we implement an efficient estimation routine that performs well when the number of attributes grows large, as with text data where the researcher has a document-term matrix with counts on thousands of n-grams for each speaker. Inference on the scaled locations for both shared and idiosyncratic subspaces is performed via bootstrapping. Finally, scaled locations are modeled as a function of covariates. This facilitates conducting inference on whether or how scaled locations relate to covariates of interest, giving a principled way to explore the estimated latent scales with substantive information.

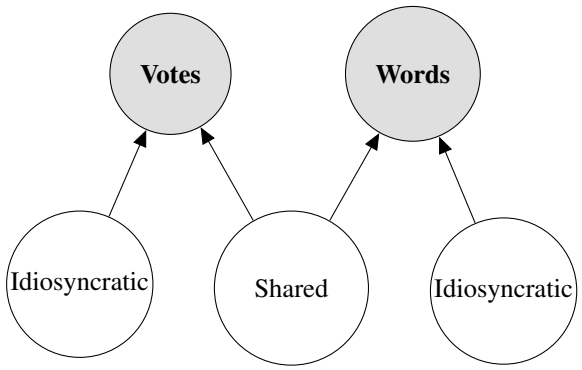
We illustrate the method's use and efficacy through a simulation exercise and an empirical application. We show in the simulation study that MD2S recovers a shared and idiosyncratic dimensions more accurately than existing methods suited to combine multiple datasets, especially as the number of attributes grows large. We then apply it to roll call votes and floor speech in the US Senate. Our shared first dimension aligns with the standard ideological dimension running from liberals to conservatives (e.g. Poole, 2005), with the selected words that differentiate Senators on economic issues. We recover a word-specific subspace that ranges from national security to budget concerns, and a vote-specific subspace with Tea Party senators on one extreme and senior committee leaders on the other.

The rest of this paper proceeds as follows: first, we motivate MD2S. Second, we describe the model and method in formal detail. Third, we present results from a simulation study, followed by an analysis of roll call and speech data from the 112th Senate. We conclude with some remarks and a discussion of possible extensions.

2 Motivation and Use Cases

To illustrate the basic problem and insight, consider the case where we observe two different streams of data, votes in a roll call matrix and word counts in a document-term matrix, that are observed on the same actors. As is common in text data, assume that we have many more words than votes. Were we to simply join the two datasets and estimate a single scale, it would be closer to the words-only scaling than the votes-only scaling. The words contain more information, but we are not interested in all of the word data. We are most interested in the word data that contribute to explaining the joint variation in both types of data. We could conduct multiple analyses after reweighting the matrix, to find a suitable balance between words and votes in the scaling procedure as in Kim, Londregan and Ratkovic (2018), but this sidesteps the problem of relative weighting rather than solving it.

MD2S solves this problem by returning three factors from the two datasets. The first is a joint factor, estimated to explain the largest amount of variance common to both datasets. The next two are idiosyncratic factors, unique to each data source and uncorrelated with the common factor:



This model is the Inter-Battery Factor Analysis (IBFA) model of Tucker (1958). As noted above, MD2S builds on this model in several regards: estimating the number of latent dimensions, providing an efficient and effective estimation algorithm for a large number of attributes, and modeling the scaled locations with covariates.

Use Cases and Scope. If the data come from a single source, or the researcher is willing to ignore the problem of one data source overwhelming the other, then a standard principal components or factor analysis should be utilized. There may be several cases, though, where the researcher may wish to model the two different data sources. Our interest in this method was motivated by combining text and votes, where the sheer volume of the textual data may overwhelm the vote data. Beyond this particular case, the method's use fall into two broad categories: combining data sets and contrasting them.

Combining data sets may involve bringing auxiliary information to bear on a problem. For example, roll call votes are not informative in legislatures with strong party systems or in the presence of pressures for unanimous or lopsided voting, so words can be used to differentiate among members (for more, see Kim, Londregan and Ratkovic, 2018; Kellerman, 2012). Relatedly, one data source may not have sufficient signal to generate a reliable scale, so a second data source can help leverage the first (e.g., Hobbs, 2017). Combining the two sources offers an additional benefit. Placing words and actors in the same space, as in our applied example, allows the researcher to use the selected features to characterize the substantive meaning of each dimension.

Bridging across different actors is another form of combining data. Jessee (2016) highlighted the problem of weighting data from different sources when bridging across different sets of respondents (e.g., Tausanovitch and Warshaw, 2013; Shor and McCarty, 2011; Bafumi and Herron, 2010). If the two groups have different item discrimination parameters, simply pooling the two sets generates ambiguity in their ideal point estimates. The estimates will vary based on either the amount of information or based on the number of respondents, in the two sets.

A third instance for combining data sets comes when constructing indices. Consider the impressive set of measures for cross-national political, civic, and institutional comparison assembled by Coppedge et al. (2015). Generating an index by aggregating from finer to coarser measures requires a method that is not

sensitive to the number of items at each level.

A second use of the method is for contrasting the two information sources. This approach differs from methods that only uncover a single scale; we discuss these methods in more detail below. MD2S offers the ability to isolate a set of factors based off whether they are informing both data sets, or exclusively one or the other. With data on word usage on the same individuals before and after an event, three sets of factors can be recovered: a factor common to both before and after, one unique to before and uninformed by after the event, and one unique to after the event and not informed by information before the event. Examples of this analysis involve contrasting Twitter data (Barbera, 2016), transcripts of Federal Open Market Committee Meetings before and after a transparency shock (Hansen, McMahon and Prat, 2014), or Weibo microblogs before or after censorship (Hobbs and Roberts, 2018). Our method offers a structured way of separating out a common factor, allowing the researcher to estimate how latent factors vary across the two datasets.

3 The Proposed Method

For each observation, we observe two vectors of outcomes, $Y^*_{(1)i}$ and $Y^*_{(2)i}$ with $i \in \{1, 2, \ldots, N\}$. To ease notation, we will use the index m to denote either 1 or 2, so $Y^*_{(m)i}$ is the generic notation for either $Y^*_{(1)i}$ or $Y^*_{(2)i}$. We assume $Y^*_{(m)i}$ is of length K_m, where K_1 may not equal K_2, and we use $Y_{(m)j}$ to denote the N values of the j^{th} feature from data source m. For example, $Y^*_{(1)i}$ may be a vector of K_1 word counts uttered by legislator i, $Y^*_{(2)i}$ may be a set of K_2 observed roll call votes for the same legislator, and $Y_{(1)j}$ is the vector of N counts for a word $j \in \{1, 2, \ldots, K_1\}$ while $Y_{(2)j}$ are the N outcomes for vote $j \in \{1, 2, \ldots, K_2\}$.

We denote the outcome data matrices as:

$$Y^*_{(1)} = \begin{bmatrix} Y^{*\top}_{(1)1} \\ Y^{*\top}_{(1)2} \\ \cdots \\ Y^{*\top}_{(1)N} \end{bmatrix}; \quad Y^*_{(2)} = \begin{bmatrix} Y^{*\top}_{(2)1} \\ Y^{*\top}_{(2)2} \\ \cdots \\ Y^{*\top}_{(2)N} \end{bmatrix}. \tag{1}$$

with individual outcomes in rows. We assume that each matrix $Y^*_{(m)}$ is on a common scale. This may be due to a natural scale, such as binary vote data, columns may be normed to have sample standard deviation one, or some other method may be used to place all columns of $Y^*_{(m)}$ on a common scale (e.g., Quinn, 2004; Hoff, 2007; Murray et al., 2013). The important point for our method is that all columns of $Y^*_{(m)}$ be on a common interval scale. While each matrix must be on a common scale, the two separate matrices may be on different scales. For example, $Y^*_{(1)}$ may contain roll call votes and $Y^*_{(2)}$ word counts.

3.1 The Model

In practice, as the intercept is rarely of interest, we pre-process the matrices by double-centering them, so that the row-mean, column-mean, and grand mean is zero. We denote the double-centered matrices as $Y_{(m)}$.[1] Thus, we model $Y_{(1)}$ and $Y_{(2)}$ in terms of their latent factors as[2]

$$Y_{(1)} = Z_S L_{(1)} W^\top_{(1)} + Z_{(1)} D_{(1)} B^\top_{(1)} + \Omega_{(1)} \tag{2}$$

$$Y_{(2)} = Z_S L_{(2)} W^\top_{(2)} + Z_{(2)} D_{(2)} B^\top_{(2)} + \Omega_{(2)}. \tag{3}$$

[1] The maximum likelihood estimates for the intercept terms are the sample analogs (Tipping and Bishop 1999), i.e. the double-centering matrix with each element the row-mean plus the column-mean, less the grand mean. See Poole and Rosenthal 1997 for a discussion of double-centering.

[2] We chose notation consistent with Murphy (2012) and Klami, Virtanen and Kaski (2013). We denote all observed outcomes as $Y_{(m)}$ instead of Y and X.

We will refer to the $N \times Q_S$ matrix Z_S as the shared subspace and the $N \times Q_{(m)}$ matrix $Z_{(m)}$ as the idiosyncratic subspace. Z_S contains latent locations on the shared subspace in columns for each of the Q_S dimensions. Similarly, each column of $Z_{(m)}$ contains the latent locations in the idiosyncratic subspace for $Q_{(m)}$ latent dimensions. $L_{(m)}$ is a $Q_S \times Q_S$ nonnegative, diagonal matrices of loadings for the shared subspace. We assume that the two matrices $L_{(1)}$ and $L_{(2)}$ are proportional, so any difference between them is attributable to the relative scales across data sources $Y_{(m)}$. $W_{(m)}$ is a $Q_{(m)} \times Q_{(m)}$ matrix of factors for the shared subspace for dataset $Y_{(m)}$. $D_{(m)}$ is a $Q_{(m)} \times Q_{(m)}$ diagonal matrix of loadings for the idiosyncratic subspace, and $B_{(m)}$ is the $K_{(m)} \times Q_{(m)}$ of factors for the idiosyncratic subspace. The matrix $\Omega_{(m)}$ is a matrix of mean-zero, independent, equivariant noise.

We have modeled each observed data matrix $Y_{(m)}$ in terms of a shared subspace Z_S and individual subspaces $Z_{(m)}$. The researcher may believe, though, that the estimated scaled locations may vary systematically with some set of known covariates available for the N observations in the data. As in Roberts et al. (2014), we allow the scaled locations to take the following form:

$$Z_S = X_S\beta_S + \Omega_{Z_S} \tag{4}$$

$$Z_{(m)} = X_{(m)}\beta_{(m)} + \Omega_{Z_{(m)}}, \text{ for } m \in \{1, 2\} \tag{5}$$

The covariates X_S, $X_{(m)}$ structure the systematic factors of Z_S, $Z_{(m)}$, respectively.

We make five assumptions for identifying the model (for a discussion of identification, see Tipping

and Bishop, 1999, Appendix A.1), where the assumptions hold for $m \in \{1, 2\}$:

$$Z_S^\top Z_S = W_{(m)}^\top W_{(m)} = I_{Q_S} \tag{6}$$

$$Z_{(m)}^\top Z_{(m)} = B_{(m)}^\top B_{(m)} = I_{Q_{(m)}} \tag{7}$$

$$Z_{(m)}^\top Z_S = \mathbf{0}_{Q_{(m)} \times Q_S} \tag{8}$$

$$L_{(1)} \propto L_{(2)} \tag{9}$$

$$L_{(m)}, D_{(m)} \text{ are diagonal with non-negative entries} \tag{10}$$

Assumptions (6) - (7) state that, within a given subspace, the latent scalings and factors are uncorrelated and length one. Assumption (8) states that the common subspace spanned by Z_S is not correlated with the idiosyncratic scalings. This assumption allows us to differentiate the shared subspace from each idiosyncratic subspace. Assumption (9) requires the variation in loadings for the shared subspace to be explained by the relative scales across data sources. Assumption (10) identifies the particular rotation that we estimate. Specifically, we are assuming that the factors $W_{(m)}$ and $B_{(m)}$ are numerically equal to singular decompositions of the shared and idiosyncratic subspaces of $Y_{(m)}$,[3] respectively. Note that we only identify the latent factors $Z_S, Z_{(m)}, W_S$ and $B_{(m)}$ up to sign.[4] We follow convention and assume the elements of L and $D_{(m)}$ are nonnegative and arranged in decreasing order. We discuss relaxations of these assumptions in Section 3.5.

[3] See (Tipping and Bishop, 1999) and our discussion below for more.

[4] This means that the data cannot differentiate between a model with estimates $\{Z_S, Z_{(m)}, W_S, B_{(m)}\}$ and $\{-Z_S, -Z_{(m)}, -W_S, -B_{(m)}\}$.

3.2 A Probabilistic Framework

We next embed our factor model in a probabilistic framework, where we recover maximum likelihood estimates of the factors. The probabilistic MD2S model can be written as

$$Y_{(m)}|W_{(m)}, B_{(m)}, L_{(m)}, D_{(m)} \sim \mathcal{N}(Z_S L_{(m)} W_{(m)} + Z_{(m)} D_{(m)} B_{(m)}, \sigma^2_{(m)} I_N) \tag{11}$$

This model is an extension of the Probabilistic Principal Components model of Tipping and Bishop (1999); see also Bach and Jordan (2005). We differ from these models as we are most interested in the actors' spatial locations $(Z_S, Z_{(m)})$, so we treat the weights $W_{(m)}$ and $B_{(m)}$ as random and the spatial locations as fixed (see also Aldrich and McKelvey, 1977, p. 117). We maintain the assumption that the errors are of equal variance, and therefore do not vary systematically across individuals or features.

Marginalizing over $W_{(m)}$ and $B_{(m)}$, gives the unconditional densities for $Y_{(m)}$ as

$$Y_{(m)} \sim \mathcal{N}(0_N, C_m) \tag{12}$$

where $C_m = Z_S L^2_{(m)} Z_S^\top + Z_{(m)} D^2_{(m)} Z^\top_{(m)} + \sigma^2_{(m)} I_N$, for $m = 1, 2$.

The data log-likelihood as a function of $(Z_S, Z_{(1)}, Z_{(2)}, L_{(1)}, L_{(2)}, D_{(1)}, D_{(2)})$ can be written as

$$\begin{aligned} &l\left(Z_S, Z_{(1)}, Z_{(2)}, L_{(1)}, L_{(2)}, D_{(1)}, D_{(2)} | Y_{(1)}, Y_{(2)}\right) = \\ &-\frac{1}{2}\left\{N(K_1 + K_2)\log(2\pi) - K_1 \log\left(|C_1|\right) - K_2 \log\left(|C_2|\right) - \operatorname{tr}\left(Y_{(1)} Y^\top_{(1)} C_1^{-1} + Y_{(2)} Y^\top_{(2)} C_2^{-1}\right)\right\} \end{aligned} \tag{13}$$

We derive an expression and results for the maximum likelihood estimates in Appendix A.

3.3 Implementation

In the single dataset setting, Tipping and Bishop (1999) show that the maximum likelihood estimates for each factor are principal components of the data. We extend the result to the MD2S model. Doing so allows for an efficient estimation strategy, whereby we can estimate $Z_S, Z_{(1)}$, and $Z_{(2)}$ directly using an

iterative algorithm, then recover the remaining estimates, $\hat{W}_{(m)}, \hat{B}_{(m)}, \hat{L}_{(m)}, \hat{D}_{(m)}$, afterwards. We prove the validity of this strategy in the following proposition:

PROPOSITION 1 *The maximum likelihood estimates for the shared and idiosyncratic subspaces can be written as singular vectors of functions of the data. Specifically:*

1. *The maximum likelihood estimates for $Z_{(m)}$ are proportional to principal components of $Y_{(m)}^{\top} M(Z_S)$ for $m = 1, 2$.*

2. *Denote $Z_{S|m}$ as the first L_S principal components of $Y_{(m)}^{\top} M(Z_{(m)})$. Then,*

   (a) $Z_S \propto Z_{S|1} w_1 + Z_{S|2} w_2$ with $w_1 + w_2 = 1; w_1, w_2 > 0$ and

   (b) Z_S is selected to maximize $\operatorname{tr}\left(Z_S^{\top} Y_{(1)} Y_{(1)}^{\top} Y_{(2)} Y_{(2)}^{\top} Z_S\right)$.

where $M(A)$ is the annihilator matrix for matrix A : $I - A(A^{\top}A)^{-1}A^{\top}$ with $(A^{\top}A)^{-1}$ denoting the generalized inverse of $(A^{\top}A)$ and I is the commensurate identity matrix.

Proof. *See Appendix A.*

The proposition leads directly to our estimation strategy.[5] Our algorithm estimates the MD2S model using an iterative procedure that updates the estimate of each subspace one at a time, enforcing the constraints in Equations 6–10 along the way. That is, for every iteration until convergence, the estimation proceeds in two steps. Given the previous iteration estimate of the shared space Z_S, we update $\hat{Z}_{(1)}$ and then $\hat{Z}_{(2)}$. Second, we partial the idiosyncratic spaces out to update $\hat{Z}_S$. After convergence in each subspace, we update our estimates of the remaining parameters.

Note that our algorithm allows the computational advantage of having to invert square matrices of whichever size is smaller, $N \times N$ or $K_{(m)} \times K_{(m)}$.[6] For example, in the main empirical application

[5] See the Supplemental Appendix for details.

[6] See Aldrich and McKelvey (1977, p. 117) and Tipping and Bishop (1999, Appendix B) for similar insights.

below we observe 100 voting members, 486 votes, but 2,532 words. Our algorithm is fit through inverting matrices of size 100 $\times$ 100 instead of 486 $\times$ 486 or 2,532 $\times$ 2,532, which gives us sizable computational gains. One advantage of our algorithm is that, at each step, it recovers estimates of the data, $\widehat{Y}_{(1)}$ and $\widehat{Y}_{(2)}$, conditional on current estimates of shared and idiosyncratic subspaces. Thus, all the information at hand is used in estimation.

3.4 Uncertainty

We estimate uncertainty for two parts of the MD2S model: the scaled locations and the number of dimensions. For the scaled locations, we rely on the bootstrapping methodology introduced by Jacoby and Armstrong II (2014). Let $\{\widetilde{Y}^{*}_{(1),b}, \widetilde{Y}^{*}_{(2),b}\}$ denote two b^{th} bootstrapped sample, with $b \in \{1, 2, \ldots, B\}$, where B is some large number, such as 1,000. The bootstrapped sample is generated by fixing the number of rows and sampling $K_{(m)}$ columns for each matrix, with replacement. Uncertainty due to sampling error can be estimated through fitting MD2S to these bootstrapped estimates.

We present a statistical method for estimating the number of dimensions while acknowledging that the first empirical consideration should be substantive interpretability of the estimated subspaces. We recommend separating signal from noise dimensions through the use of a permutation test (e.g. Keele, McConnaughy and White, 2012). A permutation test requires estimating the density of a test statistic on a set of datasets permuted such that under the null hypothesis, there is in-truth no signal in the data, and then the observed value is compared to this simulated null distribution. We are not the first to use a permutation test to separate an estimated scale from noise (see e.g., Mair, Borg and Rusch 2016 and references therein). However, these authors only compare the estimated weight on each dimension to the mean under the simulated null, rather than estimate a p-value (Figure 1 in Mair, Borg and Rusch 2016).

For the permutation test, we assume that there is no structure in the data, so the subspace loadings are

all zero. Formally,

$$\mathcal{H}^0_{(m)L,q} : L_q = 0; \tag{14}$$

$$\mathcal{H}^0_{(m)D,q} : D_{(m)q} = 0 \tag{15}$$

for all (m, q), where $L_q = L_{(1);q}$ and q indexes a given dimension. Under these hypotheses, the observed data is pure noise with no systematic structure, i.e. $Y_{(m)} = \Omega_{(m)}$.

Under these null hypotheses, any permutation of the data is equally likely. We permute the data, estimate dimension weights, and then compare the statistic under the observed data to the statistic under the null distribution. To the extent that the statistic is an outlier under the null hypothesis, we can argue that the null hypothesis is not accurate and there is, in fact, some systematic relationship in the data.

Specifically, we permute the data such that within each column of $Y_{(m)}$, the rows are shuffled. In this case, in truth, there is no systematic relationship in the permuted data. Denote the r^{th} permuted dataset out of R total as $\widetilde{Y}^r_{(m)}$, with R some large number, say 1000. For each permuted dataset, we calculate the dimension weights, $\widehat{L}^r_q$ and $\widehat{D}^r_{(m);q}$. These values are then compared to the estimated values on the non-permuted data, $\widehat{L}_q$ and $\widehat{D}_{(m);q}$.

Under this formulation, a p-value for dimension q in the shared subspace or idiosyncratic subspace can be estimated as

$$\widehat{p}_{S;q} = \frac{\sum_{r=1}^{R} \mathbf{1}\left(\widehat{L}^r_q \leqslant \widehat{L}_q\right)}{R}; \;\; \widehat{p}_{(m);q} = \frac{\sum_{r=1}^{R} \mathbf{1}\left(\widehat{D}^r_{(m),q} \leqslant \widehat{D}_{(m);q}\right)}{R}; \tag{16}$$

We adapt the test to our model by noting that the tests are not independent. The dimensions are estimated in order of decreasing loadings, such that more explanatory dimensions are estimated before less explanatory ones. Therefore, we take as our estimated dimensionality the first d dimensions such that each dimension has an estimated p-value below a given threshold. In our empirical applications, we calculate the estimated dimensionality $\widehat{d} = q$ as the largest q such that dimensions 1 to q have estimated p-

values below 0.1. However, once the permuted p-value is estimated, this threshold can be be manipulated to assess the sensitivity of the estimated number of dimensions.[7]

3.5 Extensions and Discussion of Method

By using a probabilistic model, we are able to augment the model in order to extend the MD2S model to a large class of problems. For example, we can turn the model into a quadratic utility model through utilizing the latent normal representation of a probit model (Clinton, Jackman and Rivers, 2004; Albert and Chib, 1993; Hare et al., 2015; Jackman and Trier, 2008), leaving it commensurate with popular random utility models (e.g. Ladha, 1991). We can also utilize scale- and location-mixtures of normals to accommodate ordinal and count data, as in Goplerud (Forthcoming); Albert and Chib (1993). In this framework, our probabilistic model is the "M"-step of an EM routine, with the "E"-step as an adjustment to the observed data.[8] Our concern here is not with accommodating a particular class of data, such as binary, ordinal, or count data, but instead to develop a framework for integrating multiple sources in a single coherent fashion.

Other possible extensions can be integrated into the MD2S framework. For example, rather than identifying the factors through orthogonality conditions, the researcher could instead allow for correlated factors and instead identify them with a prior; see Klami, Virtanen and Kaski (2013); Gupta et al. (2011) for recent work. Placing a prior could shrink elements of the factor, returning a set of correlated factors that may be easier to interpret, particularly in high-dimensional settings (see, e.g. Rockova and George, 2016, for work in a factor model). In addition, a sparsity prior on the dimension weights could be used

[7] Formally, $\widehat{d}_S = \text{argmin}_q\{q : \widehat{p}_{S;q} > 0.1\} - 1$; $\widehat{d}_{(m)} = \text{argmin}_q\{q : \widehat{p}_{(m);q} > 0.1\} - 1$

[8] For example, in a latent probit model, this step involves adding to the fitted values the mean of a normal covariate truncated at 0 and centered at the fitted value, with support above zero for observed values of "1" or below zero for observed values of "0," and support over the whole line for missing values. See Clinton, Jackman and Rivers (2004); Albert and Chib (1993).

to select the number of underlying dimensions (e.g. Kim, Londregan and Ratkovic, 2018; Hahn, Carvalho and Scott, 2012). Using orthogonality conditions guarantee identification and simplify several of the derivations in our estimation algorithm (see Appendix A), but placing more structure on the factors can be incorporated into our model.

We have also assumed that all of the columns in $Y_{(m)}$ are on the same scale. If an analysis requires combining data on different scales, say a combination of continuous and categorical outcomes, we have two suggestions. First, if all of the data is continuous and approximately normal, each column may be converted to a z-scale by subtracting off the mean and dividing by the sample standard deviation. Recent literature has also suggested placing data on the same scale through an inverse z-transformation of the empirical distribution function,

$$Y^z_{(m);ij} = \Phi^{-1}\left(\frac{1}{N+1}\sum_{i'=1}^{N}\mathbf{1}\left(Y_{i'j} \leqslant Y_{ij}\right)\right) \tag{17}$$

where $\Phi(\cdot)$ is the normal distribution function. For more on this and other methods, see Quinn (2004); Hoff (2007); Murray et al. (2013).

The probabilistic PCA model allows us to recover point estimates even when there are more features than observations, causing existing common factor analytic implementations to fail due to a rank deficiency. This data structure is unavoidable in text data, where word features may greatly outnumber units of observation. A second approach, Weighted Multidimensional Scaling (WMDS) (Borg et al., 2013; Borg and Groenen, 2005), returns a common index across several data sets, with a measure of how much information each dataset contributes to the common index. MD2S optimally combines the two datasets to extract a common factor, as in WMDS, as well as idiosyncratic factors, allowing the researcher to estimate the location of features along each estimated scale. Doing so greatly aids interpretation, since we can use both the observations (e.g., legislators) and their features (e.g.,words) to summarize the dimensions.

Lastly, we wish to qualify how the p-values should be incorporated into the process of interpretation.

Our permutation test offers a precise, but incomplete, measure of uncertainty; see Mair, Borg and Rusch (2016, esp. 778–779) for recent work on the topic.[9] We advocate three different criteria for ascertaining whether an uncovered dimension is systematic. First is the p-value. If a dimension is not easily distinguished from noise, it should not be favored. This, of course, is necessary but not sufficient. The second criterion we recommend is substantive significance, namely the proportion of the observed variance explained by the method. The third criterion is whether the dimension has face validity. Every positive p-value threshold leaves open the possibility of recovering a noise dimension, so the particular threshold should be selected based off the researcher's tolerance of false positives. We follow convention and implement a threshold of 0.1 below, but do so while emphasizing that the final elements of evaluating the recovered dimensions rely crucially on substantive understanding. If the three criteria we list lead to differing conclusions on whether a dimension is systematic, there is no perfect answer: we advocate that the researcher carefully and transparently adjudicate amongst the criteria, bringing theory and existing results into the process.

4 Simulation Study

In order to assess the proposed method, we conduct a simulation study which tests MD2S across two different elements: first, its ability to identify common and idiosyncratic factors, as well as its ability to distinguish systematic dimensions from noise.

4.1 Simulation Setup

The observed data consist of matrices $Y_{(1)}$ and $Y_{(2)}$ with N rows and K_1 and K_2 columns respectively. N is varied along $\{20, 50, 100\}$ and K_2 along $\{20, 100, 250, 500, 1000, 2500, 5000\}$. K_1 is held at 40. The

[9] We note that these authors advocate comparing the mean dimension weight under the null to the observed value rather than calculating a proper p-value; see, e.g., their Figure 1.

data are generated as

$$Y_{(1)} = 2Z_{S1}W_{(1)1}^{\top} + Z_{S2}W_{(1)2}^{\top} + 4Z_{(1)1}B_{(1)1}^{\top} + 2Z_{(1)2}B_{(1)2}^{\top} + \Omega_{(1)} \tag{18}$$

$$Y_{(2)} = 2Z_{S1}W_{(2)1}^{\top} + Z_{S2}W_{(2)2}^{\top} + 4Z_{(2)1}B_{(2)1}^{\top} + 2Z_{(2)2}B_{(2)2}^{\top} + 2Z_{(2)3}B_{(2)3}^{\top} + \Omega_{(2)} \tag{19}$$

There are two shared dimensions, one twice the size of the other. The matrix $Y_{(m)}$ has two idiosyncratic dimensions and $Y_{(2)}$ has three. All systematic factors $Z_{\cdot\cdot}$, $W_{(m)\cdot}$, and $B_{(m)\cdot}$ are drawn from a standard normal. The error matrices $\Omega_{(m)}$ are scaled such that the systematic component has twice the standard error of the random component, i.e. the true R^2 is $\left(2/(2+1)\right)^2 = 4/9 \approx 0.44$. All simulations were run 1,000 times.

We designed this simulation with two goals in mind. First, we wanted the common factor in Z_S to not be the largest systematic factor of $Y_{(1)}$ and $Y_{(2)}$. Uncovering the common factor involves avoiding the idiosyncratic factors. Second, we wanted to have more variables than observations in one of the matrices. We did so to mimic text data, where we have more terms than observations and regular factor analysis is computational infeasible.

We compare our proposed algorithm to two additional methods that are able to recover a shared scale from multiple datasets. First, we use a variational approximation of the Bayesian Inter-Battery Factor Analysis (V-BIBFA) model of Klami, Virtanen and Kaski (2013). The data generating process behind V-BIBFA is the same as ours, which is based on a linear latent variable model. In contrast to MD2S, V-BIBFA targets the factors or linear projections $W_{(m)}$ and $B_{(m)}$ instead of the latent factors Z_S and $Z_{(m)}$. This is done by placing a sparse prior over the linear projections in order to separate a shared linear mapping $W_{(m)}$ from a specific one for each dataset m, $B_{(m)}$. Given an estimated posterior distribution of factors, scaled locations can be recovered from a normal posterior.

We also compare MD2S to another scaling approach, weighted multidimensional scaling or "individual differences scaling" (INDSCAL) as implemented in the **R** library `smacof`. Instead of focusing

on the scaling of two matrices of size N by K_1 and N by K_2 as done by MD2S, INDSCAL recovers a shared scale from two matrices of dissimilarities of size N by N instead. First, a scale is recovered for each individual dataset and a matrix of weights is estimated to map this individual scales into a shared subspace.[10] We use the Manhattan distance (L1 norm) as the measure of dissimilarity between the rows of $Y_{(1)}$ and $Y_{(2)}$ to reduce them to square matrices of size N by N. In contrast to MD2S, INDSCAL does not directly return shared and idiosyncratic variation. In order to extract data-specific scales orthogonal to the shared subspace, we first estimate the shared latent dimension with the INDSCAL procedure. Next, with a linear mapping we partial this scale out from the K_m outcomes in each original data matrix, $Y^*_{(m)}$. Finally, each partialed out dataset is transformed into a squared dissimilarity matrix and scaled via metric multidimensional scaling, as implemented in the **R** library `smacof`.

4.2 Results

Our primary interest is in recovering a scaling informed by both sets of datasets. Figure 1 presents the results comparing estimates of the shared subspace and the true shared subspace. The figure is organized with sample size in columns ($N \in \{20, 50, 100\}$) and the number of outcomes in $Y_{(2)}$, $K_2 \in \{20, 100, \ldots, 5000\}$ in rows. The x-axis ranges from 0 to 1 and measures the correlation between the true and estimated values. The y-axis is a density scale. The methods presented include Multi-Dataset Multidimensional Scaling (MD2S), the proposed method; the variational Bayesian implementation of Klami, Virtanen and Kaski (2013) (V-BIBFA); the individual differences scaling (INDSCAL); and an estimate that is pure normal noise (Random).

Looking across the columns of figure 1, each method benefits from an increase in sample size and is clearly differentiable from random noise. We see that, across settings, either MD2S or V-BIBFA performs the best in recovering the true shared subspace. Particularly when $N = 20$ and K_2 is less than 1000,

[10] We thank an anonymous reviewer for suggesting this comparison.

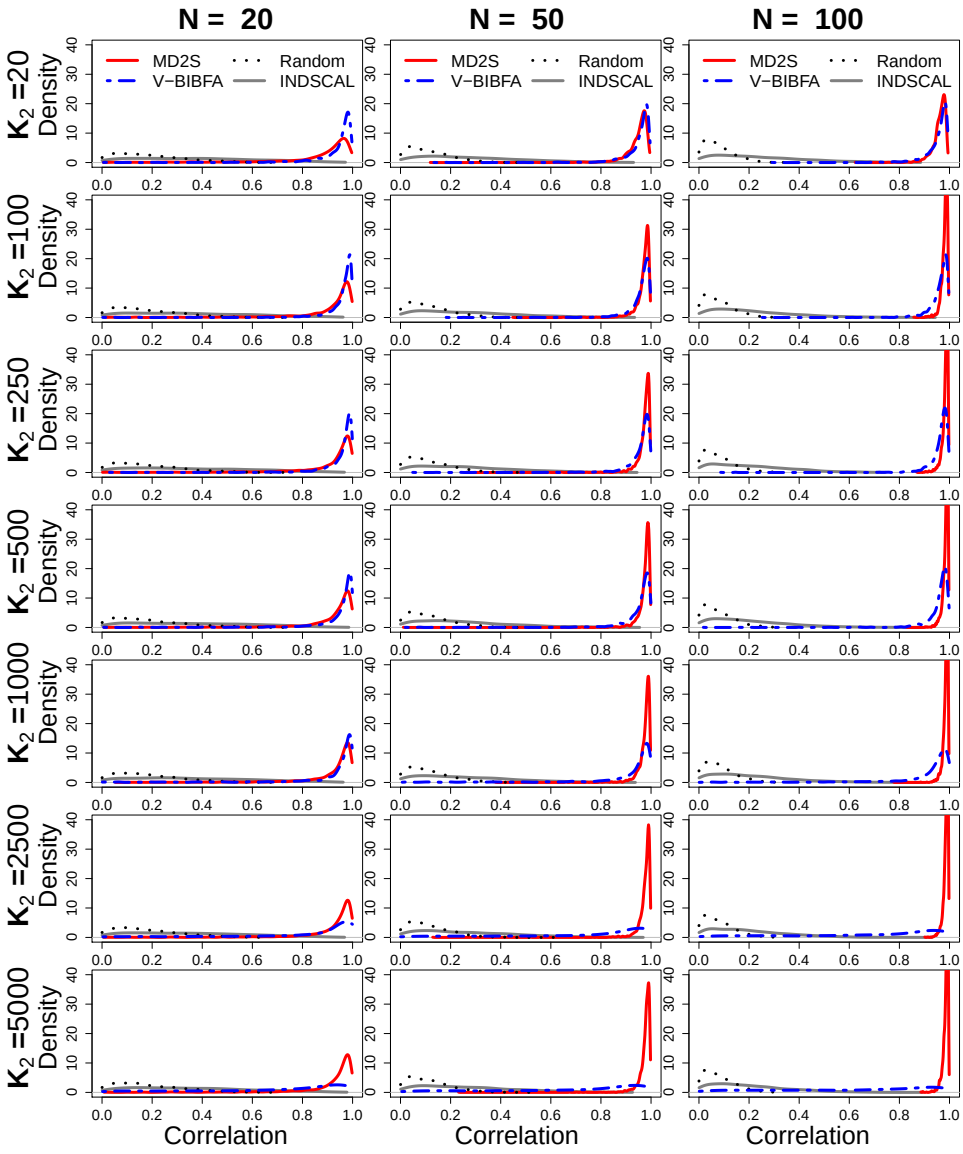


Figure 1: **Correlation between Common Subspace (Z_S) and its Estimate, by Method.** Sample size is in columns ($N \in \{20, 50, 100\}$) and $K_2 \in \{20, 100, \ldots, 5000\}$ is in rows. The x-axis ranges from 0 to 1 and measures the correlation between the true and estimated values. The proposed method, MD2S, is compared to V-BIBFA and INDSCAL, as described in the text. All methods improve in N, but MD2S alone also improves in K_2. The remaining methods deteriorate in K_2. Across settings, either MD2S or V-BIBFA performs the best.

V-BIBFA returns the best estimates of Z_S; the remainder of the time, MD2S performs the best. Note that, as K_2 increases, the estimates in V-BIBFA deteriorate in K_2 while MD2S improves. This is evidence that, with large outcome data matrices, such as the ones generated by text data, our iterative algorithm is able to recover a latent shared subspace that is closer to the true data generating process than available

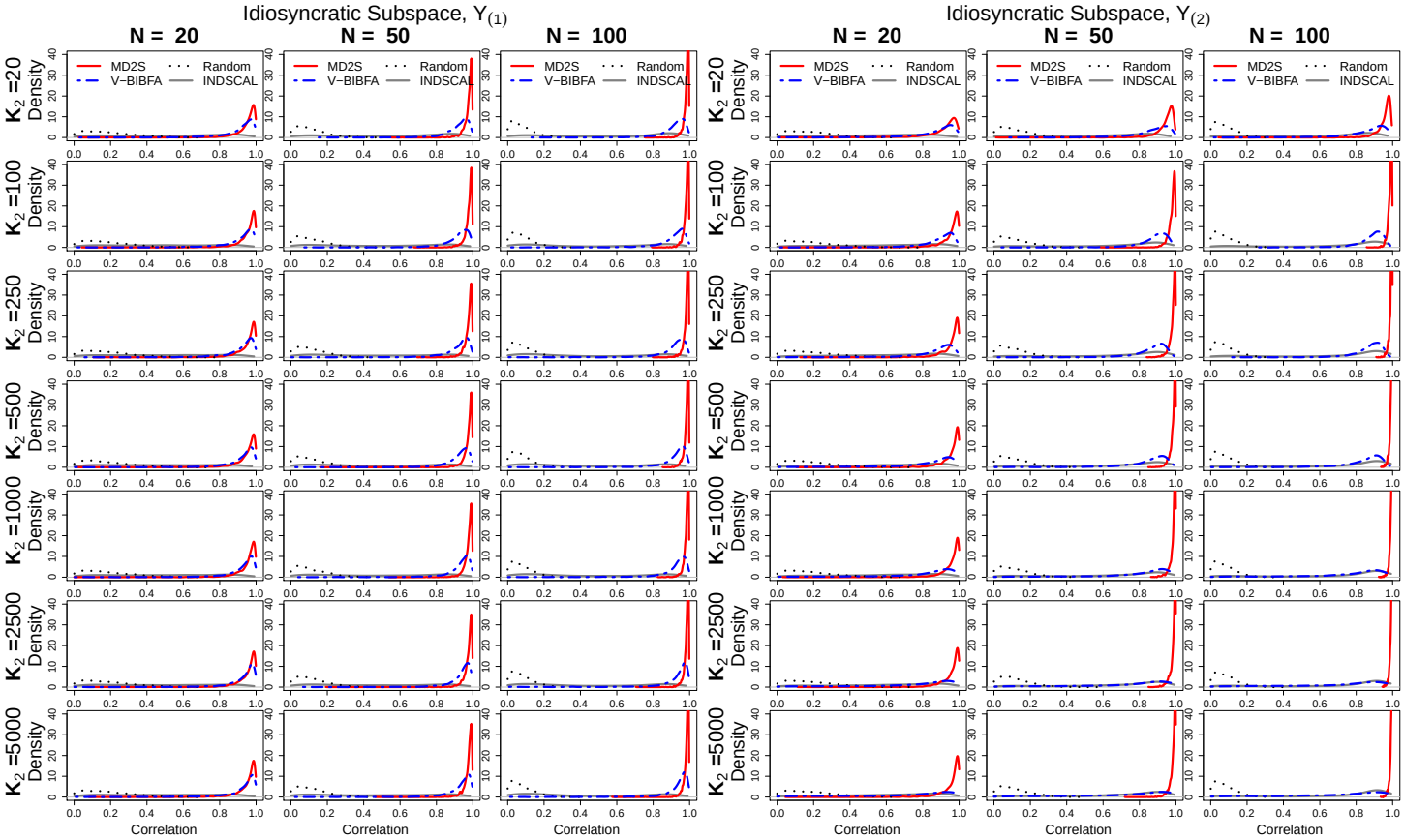


Figure 2: **Correlation between Idiosyncratic Locations for $Y_{(m)}$ and their Estimates, by Method.** The left and right figures are structured identically to Figure 1. Consider the left figure. Looking down rows, MD2S estimates for $Z_{(1)}$ change little with changes in $Y_{(2)}$. Other methods are similarly stable. Looking across columns, as N increases, MD2S and V-BIBFA perform the best and nearly equivalently, though MD2S seems to have a slight edge. Now, consider the right-hand figure. Looking across columns, all methods improve as N increases. As both N and K_2 increase, the performance of V-BIBFA deteriorates, while MD2S improves. For $N \in \{50, 100\}$ and $K > 100$, MD2S is outperforming all other methods in recovering $Z_{(2)}$. When $K_2 \geqslant 1000$ and $N \geqslant 50$, MD2S recovers $Z_{(2)}$ near exactly, and quite a bit better than all other methods.

alternatives. The solid gray line shows that INDSCAL regularly outperforms random noise, but performs notably worse than MD2S and V-BIBFA.

Figure 2 contains the same set of results as Figure 1, but for the idiosyncratic subspaces $Z_{(1)}$ from $Y_{(1)}$ (left), and $Z_{(2)}$ from $Y_{(2)}$ (right). Consider the left-side panel. The number of features in $Y_{(1)}$, K_1, is fixed across simulations, only K_2 is changing. Looking down rows, we see that the MD2S and V-BIBFA results for $Z_{(1)}$ change little with changes in $Y_{(2)}$. This is desirable: since $Z_{(1)}$ is idiosyncratic to $Y_{(1)}$, we

do not want changes in $Y_{(2)}$ to impact its estimate. Overall, INDSCAL's performance is relatively not the best, with an average correlation with $Z_{(1)}$ of just 0.33. Looking across columns, as N increases, we see that MD2S and V-BIBFA perform better, although the correlation between MD2S estimates and the true subspace $Z_{(1)}$ is significantly more precise than for the V-BIBFA results.

Next, consider the right-hand panel. Looking across columns, again, we see all methods except INDSCAL improving as N increases. However, similar to the case of the shared subspace, as K_2 increases, the performance of V-BIBFA deteriorates, while MD2S improves. In fact, MD2S outperforms V-BIBFA in recovering $Z_{(2)}$ across all configurations. Moreover, when $K_2 \geqslant 500$ and $N \geqslant 50$, MD2S recovers $Z_{(2)}$ near exactly and better than all other methods.

4.3 Estimating Dimensions

We next illustrate the proposed method's ability to separate systematic from noise dimensions through the use of the permutation test presented in section 3.4.

Figure 3 presents the results of the permutation test described in section 3.4, for which five dimensions were fit to the shared and idiosyncratic subspaces, with $K_1 = 40$. To evaluate the ability of MD2S to recover the correct number of systematic dimensions per subspace, we use three settings. First, we set $N = 50$ and $K_2 = 100$ in panel (a). Moving from panel (a) to (b), we increased N to 100 but kept K_2 fixed at 100. Finally, moving from panel (b) to (c), we kept $N = 100$, but increase K_2 to 1,000. For each of the above-mentioned settings, 500 permuted datasets were used to estimate the p-value, and 500 total simulations were run.

Across panels, if we classify values below p =0.10 as successful instances of uncovering signal from noise, MD2S consistently recovers the first dimension of the the shared subspace. However, if the number of observations (N) is small as in panel (a), MD2S recovers the second shared dimension, which is not

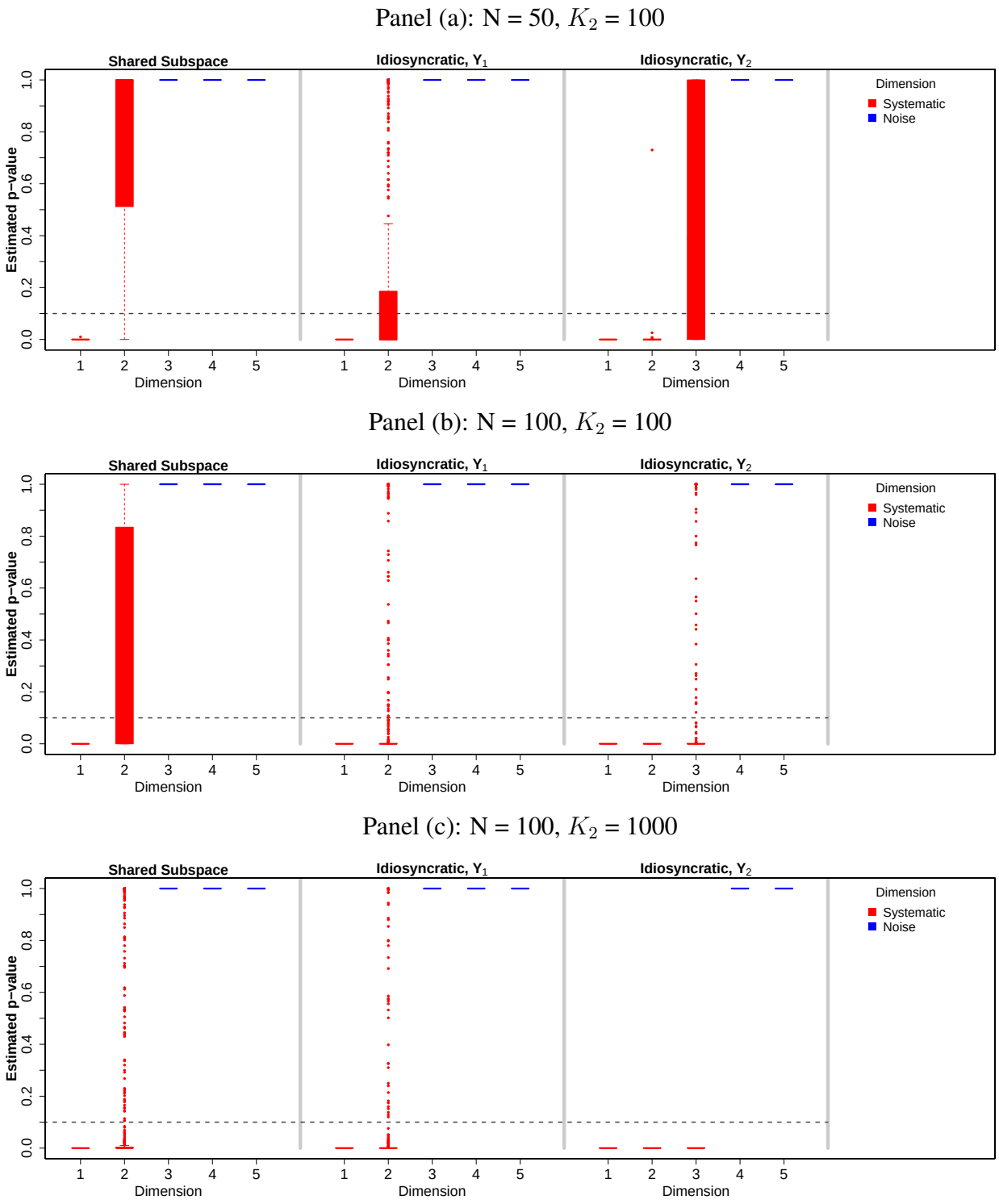


Figure 3: **Estimating Dimensions.** The three panels above present the results from the permutation test for the number of dimensions under three different settings. We use the same data generation process described in section 4, fixing $K_1 = 40$. For panel (a), $N = 50$ and $K_2 = 100$; for panel (b), $N = 100$ and $K_2 = 100$; and for panel (c), $N = 100$ and $K_2 = 1000$. Values in the boxplot that fall below the dotted line at p =0.10 were estimated as systematic dimensions; those above were considered noise. As the three panels show, as we increase both the number of observations N and the number of features (K_2) of the larger data set, MD2S is able to detect the correct dimensionality. Note that for each of the 500 simulations per setting, 500 permuted datasets were used to recover a p-value.

noise, only 17% of the time. Increasing the number of observations, as done in panel (b), improves MD2S' ability to recover the second shared dimension, as it is now detected 53% of the time. If both N and K_2 are increased, as in panel (c), MD2S classifies the second shared subspace as signal 83% of the time.

A similar but less pronounced pattern, is observed for the two idiosyncratic subspaces. The noise dimensions, 3-5 in $Y_{(1)}$ and 4 and 5 in $Y_{(2)}$ are never selected. However, dimensions 2 for $Y_{(1)}$ and 3 for $Y_{(2)}$ which contain systematic information, are difficult to recover for MD2S when both N and K_2 are small. In panel (a), MD2S correctly classifies dimensions 2 of $Y_{(1)}$ and 3 of $Y_{(2)}$ as signal 73% and 46% of the time, respectively. If the number of observations and features of our larger data set are increased, as in panel (c), then MD2S correctly classifies dimension 2 of $Y_{(1)}$ as signal 93% of the time, while dimension 3 of $Y_{(2)}$ is always correctly classified as signal.

5 Combining Senate Roll Call and Text Data

In this empirical exercise, we apply MD2S to data from speech and roll call votes in the 112th U.S. Senate. The data consist of a term document matrix of 2532 unique terms constructed from senators' floor speech found in the *Congressional Record*, as well as the final roll call vote matrix of 486 binary votes taken during this session.[11] The data was previously analyzed using the Sparse Factor Analysis (SFA) methodology introduced in Kim, Londregan and Ratkovic (2018), who found two dimensions in the space jointly informed by words and votes. The primary dimension was qualitatively the same as the

[11] Speech data is collected by the Sunlight Foundation and roll call votes are obtained from VoteView. To create the document-term matrix, senators' floor speech is preprocessed following standard practices by stemming, eliminating stopwords, and analyzing all unigrams and bigrams in the text data. Infrequent terms that are not used by at least ten senators are trimmed. See Kim, Londregan and Ratkovic (2018) for a detailed discussion of the construction of the document-term matrix. Our results are robust to different trimming rules and the inclusion of bigrams (Denny and Spirling, 2018); see Supplemental Appendix B.2 for details.

ideological dimension identified by any common scaling method applied to roll call votes from the U.S. Senate (e.g. Clinton, Jackman and Rivers, 2004, Figure 1). The second dimension was a "leadership" dimension ranging from party leaders, on one end, to rank and file members on the other.

We present the results obtained from MD2S in two parts. First, we use our permutation test to assess which latent dimensions may not be noise. Second, we examine the substance of the scaled locations in the first shared subspace informed by both words and votes, as well as the first idiosyncratic dimensions specific to each type of data. We show the point estimates of the scaled locations for each senator in the results below. Results of the bootstrapped confidence intervals as described in Section 3.4 can be found in the Supplemental Appendix.

We inform the scaled locations with available senators' characteristics. In particular, we allow the latent variables to be a function of senators' `party`, `gender`, and `seniority`. We also account for measures of the number and type of committee assignments of each senator in this session.[12] We include `membership`, which is given by the total number of committees a senator belongs to. The variable `leadership` is a representation of the number of committees where a senator holds a leadership position. The remaining covariates: `agricultural`, `economics`, and `security`, measure the proportion of committees a senator belongs to that deal with these issues.[13] In the Supplemental Appendix, we show the results from regressing the estimated locations for each subspace on this set of covariates.

Estimating dimensionality. Table 1 presents the results from the permutation test presented in section 3.4 applied to the Senate data. The table contains the results for the shared subspace and the two idiosyn-

[12] Senate committee assignments are obtained from Stewart and Woon (1998).

[13] `agricultural` include the committees of Agriculture, Nutrition and Forestry, Energy and Natural Resources, and Environment and Public Works. `economics` include the committees of Appropriations, Banking, Housing, and Urban Affairs, Budget, and Finance. `security` includes the committees of Armed Services and Homeland Security and Governmental Affairs.

cratic dimensions.

		Dimension				
		(1)	(2)	(3)	(4)	(5)
Shared subspace	p-value	0.00	1.00	1.00	1.00	1.00
	%	95.72	1.94	0.89	0.82	0.63
Word subspace	p-value	0.00	0.72	1.00	1.00	1.00
	%	40.87	20.43	15.53	12.13	11.04
Vote subspace	p-value	0.00	1.00	1.00	1.00	1.00
	%	42.54	19.15	14.84	13.08	10.40

Table 1: Permutation test. % represents the percentage of explained variance for each dimension. Using any p-value threshold between 0.01 and 0.71 gives us two dimensions for the shared subspace and three for the idiosyncratic word and votes subspaces.

Using any p-value threshold between 0.01 and 0.71 gives us one statistically significant dimension across the shared and idiosyncratic subspaces. In terms of explained variance, the first shared subspace explains most of the joint variance across votes and words (i.e., 96%). For the idiosyncratic subspaces the first dimension explains 43% and 41% of the variance unique to votes and words, respectively.

Scaled locations. Since we are placing the words and the senators in the same subspace, words at one extreme are most used by legislators at the same extreme. Connecting the words with the legislators greatly aids in interpretation, as we do not only have the locations of the legislators to go by in ascertaining the substantive meaning of the dimension.

We present the scaling estimates of the first shared dimension in Figure 4. On the left panel, we present word clouds containing the top 100 positive (in red) and top 100 negative (in blue) words according to their weights in the estimated matrix of factors for the text data, $\widehat{W}_{(words)}$. The size and color intensity of each word in the wordclouds are proportional to the absolute value of the estimated weights, so words of one color are estimated as near legislators of the same color.

The right panel of figure 4 presents the estimated location of senators in the shared subspace $\hat{Z}_s$.[14]

[14]Section **B.2** in the Appendix presents the estimated locations with their corresponding 95% bootstrapped percentile confi-

The shared subspace differentiates the party, placing Republicans towards the top and Democrats towards the bottom. Our estimates are highly correlated with the SFA ideal point estimates at 0.97. With respect to scaling approaches using only data on roll call votes, our first shared dimension correlates with DW-NOMINATE (Poole, 2005) and IDEAL (Clinton, Jackman and Rivers, 2004) at 0.94 and 0.95, respectively. Therefore, the shared scale is consistent with the ideological dimension uncovered from a spatial vote choice model and from its extension to word choice.[15] These correlations serve as a validation exercise, as DW-NOMINATE and IDEAL have proven to be robust methods to extract information exclusively from roll call votes. Thus, by adding words into the equation in a structured fashion, MD2S is able to recover other interesting patterns while also recovering the expected ideological dimension from the vote data.

The words anchoring each dimension are similar to those identified in Kim, Londregan and Ratkovic (2018, see Figure 3, righthand plot). In particular, we find parliamentary control terms on the side associated with the governing Democratic majority (*meet session*, *author meet*, *conduct hear*) with fiscal terms on the side associated with the Republican minority, (*stimulus*, *trillion*, *budget*, *rais tax*, *debt*) commensurate with the party's professed fiscal concerns. If we move past the parliamentary terms, the first set of substantive terms on the Democratic side are also fiscal in nature but diametrically opposite the Republicans: *wealthiest*, *middleclass*, *tax break*, *tax cut*, and *hear entitl*. Therefore, by recovering the associated terms with each scaled location, we find that the first shared dimension captures well the main differences between Democrats and Republicans in terms of fiscal policy, as identified by the words most associated with each side of the scale.

In terms of idiosyncratic subspaces, we first focus our attention on the vote subspace. As illustrated by Figure 5, we find a significant first dimension that organizes voting, but is unrelated to floor speech.

dence intervals.

[15]The shared scale recovered by V-BIFBA applied to the Senate data correlates with our shared scale at 0.89. The correlation of V-BIFBA with SFA, DW-Nominate and IDEAL is 0.91, 0.91 and 0.95, respectively.

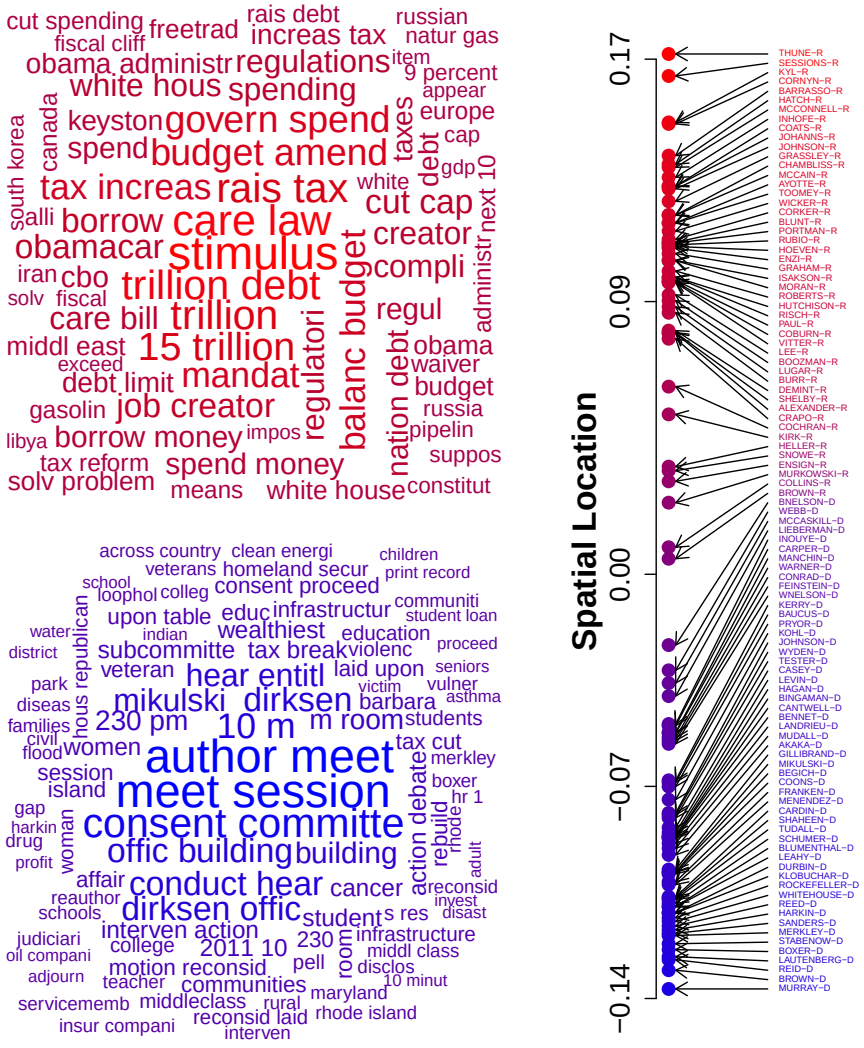


Figure 4: **Shared Subspace Locations Estimated via MD2S for the Members of the 112th U.S. Senate.**

On one extreme, this dimension is anchored by fiscal conservative senators DeMint, Lee, Toomey, Paul, and Risch, noted Tea Party and small government enthusiasts. In terms of the covariates included in the estimation, senators assigned to a leadership position in committees related to `agricultural` and `economic` issues are significantly correlated with this extreme of the scale. The other extreme of the vote subspace is anchored by prominent and more moderate senior senators like Schumer, Boxer, and Collins, who have been reelected at least once and hold a leadership position in a Senate committee. As shown in the appendix, we find that `seniority` and more `leadership` assignments, as well as membership in

committees focused on national `security` issues, are systematically associated with positive locations in the vote subspace.

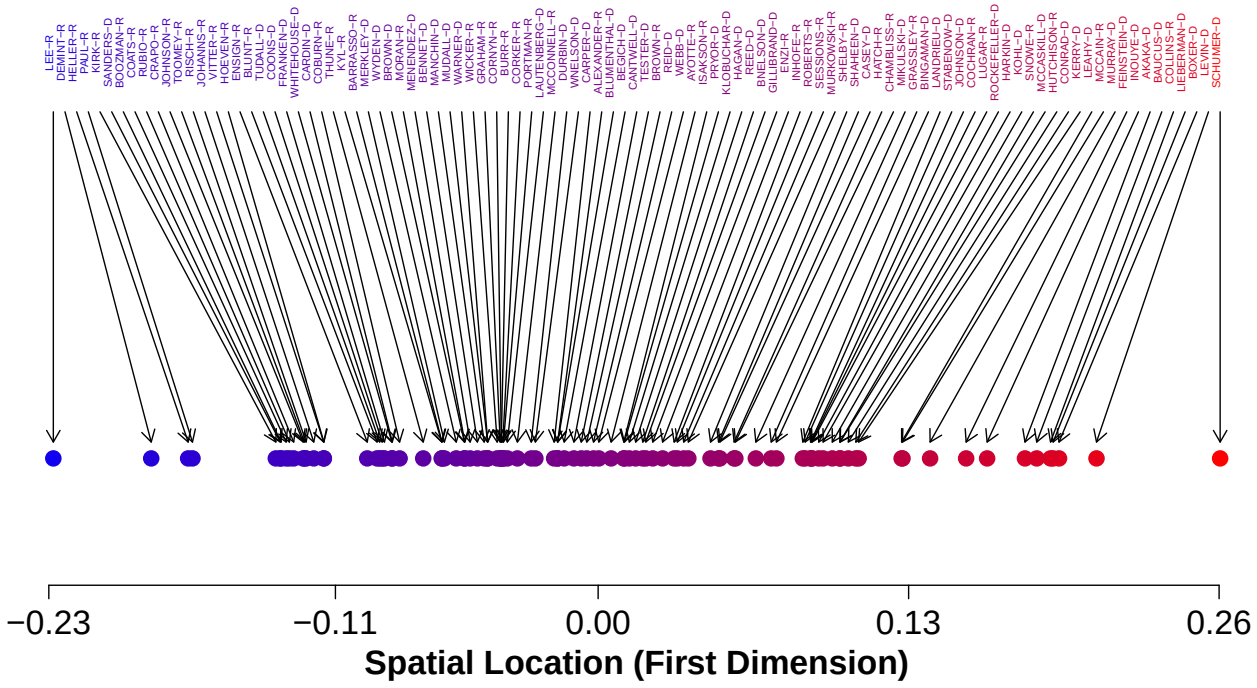


Figure 5: **Idiosyncratic Vote Subspace Locations Estimated via MD2S for the Members of the 112th U.S. Senate.**

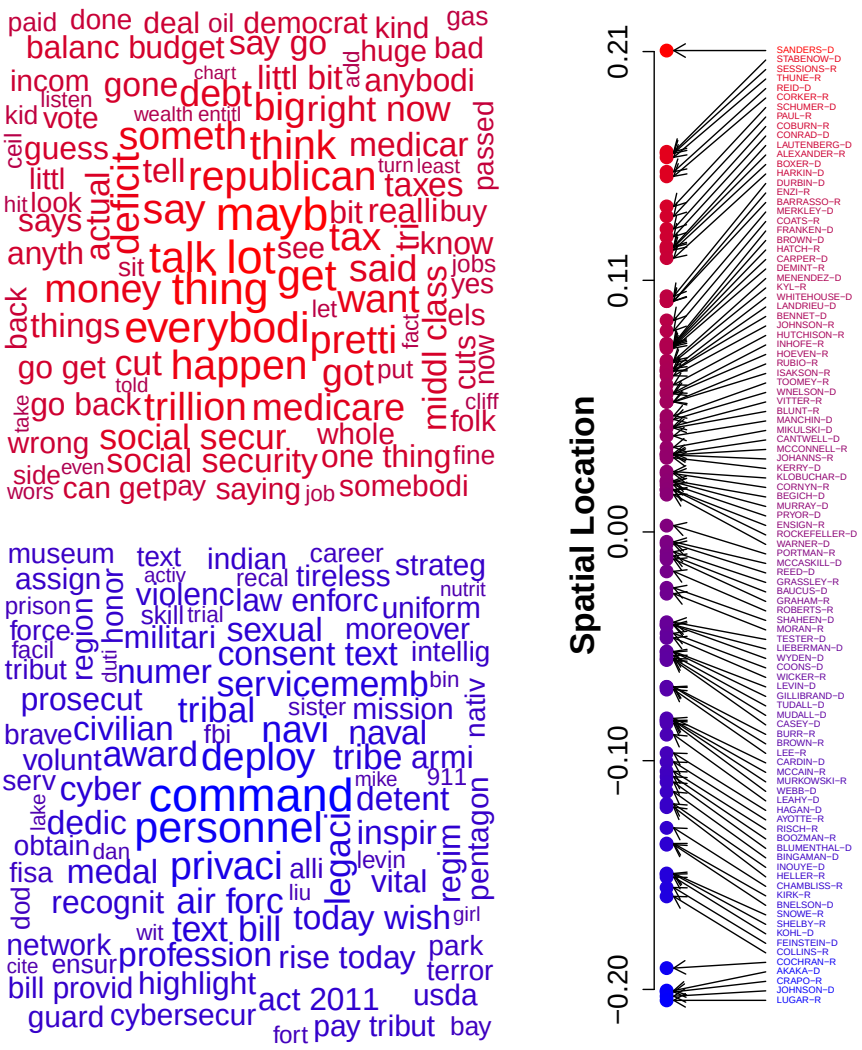


Figure 6: **Idiosyncratic Word Subspace Locations Estimated via MD2S for the Members of the 112th U.S. Senate. First Dimension.**

Similar to the results for the shared subspace, Figure 6 shows senators' locations in the first dimension of the word subspace along with the word clouds of the top 100 words associated with each side of the scale given by the estimated text factor $\widehat{B}_{(words)}$. In terms of the scaled locations, we have on one end, senators who put emphasis on national security issues, such as prominent members of the Committees on Armed Services and Homeland Security, as well as Governmental Affairs like senators Johnson, Akaka, and Collins. The associated terms on this extreme relate to the military (e.g, *command*, *deploy*, *navi*, *air*

forc), as well as personnel and privacy-protection issues (e.g., *privaci*, *personnel*, *civilian*). The opposite side of the words subspace is anchored by senators of both parties (Sanders, Stabenow, Sessions and Thune) addressing budget issues, with associated terms such as *tax*, *deficit*, *debt*, *money* and *medicare*.

6 Conclusion

As we enter a period of "big data," we encourage political scientists to think not just of analyzing large datasets but also how to combine data from disparate sources. We present such a method here, for scaling data from two separate datasets. The method, MD2S, successfully incorporates information from two different data sources, generating scaled locations with a higher internal validity than analyzing the two datasets separately. We include methods for checking validity, separating systematic dimensions from noise, and a way to relate scaled locations to covariates, all fit using an efficient statistical algorithm.

The method also allows the user to use the scaled locations from both datasets to help infer the meaning of the latent dimensions. In our empirical application, scaled locations were also associated with words that let us better interpret the meaning of the latent scale. The idiosyncratic subspaces also offer a new mode of analysis, allowing us to identify ways in which the members at the extremes of the shared subspace differed.

We anticipate several ways in which this project can be moved forwards. First, we have presented the method in a geometric, least squares framework. Placing the method in a probabilistic framework will allow for an extension to commonly used Bayesian techniques (Hare et al., 2015; Tipping and Bishop, 1999). We also plan to extend the method to allow for cross-time comparisons, so as to place multiple observations in the same space over time.

A Proof

We first derive score conditions for the IBFA, extending the model of Tipping and Bishop (1999). We then implement a Minorize-Maximization (MM) algorithm for estimation (for the use of this class of algorithm in scaling, see Borg and Groenen, 2005). The estimation procedure works by deriving a minorizing function that lies weakly below the true objective, maximizes, and iterates to convergence.[16]

The model and likelihood. This section follows Tipping and Bishop (1999). The two data sets, $Y_{(1)}$ and $Y_{(2)}$ are modeled in terms of a shared subspace Z_S as well as dataset specific subspaces, $Z_{(1)}$ and $Z_{(2)}$, such that for column j:

$$Y_{(m)j}|W_{(m)j}, B_{(m)j}, L_{(m)}, D_{(m)} \sim \mathcal{N}(Z_S L_{(m)} W_{(m)j} + Z_{(m)} D_{(m)} B_{(m)j}, \sigma^2_{(m)} I_N) \tag{20}$$

for $m \in \{1, 2\}$. Marginalizing over $W_{(m)}$ and $B_{(m)}$ gives the unconditional densities for $Y_{(m)j}$ as

$$Y_{(m)j} \sim \mathcal{N}(0_N, C_m) \tag{21}$$

where $C_m = Z_S L^2_{(m)} Z_S^\top + Z_{(m)} D^2_{(m)} Z^\top_{(m)} + \sigma_1^2 I_N$

The log-likelihood of $(Z_S, Z_{(1)}, Z_{(2)}, L_{(1)}, L_{(2)}, D_{(1)}, D_{(2)})$ is then

$$\begin{aligned} &l\left(Z_S, Z_{(1)}, Z_{(2)}, L_{(1)}, L_{(2)}, D_{(1)}, D_{(2)}|Y_{(1)}, Y_{(2)}\right) = \\ &-\frac{1}{2}\left\{N(K_1 + K_2)\log(2\pi) - K_1 \log\left(|C_1|\right) - K_2 \log\left(|C_2|\right) - \mathrm{tr}\left(Y_{(1)} Y_{(1)}^\top C_1^{-1} + Y_{(2)} Y_{(2)}^\top C_2^{-1}\right)\right\}. \end{aligned} \tag{22}$$

Denoting $L^2_{(2)} = \lambda^2 L^2_{(1)}$, the score conditions for the shared subspace model are

$$\frac{\partial l(\cdot)}{\partial Z_S} = \left\{\frac{1}{2}\left(K_1 C_1^{-1} + K_2 C_2^{-1}\right) - \frac{1}{2}\left\{C_1^{-1} Y_{(1)} Y_{(1)}^\top C_1^{-1} + \lambda^2 C_2^{-1} Y_{(2)} Y_{(2)}^\top C_2^{-1}\right\}\right\} Z_S L^2_{(1)} \tag{23}$$

$$\frac{\partial l(\cdot)}{\partial Z_{(m)}} = \left\{\frac{1}{2} K_m C_m^{-1} - \frac{1}{2} C_m^{-1} Y_{(m)} Y_{(m)}^\top C_m^{-1}\right\} Z_{(m)} D^2_{(m)} \tag{24}$$

[16] The Q function in the popular EM algorithm is a minorizing function.

It may appear at first that $Z_{(m)}$ and Z_S are solutions to an eigen problem of the form $AZ = \lambda Z$. This does not immediately follow from Equations (23) - (24), though, because $Z_{(m)}$ and $Z_{(S)}$ enter into C_m nonlinearly. The work in the proof below comes from using the identification conditions (equations 6 to 10) and making use of the Woodbury identity to isolate Z_S in C_m^{-1}. With this done, it is apparent that the maximum likelihood estimates are indeed singular vectors. We formalize that result in the Proposition 1, which is given in the text.

Proof of Proposition 1.

1. We proceed in two steps. First, we simplify the term $C_m^{-1} Z_{(m)} D_{(m)}^2$, leaving it a function of only $Z_{(m)}$ and not Z_S. Second, we substitute this simplified term back into the score conditions, showing that $Z_{(m)}$ are singular vectors.

   First, denote $A = (Z_m D_{(m)}^2 Z_{(m)}^\top + \sigma_{(m)}^2 I_n)$ and $U = (L_{(m)}^{-2} + Z_S^\top A^{-1} Z_S)$. Then,

$$\begin{aligned} C_m^{-1} Z_{(m)} D_{(m)}^2 &= C_m^{-1} M(Z_S) Z_{(m)} D_{(m)}^2 && Z_S \perp Z_{(m)} \\ &= \left\{ A^{-1} - A^{-1} Z_S U^{-1} Z_S^\top A^{-1} \right\} M(Z_S) Z_{(m)} D_{(m)}^2 && \text{Woodbury identity to } C_m^{-1} \\ &= (Z_{(m)} D_{(m)}^2 Z_{(m)}^\top + \sigma_{(m)}^2 I_N)^{-1} M(Z_S) Z_{(m)} D_{(m)}^2 \end{aligned}$$

   where the last line follows from distributing and that A^{-1} is not a function of Z_S, leaving the second summand linear in Z_S and therefore annihiliated by $M(Z_S)$. We further simplify through reapplying the Woodbury identity to $(Z_{(m)} D_{(m)}^2 Z_{(m)}^\top + \sigma_{(m)}^2 I_N)^{-1}$:

$$\begin{aligned} &= \frac{1}{\sigma_{(m)}^2} \left\{ I_N - Z_{(m)} \left(\sigma_{(m)}^2 D_{(m)}^{-2} + Z_{(m)}^\top Z_{(m)} \right)^{-1} Z_{(m)}^\top \right\} M(Z_S) Z_{(m)} D_{(m)}^2 \\ &= M(Z_S) Z_{(m)} \widetilde{D}_{(m)} \end{aligned}$$

   where we denote the diagonal matrix $\widetilde{D}_{(m)} = \frac{1}{\sigma_{(m)}^2} \{ I_N - (\sigma_{(m)}^2 D_{(m)}^{-2} + I_{Q_{(m)}})^{-1} \}$. That this matrix is diagonal is crucial to our result, illustrating where the advantage of the PPCA enters our results.

Substituting into the score conditions gives:

$$
\begin{aligned}
& K_m C_m^{-1} Z_{(m)} - \frac{1}{\sigma^2_{(m)}} C_m^{-1} Y_{(m)} Y_{(m)}^\top M(Z_S) Z_{(m)} \widetilde{D}_m = 0_{N \times Q_m} \\
& \Rightarrow K_m Z_{(m)} - \frac{1}{\sigma^2_{(m)}} Y_{(m)} Y_{(m)}^\top M(Z_S) Z_{(m)} \widetilde{D}_m = 0_{N \times Q_m} && \text{Pre-multiply } C_m \\
& \Rightarrow K_m M(Z_S) Z_{(m)} - \frac{1}{\sigma^2_{(m)}} M(Z_S) Y_{(m)} Y_{(m)}^\top M(Z_S) Z_{(m)} \widetilde{D}_m = 0_{N \times Q_m} && \text{Pre-multiply } M(Z_S) \\
& \Rightarrow K_m Z_{(m)} - \frac{1}{\sigma^2_{(m)}} M(Z_S) Y_{(m)} Y_{(m)}^\top M(Z_S) Z_{(m)} \widetilde{D}_m = 0_{N \times Q_m} && M(Z_S) Z_{(m)} = Z_{(m)} \\
& \Rightarrow K_m Z_{(m)} I_{L_m} - \frac{1}{\sigma^2_{(m)}} W_{(m)} Z_{(m)} \widetilde{D}_m = 0_{N \times Q_m},
\end{aligned}
$$

where we define $W_{(m)} \equiv M(Z_S) Y_{(m)} Y_{(m)}^\top M(Z_S)$ in the last line. Considering this last equality columnwise shows that each column of $Z_{(m)}$ is a singular vector of $W_{(m)}$, which was to be shown.

2. Denote $Z_{S|m}$ as the first L_S principal components of $Y_{(m)}^\top M(Z_{(m)})$. To prove part (a), just repeat the proof for point 1 using $M(Z_{(1)})$ and $M(Z_{(2)})$ in equation (23). Then, by a similar argument, the maximum likelihood estimates of Z_S are proportional to singular vectors of a weighted average of $Y_{(1)} Y_{(1)}^\top$ and $Y_{(2)} Y_{(2)}^\top$. To prove part (b), we maximize a minorizing function that lies weakly below the true likelihood function. To generate the minorizing function, note

$$
\begin{aligned}
& Y_{(m)}^\top Y_{(m)} \geqslant \mathbb{E}\left(Y_{(m)}^\top Y_{(m)} | Z_S, Z_{(m)}, \sigma^2_{(m)} \right) = C_{(m)} \\
& \Rightarrow \left(Y_{(m)}^\top Y_{(m)} \right)^{-1} \leqslant C_{(m)}^{-1}
\end{aligned}
$$

with the inequalities meant in a matrix sense. Define

$$
\begin{aligned}
& Q\left(Z_S, Z_{(1)}, Z_{(2)} \middle| Y_{(1)}, Y_{(2)}, C_{(1)}^{-1}, C_{(2)}^{-1} \right) = \\
& \quad - \frac{1}{2} \Big\{ N(K_1 + K_2) \log(2\pi) + K_1 \log\left(|C_1|\right) + K_2 \log\left(|C_2|\right) \\
& \quad + \operatorname{tr}\left(C_2^{-1} Y_{(2)} Y_{(2)}^\top Y_{(1)} Y_{(1)}^\top C_1^{-1} + C_1^{-1} Y_{(1)} Y_{(1)}^\top Y_{(2)} Y_{(2)}^\top C_2^{-1} \right) \Big\}
\end{aligned}
$$

By construction,

$$Q\left(Z_S, Z_{(1)}, Z_{(2)} \middle| Y_{(1)}, Y_{(2)}, C_{(1)}^{-1}, C_{(2)}^{-1}\right) \leqslant l\left(Z_S, Z_{(1)}, Z_{(2)} \middle| Y_{(1)}, Y_{(2)}\right). \tag{25}$$

Following the steps in the proof of part (1), Z_S is clearly proportional to a left singular vector of a weighted average of A and $A^\top$, where $A = Y_{(2)}Y_{(2)}^\top Y_{(1)}Y_{(1)}^\top$. That the maximizer of the minorizing function at convergence is also the ML estimate follows from invariance of the ML estimator.

References

Albert, James H. and Siddhartha Chib. 1993. "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association* 88:669–679.

Aldrich, John and Richard McKelvey. 1977. "A Method of Scaling with Applications to the 1968 and 1972 Presidential Elections." *American Political Science Review* 71(1):111–130.

Anderson, T.W. 1989. "Linear Latent Variable Models and Covariance Structures." *Journal of Econometrics* 41:91–119.

Bach, F. and M. Jordan. 2005. A probabilistic interpretation of canonical correlation analysis. Technical Report 688 Department of Statistics, University of California at Berkeley.

Bafumi, Joseph and Michael Herron. 2010. "Leapfrog Representation and Extremism: A Study of American Voters and Their Members in Congress." *American Political Science Review* 104(3):519–542.

Barbera, Pablo. 2016. "Birds of the Same Feather Tweet Together. Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* Forthcoming.

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

Borg, Ingwer, Groenen, Patrick J.F. and Patrick Mair. 2013. *Applied Multidimensional Scaling*. Springer.

Borg, Ingwer and Patrick J.F. Groenen. 2005. *Modern Multidimensional Scaling:Theory and Applications*. Springer.

Browne, M. W. 1979. "The maximum-likelihood solution in inter-battery factor analysis." *British Journal ofMathematical and Statistical Psychology* 32(75–86).

Clinton, Joshua, Simon Jackman and Doughlas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98:355–370.

Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, with David Altman, Michael Bernhard, M. Steven Fish, Adam Glynn, Allen Hicken, Carl Henrik Knutsen, Kelly McMann, Pamela Paxton, Daniel Pemstein, Jeffrey Staton, Brigitte Zimmerman, Rachel Sigman, Frida Andersson, Valeriya Mechkova and Farhad Miri. 2015. "V-Dem Codebook v5." Varieties of Democracy (V- Dem) Project.

Denny, Matthew J. and Arthur Spirling. 2018. "Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It." *Political Analysis* 26(2):168–189.

Gentzkow, Matthew and Jesse M Shapiro. 2010. "What drives media slant? Evidence from US daily newspapers." *Econometrica* 78(1):35–71.

Goplerud, Max. Forthcoming. "A Multinomial Framework for Ideal Point Estimation." *Political Analysis* .

Groseclose, Tim and Jeffrey Milyo. 2005. "A measure of media bias." *The Quarterly Journal of Economics* 120(4):1191–1237.

Gupta, Sunil Kumar, Dinh Phung, Brett Adams and Svetha Venkatesh. 2011. *Advances in Knowledge Discovery and Data Mining*. Lecture Notes in Computer Science Springer-Verlag chapter "A Bayesian Framework for Learning Shared and Individual Subspaces from Multiple Data Sources", pp. 136–147.

Hahn, P. Richard, Carlos M. Carvalho and James G. Scott. 2012. "A Sparse factor Analytic Probit Model for Congressional Voting Patterns." *Journal of the Royal Statistical Society, Series A* 61(4):619–635.

Hansen, Stephen, Michael McMahon and Andrea Prat. 2014. "Transparency and Deliberation within the FOMC: A Computational Linquistics Approach." Working Paper. Available at http://eprints.lse.ac.uk/58072/.

Hare, Christopher, David A. Armstrong II, Ryan Bakker Royce Carroll and Keith T. Poole. 2015. "Using Bayesian Aldrich-McKelvey Scaling to Study Citizens' Ideological Preferences and Perceptions." *American Journal of Political Science* 59(3):759–774.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2013. *The Elements of Statistical Learning.* 10 ed. New York: Springer-Verlag.

Hobbs, William. 2017. "Pivoted Text Scaling for Open-Ended Survey Responses." Unpublished manuscript.

Hobbs, William R and Margaret E Roberts. 2018. "How sudden censorship can increase access to information." *American Political Science Review* 112(3):621–636.

Hoff, Peter D. 2007. "Extending the Rank Likelihood for Semiparametric Copula Estimation." *The Annals of Applied Statistics* 1(1):265–283.

Jackman, Simon and Shawn Trier. 2008. "Democracy as a Latent Variable." *American Journal of Political Science* 52(1):201–17.

Jacoby, William G. 1986. "Levels of conceptualization and reliance on the liberal-conservative continuum." *The Journal of Politics* 48(2):423–432.

Jacoby, William G. 2009. "Public opinion during a presidential campaign: Distinguishing the effects of environmental evolution and attitude change." *Electoral Studies* 28(3):422–436.

Jacoby, William G. and David A. Armstrong II. 2014. "Bootstrap Confidence Regions for Multidimensional Scaling Solutions." *American Journal of Political Science* 58(1):264–278.

Jessee, Stephen. 2016. "(How) Can We Estimate the Ideology of Citizens and Political Elites on the Same Scale?" *American Journal of Political Science* Forthcoming.

Keele, Luke, Corrine McConnaughy and Ismail White. 2012. "Strengthening the Experimenter's Toolbox: Statistical Estimation of Internal Validity." *American Journal of Political Science* 56(2):484–499.

Kellerman, Michael. 2012. "Estimating Ideal Points in the British House of Commoms Using Early Day Motions." *American Journal of Political Science* 56(3):757–771.

Kim, In Song, John Londregan and Marc Ratkovic. 2018. "Estimating Spatial Preferences from Votes and Text." *Political Analysis.* .

Klami, Arto, Seppo Virtanen and Samuel Kaski. 2013. "Bayesian Canonical Correlation Analysis." *Journal of Machine Learning Research* 14(965–1003).

Ladha, Krishna. 1991. "A Spatial Model of Leglslative Voting with Perceptual Error." *Public Choice* 68:151–74.

Lauderdale, Benjamin and Tom Clark. 2014. "Scaling Politically Meaningful Dimensions Using Texts and Votes." *American Journal of Political Science* Forthcoming.

Mair, Patrick, Ingwer Borg and Thomas Rusch. 2016. "Goodness-of-Fit Assessment in Multidimensional Scaling and Unfolding." *Multivariate Behavioral Research* 51(6):772–789.

Martin, Gregory J and Ali Yurukoglu. 2017. "Bias in cable news: Persuasion and polarization." *American Economic Review* 107(9):2565–99.

Murphy, Kevin P. 2012. *Machine learning: a probabilistic perspective*. MIT press.

Murray, Jared S., David B. Dunson, Lawrence Carin and Joseph E. Lucas. 2013. "Bayesian Gaussian Copula Factor Models for Mixed Data." *Journal of the American Statistical Association* 108(502).

Poole, Keith and Howard Rosenthal. 1997. *Congress: A Political Economic History of Roll Call Voting*. New York: Oxford University Press.

Poole, Keith T. 2005. *Spatial Models of Parliamentary Voting*. Analytical Methods for Social Research Cambridge: Cambridge University Press.

Quinn, Kevin M. 2004. "Bayesian Factor Analysis for Mixed Ordinal and Continuous Responses." *Political Analysis* 12(4):338–353.

Roberts, Molly, Brandon Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Gadarian, Bethany Albertson and David Rand. 2014. "Structural topic models for open ended survey responses." *American Journal of Political Science* 58:1064–1082.

Rockova, V. and E. I. George. 2016. "Fast Bayesian factor analysis via automatic rotations to sparsity." *Journal of the American Statistical Association* 111(516):1608–1622.

Shor, Boris and Nolan McCarty. 2011. "The Ideological Mapping of American Legislatures." *American Political Science Review* 105(03):530–551.

Spence, Ian and John C Ogilvie. 1973. "A table of expected stress values for random rankings in nonmetric multidimensional scaling." *Multivariate Behavioral Research* 8(4):511–517.

Stenson, Herbert H and Ronald L Knoll. 1969. "Goodness of fit for random rankings in Kruskal's nonmetric scaling procedure." *Psychological Bulletin* 71(2):122.

Stewart, Charles III and Jonathan Woon. 1998. "Congressional Committee Assignments, 103rd to 114th Congresses, 1993–2017:Senate, 11/17/2017.".

Tausanovitch, Chris and Christopher Warshaw. 2013. "Measuring Constituent Policy Preferences in Congress, State Legislatures, and Cities." *The Journal of Politics* 75(02):330–342.

Tipping, Michael E. and Christopher M. Bishop. 1999. "Probabilistic Prncipal Component Analysis." *Journal of the Royal Statistcal Society, Series B* 61(3):611–622.

Tucker, Ledyard R. 1958. "An Inter-Battery Method of Factor Analysis." *Psychometrika* 23(2):111–136.

Abstract: 163 Words

Body of Paper: 7109 Words