Evaluating U.S. Electoral Representation with a Joint Statistical Model of Congressional Roll-Calls, Legislative Text, and Voter Registration Data

Zhengming Xing
Criteo Labs
Palo Alto, CA 94301
zh.xing@criteo.com

Sunshine Hillygus
Political Science
Duke University
Durham, NC 27708
hillygus@duke.edu

Lawrence Carin
Department of Electrical and Computer Engineering
Duke University
Durham, NC
lcarin@duke.edu

ABSTRACT

Extensive information on 3 million randomly sampled United States citizens is used to construct a statistical model of constituent preferences for each U.S. congressional district. This model is linked to the legislative voting record of the legislator from each district, yielding an integrated model for constituency data, legislative roll-call votes, and the text of the legislation. The model is used to examine the extent to which legislators' voting records are aligned with constituent preferences, and the implications of that alignment (or lack thereof) on subsequent election outcomes. The analysis is based on a Bayesian formalism, with fast inference via a stochastic variational Bayesian analysis.

CCS CONCEPTS

• **Mathematics of computing → Bayesian nonparametric models**; **Variational methods**; • **Theory of computation → Bayesian analysis**;

KEYWORDS

stochastic variational inference; ideal point; topic model; hierarchical Dirichlet process; matrix factorization; multiplicative gamma process

ACM Reference format:
Zhengming Xing, Sunshine Hillygus, and Lawrence Carin. 2017. Evaluating U.S. Electoral Representation with a Joint Statistical Model of Congressional Roll-Calls, Legislative Text, and Voter Registration Data. In *Proceedings of KDD '17, August 13-17, 2017, Halifax, NS, Canada, ,* 10 pages.
DOI: 10.1145/3097983.3098151

1 INTRODUCTION

One of the fundamental research topics in political science is the extent to which elected officials represent the preferences of the citizens who elect them. Although democratic theorists assume an electoral connection between representatives and their constituents, data limitations have historically made it difficult to empirically evaluate both legislators and the public within the same policy space. A long line of research has estimated the ideological preferences of legislators from their voting records, using an "ideal point" model [8, 16]. Such a model typically assumes each legislator and each piece of legislation can be represented by a point in a one-dimensional latent space. More recently [10, 19, 20, 24–26] have offered approaches for incorporating information beyond roll-call votes. For example, in [10, 24] a latent factor model is proposed to jointly analyze the congressional votes and the legislative text. In [11] the authors improve the model by allowing the ideological position of legislators to vary on specific issues. Further, in [20, 25] a spatio-temporal model is proposed, accounting for the time of the votes and the spatial location of the legislators' districts. However, these methods do not explicitly account for properties of the constituents living within a given electoral district.

Estimating the ideological preferences of a member's ideological district is far more difficult. Some researchers rely on crude proxies such as presidential vote share [7]. More recently, scholars have turned to public opinion polls – often pooling many different national surveys to increase sample sizes [2, 9, 13]. For example, in [13] over 100 surveys are aggregated to estimate state-level ideological preferences. Unfortunately, these works are limited by the relatively small number of survey respondents, which causes inaccuracy in parameter estimation, while also hindering access to finer-scale (district level) constituency information.

Motivated by these challenges, we propose a new scalable Bayesian model to jointly analyze individual-level constituency information, congressional roll-call votes, and associated legislative text. For the constituent information, we leverage a random, de-identified sample of 3 million individuals from the political data vendor Catalist, which collects, maintains, and updates a database with political, demographic, and commercial characteristics on 280 million Americans. Matrix factorization [18] is integrated with the hierarchical Dirichlet process (HDP) [22], yielding a statistical characterization of people living within each US congressional district. Further, a topic model is employed on the text of the legislation. The inferred district-level feature vectors of the people living in each district and the topic distribution on a given piece of legislation are employed to infer roll-call votes. Within the model is a novel component that allows inference of the degree to which a given legislator votes in a manner aligned with the interests of his/her constituents. The

KDD '17, August 13-17, 2017, Halifax, NS, Canada

978-1-4503-4887-4/17/08…$15.00
DOI: 10.1145/3097983.3098151

inferred value of this parameter is examined in the context of the success of the legislator in the next election, yielding a new means to evaluate the relationship between legislative behavior, constituent preferences, and electoral outcomes. To address the massive scale of the constituency data, stochastic variational Bayesian inference [6, 12, 23] is utilized.

While the explicit data considered here are associated with politics, the basic model setup is more general. One may envision trying to assess whether specific individuals, from a region or group with particular demographics, will like/dislike given products. The binary legislative votes are analogous to like/dislike of particular products (here legislation), targeted toward specific people. The text of the legislation is like a document describing the product in question. Given a new product/legislation, with an associated text description, we wish to predict whether it will be liked/disliked by particular people (here, whether legislators will vote yes/no on a new piece of legislation).

2 MODEL CONSTRUCTION

2.1 Data and notation

We jointly analyze congressional roll call votes and constituent information for the $J = 435$ congressional districts across the United States. Individual-level constituent information comes from Catalist, a political data vendor (www.catalist.us). An academic subscription provided a 1% random sample of their database (3 million cases) in 2012, and includes a wide range of demographic, political, and commercial characteristics about each individual. For each (anonymous) individual in the Catalist data, there is an associated vector of attributes, describing personal information, such as race, income, education level and voting-turnout history; these features are mixed, real and binary. Let $\mathbf{X}_j \in \mathbb{R}^{P^r \times N_j}$ denote real-valued attributes for individuals in district $j \in \{1, \ldots, J\}$, where N_j denotes the number of individuals from district j for whom we have Catalist data, and P^r represents the number of real attributes. Let $\mathbf{B}_j \in \{0,1\}^{P^b \times N_j}$ denote the binary attributes for the same individuals. Additionally, we have a series of Congressional votes on pieces of legislation, for legislators elected around the time the Catalist data were collected (we consider roll-call data in 2009-2011). Let $\mathbf{R} \in \{0,1\}^{J \times L}$ denote Congressional roll-call votes on bills reaching the House floor (there are 6% missing votes). Finally, for each piece of legislation, we have the associated text of the bill. The lth piece of legislation is denoted $\boldsymbol{w}_l$, where $\boldsymbol{w}_l \in \mathbb{Z}_+^V$ represents the count of each word in the text (a vector of nonnegative integers), where the vocabulary dimension is V.

2.2 Matrix factorization of constituent data

The matrix of real-valued individual-level data from people in district j is factorized as

$$\mathbf{X}_j = \mathbf{D}^r \boldsymbol{\Lambda}^r \mathbf{S}_j^r + \mathbf{E}_j^r, \tag{1}$$

where $\mathbf{D}^r \in \mathbb{R}^{P^r \times K^r}$, $\mathbf{S}_j^r \in \mathbb{R}^{K^r \times N_j}$, $\boldsymbol{\Lambda}^r = \text{diag}(\lambda_1^r, \ldots, \lambda_{K^r}^r)$, and $\mathbf{E}_j^r \in \mathbb{R}^{P^r \times N_j}$. Each column of $\mathbf{E}_j^r$ is drawn from $\mathcal{N}(0, \sigma_j^{-1}\mathbf{I})$ and a diffuse gamma prior is placed on σ_j, *i.e.*, $\text{Ga}(10^{-6}, 10^{-6})$. Note that $\mathbf{D}^r$ and $\boldsymbol{\Lambda}^r$ are shared for all districts j. Each column of $\mathbf{D}^r$ is drawn from $\mathcal{N}(0, \mathbf{I}_{P^r})$, where $\mathbf{I}_{P^r}$ is the $P^r \times P^r$ identity matrix. We wish to impose that $|\lambda_k^r|$ decreases as index k increases; hence, while we truncate the model to K^r factors, through the λ_k^r we infer the subset of factors that are needed to represent the data. To achieve this, we employ the multiplicative gamma process (MGP) proposed in [4]: $\lambda_k^r \sim \mathcal{N}(0, 1/\tau_k^r)$, $\tau_k^r \sim \prod_{h=1}^k \varphi_h^r$, and $\varphi_h^r \sim \text{Ga}(a_1, 1)$. By choosing $a_1 > 1$, $\mathbb{E}(\varphi_h^r) > 1$, encouraging τ_k^r to increase with k; this in turn results in increasing encouragement of shrinking the amplitude of λ_k^r as k increases.

For the observed matrix of binary data for people in district j, $\mathbf{B}_j$, we employ a probit model, and a latent $\tilde{\mathbf{B}}_j \in \mathbb{R}^{P^b \times N_j}$ [1]. Let $\tilde{b}_{jpn}$ be element (p, n) in $\tilde{\mathbf{B}}_j$ and let b_{jpn} represent element (p, n) in $\mathbf{B}_j$; these are related via the probit link: $b_{jpn} = 0$ if $\tilde{b}_{jpn} + \epsilon_{jpn}^b \geq 0$, and $b_{jpn} = 1$ if $\tilde{b}_{jpn} + \epsilon_{jpn}^b < 0$, where $\epsilon_{jpn}^b \sim \mathcal{N}(0, 1)$. We factorize the latent matrix as $\tilde{\mathbf{B}}_j = \mathbf{D}^b \boldsymbol{\Lambda}^b \mathbf{S}_j^b$, where $\mathbf{D}^b \in \mathbb{R}^{P^b \times K^b}$ and $\mathbf{S}_j^b \in \mathbb{R}^{K^b \times N_j}$. The columns of $\mathbf{D}^b$ are drawn with the same class prior as employed above for $\mathbf{D}^r$, and the MPG prior is employed for $\boldsymbol{\Lambda}^b = \text{diag}(\lambda_1^b, \ldots, \lambda_{K^b}^b)$.

2.3 Clustering the constituency latent features

Individual n sampled from district j is characterized by the nth column of $\mathbf{S}_j^r$ and $\mathbf{S}_j^b$. Assuming that people are likely clustered with respect to the attributes included in the Catalist database, we develop a joint mixture model for the columns of $\mathbf{S}_j^r$ and $\mathbf{S}_j^b$. Let $\boldsymbol{s}_{jn}^r$ and $\boldsymbol{s}_{jn}^b$ denote the nth columns of $\mathbf{S}_j^r$ and $\mathbf{S}_j^b$, respectively. We impose the following hierarchical Dirichlet process (HDP) [22] model:

$$\begin{aligned} \boldsymbol{s}_{jn}^r \sim f(\boldsymbol{\theta}_{jn}),\ \boldsymbol{s}_{jn}^b \sim f(\boldsymbol{\psi}_{jn}),\ \{\boldsymbol{\theta}_{jn}, \boldsymbol{\psi}_{jn}\} \sim G_j, \\ G_j \sim \text{DP}(\kappa, G_0),\ G_0 \sim \text{DP}(\kappa_0, H) \end{aligned} \tag{2}$$

where $H(\boldsymbol{\theta}, \boldsymbol{\psi}) = H_r(\boldsymbol{\theta}) H_b(\boldsymbol{\psi})$, and therefore $G_0 = \sum_t \nu_t \delta_{(\boldsymbol{\theta}_t^*, \boldsymbol{\psi}_t^*)}$, with $\nu_t > 0$, $\sum_t \nu_t = 1$ and $\delta_{(\boldsymbol{\theta}_t^*, \boldsymbol{\psi}_t^*)}$ a unit point measure concentrated at the pair $(\boldsymbol{\theta}_t^*, \boldsymbol{\psi}_t^*)$. The distribution $f(\cdot)$ here corresponds to multivariate Gaussian, and H_r and H_b are each Normal-Wishart distributions. Diffuse gamma priors are placed on κ and κ_0. We employ the stick-breaking representation [21] of the HDP developed in [22] and a point estimate of $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots)^T$ [6, 14] to simplify the variational derivations (discussed in Section 3). The number of components ("sticks") used to approximate G_0 and each of the G_j is truncated to T. Each district j is characterized by $G_j = \sum_{t=1}^T \pi_{jt} \delta_{(\boldsymbol{\theta}_t^*, \boldsymbol{\psi}_t^*)}$. The "atoms" $\{\boldsymbol{\theta}_t^*, \boldsymbol{\psi}_t^*\}$ are shared across all J districts, and hence the jth district is distinguished by the probability vector $\boldsymbol{\pi}_j = (\pi_{j1}, \ldots, \pi_{jT})^T$.

2.4 Modeling the text of legislation

Consider a corpus of L pieces of legislation, voted on during a Congressional session. A probability vector $\boldsymbol{\beta}_l$ is inferred to represent the lth piece of legislation. Specifically, we employ a basic topic model, latent Dirichlet allocation (LDA) [5] to model each of the L documents, from which we constitute $\boldsymbol{\beta}_l$, a probability vector over topics (assumed here to be truncated to K topics). Topic $k \in \{1, \ldots, K\}$ is characterized by a V-dimensional probability vector $\boldsymbol{\phi}_k$, and a word from document/legislation l is associated with

topic k with probability β_{lk}. If a word is drawn from topic k, the specific word is drawn Mult$(1, \boldsymbol{\phi}_k)$ [5].

The vote of the jth legislator on bill l is modeled in terms of $\boldsymbol{\pi}_j$ and $\boldsymbol{\beta}_l$, coupling the constituency data and the text of legislation to predict roll-call votes. Rather than predicting roll-call votes directly based on $\boldsymbol{\pi}_j$ and $\boldsymbol{\beta}_l$ (the doing of which significantly complicates inference), we introduce surrogates for $\boldsymbol{\pi}_j$ and $\boldsymbol{\beta}_l$ [15]. Specifically, individual $n \in \{1, \ldots, N_j\}$ in district j has an associated latent variable $c_{jn} \in \{1, \ldots, T\}$, identifying which model parameters $(\boldsymbol{\theta}^*_{c_{jn}}, \boldsymbol{\psi}^*_{c_{jn}})$ are used for his/her representation. This assigns individual n in district j to a cluster, with cluster t characterized by $(\boldsymbol{\theta}^*_t, \boldsymbol{\psi}^*_t)$. The VB analysis yields the expected probability of which of the T clusters person n in district j is associated with, this probability vector denoted $\tilde{\boldsymbol{\pi}}_{jn}$.

Similarly, we introduce latent variable $z_{il} \in \{1, ..., K\}$, assigning a topic to word i in document l. Within the VB inference of LDA, we manifest $\tilde{\boldsymbol{\beta}}_{li}$, the expected probability vector for which topic word i in document l is associated with. We predict the roll call vote associated with district j for legislation l in terms of the two probability vectors $\tilde{\boldsymbol{\pi}}_j = \frac{1}{N_j} \sum_{n=1}^{N_j} \tilde{\boldsymbol{\pi}}_{jn}$ and $\tilde{\boldsymbol{\beta}}_l = \frac{1}{W_l} \sum_{i=1}^{W_l} \tilde{\boldsymbol{\beta}}_{li}$, assuming W_l total words in document l.

2.5 Coupling constituency characteristics and legislative text: Roll-call analysis

Like for the binary attributes $\mathbf{B}_j$ discussed above, for the binary roll-call votes we assume a latent matrix $\tilde{\mathbf{R}} \in \mathbb{R}^{J \times L}$ which we factorize as $\tilde{\mathbf{R}} = \mathbf{D}^\ell \boldsymbol{\Lambda}^\ell \mathbf{S}^\ell + \mathbf{E}^\ell$. The MPG prior is imposed for the elements of the diagonal matrix $\boldsymbol{\Lambda}^\ell$.

Row j of $\mathbf{D}^\ell$, denoted by the column vector $\boldsymbol{d}^\ell_j$, is a feature vector associated with district j, from the standpoint of voting on legislation. The lth column of $\mathbf{S}^\ell$, denoted by the column vector $\boldsymbol{s}^\ell_l$, is similarly a feature vector for legislation l (from the standpoint of how the text affects the voting). We connect the voting characteristics of the legislators from district j to the constituency characteristics of his/her district by modeling $\boldsymbol{d}_j$ in terms of $\tilde{\boldsymbol{\pi}}_j$. Similarly, we connect votes to the properties (text) of the legislation by modeling $\boldsymbol{s}^\ell_l$ in terms of $\tilde{\boldsymbol{\beta}}_l$. Specifically, we impose the models

$$\boldsymbol{d}^\ell_j = \mathbf{U}^d \tilde{\boldsymbol{\pi}}_j + \boldsymbol{d}^\ell_0 + \boldsymbol{\xi}_j \,, \; \boldsymbol{s}^\ell_l = \mathbf{U}^s \tilde{\boldsymbol{\beta}}_l + \boldsymbol{s}^\ell_0 \,, \tag{3}$$

where $\mathbf{U}^d \in \mathbb{R}^{K^\ell \times T}$, $\boldsymbol{\xi}_j \in \mathbb{R}^{K^\ell}$, $\boldsymbol{d}^\ell_0 \in \mathbb{R}^{K^\ell}$, $\mathbf{U}^s \in \mathbb{R}^{K^\ell \times K}$ and $\boldsymbol{s}^\ell_0 \in \mathbb{R}^{K^\ell}$. The elements of $\mathbf{U}^d$, $\boldsymbol{d}^\ell_0$, $\mathbf{U}^s$ and $\boldsymbol{s}^\ell_0$ and are drawn i.i.d. from, respectively, $\mathcal{N}(0, \alpha_d^{-1})$, $\mathcal{N}(0, \alpha_{d0}^{-1})$, $\mathcal{N}(0, \alpha_s^{-1})$ and $\mathcal{N}(0, \alpha_{s0}^{-1})$, with diffuse gamma priors on α_d, α_{d0}, α_s and α_{s0}.

The vector $\boldsymbol{\xi}_j$ is employed to identify legislators who may be voting against the interests of their constituents, as defined by the attributes in the Catalist database. Since it is hoped that most of $\boldsymbol{d}^\ell_j$ is captured by these features, we impose a prior on $\boldsymbol{\xi}_j$ that encourages (near) sparsity. Therefore, we impose the hierarchical shrinkage prior $\xi_{jk} \sim \mathcal{N}(0, \alpha_{jk}^{-1})$, $\alpha_{jk} \sim$ InvGa$(1, \gamma_{jk}/2)$, $\gamma_{jk} \sim$ Ga$(10^{-6}, 10^{-6})$.

The matrix $\mathbf{E}^\ell \in \mathbb{R}^{J \times L}$ models "random effects." Let E^ℓ_{jl} represent component (j, l) of $\mathbf{E}^\ell$. We impose $E^\ell_{jl} = \delta_l + \delta_{jl}$, where δ_l is a random effect associated with legislation l and δ_{jl} is a random effect associated with the legislation-legislator pair. We further connect δ_l to the legislative text by modeling it in terms of $\tilde{\boldsymbol{\beta}}_l$: $\delta_l = \boldsymbol{w}^T \tilde{\boldsymbol{\beta}}_l + w_0$, where $\boldsymbol{w} \in \mathbb{R}^K$ and $w_0 \in \mathbb{R}$ are i.i.d draw from $\mathcal{N}(0, \alpha_w^{-1})$ and $\mathcal{N}(0, \alpha_{w0}^{-1})$. Diffuse gamma prior is placed on α_w and α_{w0}. There are ceremonial pieces of legislation, for which every legislator tends to vote "yes," and for such legislation δ_l tends to be large and positive. There are also pieces of legislation l for which the jth legislator may vote idiosyncratically, for which δ_{jl} may be large negative or positive (meaning that legislator votes uncharacteristically "no" or "yes," respectively). We don't assume a random effect δ_j, which would imply that the jth legislator tends to always vote one way ("yes" or "no"), *independent* of the legislation.

We expect $\{\delta_{jl}\}$ to be sparse (or nearly sparse), and therefore on each we impose a shrinkage prior (in the same hierarchical manner discussed above for $\boldsymbol{\xi}_j$). We could impose similar random effects on the demographic data model, for representation of $\tilde{\mathbf{B}}_j$, but this proved unnecessary, as there we were model binary traits (*e.g.*, gender), rather than votes.

2.6 Model summary

Figure 1 provides a graphical representation of the model, with shaded and unshaded nodes indicating observed and latent variables, respectively. To assist with understanding the multiple components of the model, and their motivations, we provide an overarching summary below.

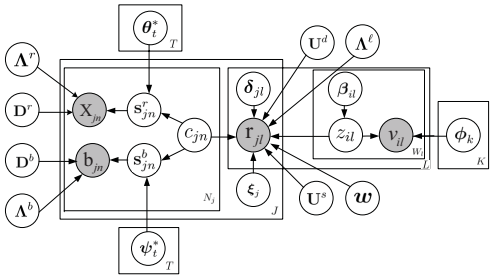


Figure 1: Graphical representation of the model.

The demographic data from distict j are represented by matrix factorizations (factor analysis), where column n of the factor-score matrices $\mathbf{S}^r_j$ (real data) and $\mathbf{S}^b_j$ (binary data) characterize person n in district j. For both matrix factorizations, the multiplicative gamma process is employed to encourage that only a relatively small number of factors are expected to define person choices.

We assume that the people (columns of $\mathbf{S}^r_j$ and $\mathbf{S}^b_j$) in each district will cluster into types of preferences. A truncated HDP is employed to infer this clustering. The probability of each of the T clusters is represented for district j by probability vector $\boldsymbol{\pi}_j$; $\{\boldsymbol{\theta}^*_t, \boldsymbol{\psi}^*_t\}_{t=1,T}$ represent the cluster-dependent parameters.

The vote $r_{jl} \in \{0, 1\}$ of congressman j on legislation l is characterized, via a probit matrix factorization, as an inner product between a feature vector for legislator j, $\boldsymbol{d}^\ell_j$, and a feature vector for legislation l, $\boldsymbol{s}^\ell_l$. To infer the relationship between how the congressman from district j votes relative to the interests of her/his constituents, we relate $\boldsymbol{d}^\ell_j$ to $\boldsymbol{\pi}_j$ via linear regression. We similarly

wish to relate feature vector legislation $\boldsymbol{s}_l^\ell$ to the text of the associated legislation; in this case a regression is performed between $\boldsymbol{s}_l^\ell$ and $\boldsymbol{\beta}_l$, the latter the text-dependent distribution over topics (inferred here for simplicity via LDA, but any topic model may be used).

A key novelty of the model is a term $\boldsymbol{\xi}_j$, constituting a "random effect" in the regression between $\boldsymbol{\pi}_j$ and $\boldsymbol{d}_j^\ell$; $\boldsymbol{\xi}_j$ allows inference of the degree to which the congressman from district j appears to vote in a manner inconsistent with the preferences of her/his constituents. A random effect δ_l also allows identification of atypical legislation, linked to the text of the legislation via $\boldsymbol{\beta}_l$.

The regressions above were discussed in terms of $\boldsymbol{\pi}_j$ and $\boldsymbol{\beta}_l$. For technical reasons, discussed in the preceding sections, it is significantly more convenient to employ closely related surrogates $\tilde{\boldsymbol{\pi}}_j$ and $\tilde{\boldsymbol{\beta}}_l$; these are defined in terms of the relative counts of indicator variables c_{jn} and z_{il}, for person n in district j, and word i in document/legislation l.

3 SCALING UP: VARIATIONAL BAYES AND STOCHASTIC GRADIENT DESCENT INFERENCE

The Catalist data considers 2,969,925 people, and to handle data of this size we employ a mini-batch-based inference algorithm, stochastic variational Bayesian (VB) analysis [6, 12, 23]. Unlike traditional VB inference [3], which includes the whole dataset when updating the parameters, the stochastic variational inference method samples a subset of the data (mini-batch), and calculates a noisy natural gradient to optimize the variational objective function. Specifically, the individuals in the Catalist data are partitioned into $N^* = 15$ mini-batches, and each mini-batch contains individuals from all $J = 435$ congressional districts. The congressional votes and associate text are considered as a whole, since the size of that data is relatively small. The variational parameters specific to each individual mini-batch (in our case, the variational parameters associated with $\{\boldsymbol{s}_{jn}^r, \boldsymbol{s}_{jn}^b, c_{jn}\}$), are called "local" parameters, denoted $\boldsymbol{\Theta}^l$. The remaining variational parameters, not specific to the mini-batch, are called "global" parameters, denoted $\boldsymbol{\Theta}^g$. At the hth iteration, the hth mini-batch is selected, and local variational parameters of the mini-batch $\boldsymbol{\Theta}^l$ are optimized; intermediate global parameters $\tilde{\boldsymbol{\Theta}}^g$ are then estimated with the most recent mini-batch. The new estimated global parameters are updated by computing the weighted average of previous value and $\tilde{\boldsymbol{\Theta}}^g$, $\boldsymbol{\Theta}^g \leftarrow (1-\omega_h)\boldsymbol{\Theta}^g + \omega_h\tilde{\boldsymbol{\Theta}}^g$, where $\omega_h \in (0,1)$ is the weight given to each new batch, and also called the learning rate. Following [12], we let $\omega_h = (a_3 + h)^{-b_3}$, where $b_3 \in (0.5, 1]$ controls the rate of decay of the contribution from old mini-batches and $a_3 \geq 0$ serves to slow down the decay rate for initial iterations. In the experiments, we set $a_3 = 1$ and $b_3 = 0.8$. One may employ the method proposed in [17] to adapt the learning step.

Details of the VB update equations are presented in the Appendix. In the following, we examine two of the update equations, as they provide insight into how different parts of model relate to one another.

Variational Distribution for c_{jn}: The posterior-approximating distribution for the indicator variable c_{jn}, $q(c_{jn})$, is a categorical distribution with parameter $\tilde{\boldsymbol{\pi}}_{jn}$, the components of which satisfy $\tilde{\pi}_{jnt} \propto \exp\{\mathbb{E}[\log p(\boldsymbol{s}_{jn}^r|\boldsymbol{\theta}_t^*)] + \mathbb{E}[\log p(\boldsymbol{s}_{jn}^b|\boldsymbol{\psi}_t^*)] + \mathbb{E}[\log(\pi_{jt})] + \sum_{l=1}^L \mathbb{E}[\log p(\tilde{r}_{jl}|c_{jn} = t, -)]\}$. The term $\mathbb{E}[\log(\pi_{jt})]$ characterizes the clustering characteristics of district j, where $p(\boldsymbol{s}_{jn}^r|\boldsymbol{\theta}_t^*)$ and $p(\boldsymbol{s}_{jn}^b|\boldsymbol{\psi}_t^*)$ characterize the properties of cluster t. The term $p(\tilde{r}_{jl}|c_{jn} = t, -)$ characterizes the latent real matrix associated with the binary legislative votes of the representative from district j on all L pieces of legislation.

Variational Distribution for z_{il}: The approximating distribution for the latent topic associated with word i in legislation l, $q(z_{il})$, is a categorical distribution with parameter $\tilde{\boldsymbol{\beta}}_{il}$, and $\tilde{\beta}_{ilk} \propto \phi_{v_{il},k}\exp\{\mathbb{E}[\log\ \beta_{lk}] + \sum_{j=1}^J \mathbb{E}[\log p(\tilde{r}_{jl}|z_{il} = k, -)]\}$. Note that this update equation is affected by the fit of the word to the topic (first term) plus the impact of that topic to the roll-call votes from the J legislators (second term).

4 EXPERIMENTAL RESULTS

We employ the proposed model on the Catalist data discussed above ($P^r = 28$ and $P^b = 51$; 2,969,925 total people across the $J = 435$ Congressional districts, with typically 5,000 to 7,000 people from each district). The roll-call data are from the 111th US Congress (January 3, 2009 - January 3, 2011), consistent with the time period of the Catalist data. Roll-call votes on a total of $L = 802$ bills are considered. For the text of the bill, we follow the n-gram preprocessing procedure described in [10], and obtain a bag of words with vocabulary size $V = 4743$.

The election for the 112th Congress took place on November 2, 2010, and we used votes on bills in the 111th Congress that occurred before then to examine the party affiliation of each winner of that election, and examine the vote share, relative to the roll-call data.

For model initialization, we first consider each data source separately. For example, we take a subset of the Catalist data, to infer $\mathbf{D}^r$, $\mathbf{D}^b$, $\boldsymbol{\Lambda}^r$ and $\boldsymbol{\Lambda}^b$. Then K-means was performed on the learned latent features for the individuals, to initialize the HDP model. Similarly, LDA was first applied to the legislative text to infer initial topics. The results are repeatable for different related forms of this initialization.

We set $K = 30$, $T = 15$, $K^r = K^b = 20$, $K^l = 10$ and the MGP hyperparameter is $a_1 = 2$. The Catalist data are randomly partitioned into 15 mini-batches, each of size 197,995. We implemented the proposed model in MATLAB, and ran the code on a PC with 8 cores, 3.2GHz CPU, and 128 GB memory. We considered 40 VB iterations per mini batch, and the total computation time for these data was 16 hours.

4.1 Inferred district-level characteristics and Congressional election results

Using the full model, we infer $\mathbb{E}[\boldsymbol{\pi}_j]$, the expected probability of demographic clusters for district j. The characteristics of cluster t may be interpreted by mapping the cluster center $\{\mathbb{E}[\boldsymbol{\theta}_t^{\mu*}], \mathbb{E}[\boldsymbol{\psi}_t^{\mu*}]\}$ back to the original data space $\{\mathbb{E}[\mathbf{D}^r\Lambda^r\boldsymbol{\theta}_t^{\mu*}], \Phi(\mathbb{E}[\mathbf{D}^b\Lambda^b\boldsymbol{\psi}_t^{\mu*}])\}$, where $\Phi(\cdot)$ is the cumulative probability function of standard normal (from the probit model). In Figure 2, we plot $\mathbb{E}[\pi_{tj}]$ of four example clusters, for 432 congressional districts (excluding Alaska and Hawaii).

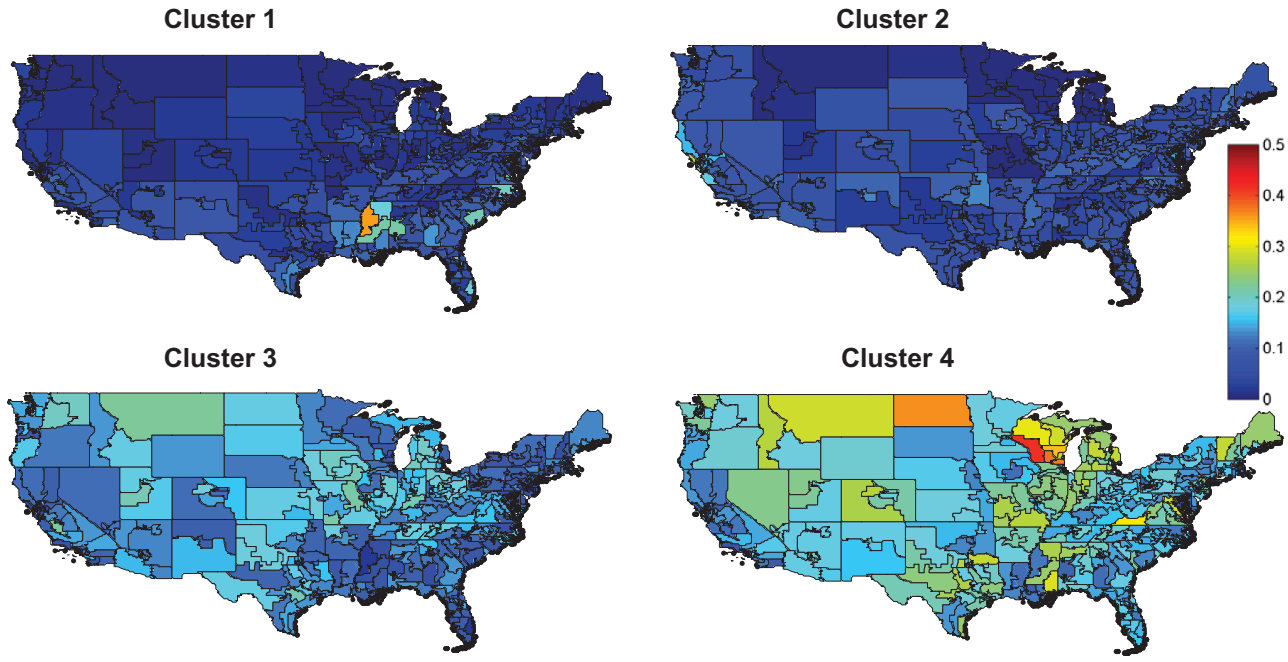


Figure 2: The expected probability of demographic clusters $\mathbb{E}[\pi_{tj}]$ $(t = 1, 2, 3, 4)$ for the 432 congressional districts across US (excluding Alaska and Hawaii).

Table 1: Center of clusters in original space $\{\Phi(\mathbb{E}[\mathbf{D}^b\Lambda^b\boldsymbol{\theta}^{\mu*}]), \mathbb{E}[\mathbf{D}^r\Lambda^r\boldsymbol{\psi}^{\mu*}]\}$. First 7 columns are the probability of answer "yes" for the corresponding attributes.

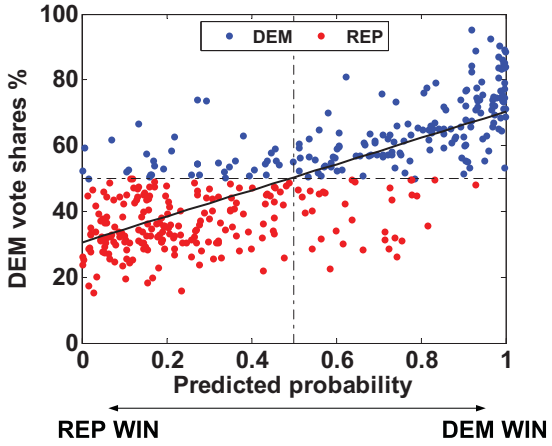
Cluster	Male	2006 Election	2008 Election	Black	Caucasian	Hispanic	Democrat	Republican	Age	Purchase Power
1	0.38	0.07	0.27	0.57	0.19	0.18	0.93	0.01	52	11509
2	0.39	0.63	0.87	0.29	0.55	0.08	0.93	0.04	49	76843
3	0.49	0.09	0.27	0.03	0.90	0.05	0.10	0.36	28	59999
4	0.49	0.07	0.22	0.04	0.88	0.06	0.11	0.34	48	74286

The corresponding $\{\mathbb{E}[\mathbf{D}^r\Lambda^r\boldsymbol{\theta}_t^{\mu*}], \Phi(\mathbb{E}[\mathbf{D}^b\Lambda^b\boldsymbol{\psi}_t^{\mu*}]\}$ are shown in Table 1 (this table provides mean values of a subset of Catalist parameters). From Table 1 and Figure 2, individuals in Clusters 1 and 2 are more likely to be Democrats. Cluster 1 seems to capture low-income Black and Hispanic Democrats, with poor turnout in the past election. In contrast, Cluster 2 is more likely to include high-income Democrats, with high turnout in previous elections. Cluster 1 is found to have high probability in many of the southern districts, especially these close to the border. Cluster 2 tends to appear in metropolitan areas, such as San Francisco, Los Angels, DC and New York. In a similar manner, Clusters 3 and 4 are more likely to include whites and Republicans (or undeclared voters). Age and purchasing power (in U.S. dollars) seem to distinguish Clusters 3 and 4 from Clusters 1 and 2.

To further assess how well the latent estimates capture constituent preferences, we examine the ability of the model to predict the party affiliation of the district's House member, based on the constituent characteristics in the Catalist datafile. Specifically, we use $\mathbb{E}[\boldsymbol{\pi}_j]$ as a feature vector, and build a linear probit-regression classifier (similar results can be obtained with other probabilistic classifiers), where shrinkage is imposed on the regression weights, using the same prior as imposed in the full model on $\boldsymbol{\xi}_j$.

In Figure 3, we plot the probit-regression-based probability that a given district will select a Democratic legislator, and along the vertical axis is plotted the fraction of vote share received in the district for the Democratic candidate (in the 2010 election). We consider the 406 (of 435) districts for which there was a contested election, with two candidates. We partitioned the districts into 5 folds, and iteratively train on 4 folds and test on the rest. Note, for example, when the model predicted that the probability of a Democratic win was 0.5, the fraction of vote received on average was about 50%. In the table in Figure 3, we note that the predictions of the model are in close alignment with actual district-level voting. These results indicate that the characterization of people in each district based on the Catalist data is a good representation of voter preferences. This provides further insight into why the Catalist data are useful for inferring more-confident prediction of roll call votes based on held-out text of the legislation (see Table 2), and also

Figure 3: Left column: Probability of Democratic win vs the vote share received for Democratic candidates. The solid line is a linear regression fit with vote share and predicted probability. Right column: Actual (empirical) probability of Democratic candidates win in each predicted probability bin.



Predicted Prob. bin of DEM win	Actual Prob. of DEM win
0-0.2	0.09
0.2-0.4	0.21
0.4-0.6	0.43
0.6-0.8	0.74
0.8-1	0.96

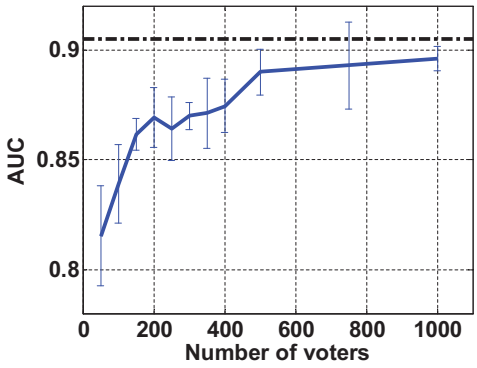


Figure 4: AUC versus number of voters in each districts. Black dash line corresponds to using all the data.

why a legislator tends to perform poorly in the next election when her voting record is inconsistent with the district-level preferences (reflected by large ξ_j, as depicted in Figure 7).

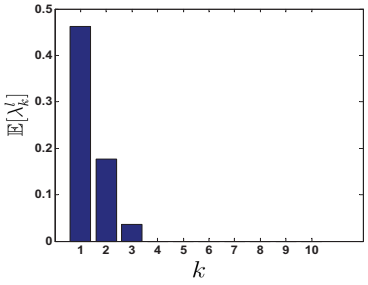


Figure 5: ($\mathbb{E}[\mathbf{diag}(\Lambda^\ell)]$.

It is of interest to examine the quality of the model as a function of the number of people per district we have demographic data from. Specifically, we train the whole model with a subset of randomly selected voters from Catalist dataset. We use $\mathbb{E}[\boldsymbol{\pi}_j]$ inferred from the subset as a feature vector, and perform the same prediction experiment discussed above. AUC (area under ROC curve) is employed as the metric for assessing the performance. In Figure 4, we plot the AUC as a function of average number of voters selected per district. The result is the average of 5 runs, and the error bar correspond to one standard deviation.

4.2 Insights on relationships between constituents and representatives

In the political science literature [8] and in recent machine learning research [10], it has been assumed that the latent space of the legislators and legislation is one-dimensional based on roll call votes (*i.e.*, feature vectors like $\boldsymbol{d}_j^\ell$ and $\boldsymbol{s}_l^\ell$ are *assumed* to be one-dimensional). Via the MPG prior on $\boldsymbol{\Lambda}^\ell$, we may *infer* the dimensions of these vectors. In Figure 5, we depict $\mathbb{E}[\text{diag}(\boldsymbol{\Lambda}^\ell)]$, which indicates that there is indeed one dominant latent dimension, but also two additional weaker dimensions.

To illustrate the connection of the dominant latent feature dimension to the characteristics of the representatives in each district, and to the characteristics of the people they represent, in Figure 6(a) we plot the principal dimension of $\boldsymbol{d}_j^\ell$ for each legislator, and note that Democrats tend to be positive in this dimension and Republicans negative. This result agrees with the ideal point obtained with the model in [8, 10]. In Figure 6(b) we plot the principal dimension of $\mathbb{E}[\mathbf{U}^d\tilde{\boldsymbol{\pi}}_j + \boldsymbol{d}_0^\ell]$, which within the model captures the roll-call-related preferences of the people who live in district j. Note that the Republican representatives (Figure 6(a)) appear to often be more negative in this dimension than their constituents (Figure 6(b)). Finally, in Figure 6(c) we plot $\mathbb{E}[\xi_j]$ in the principal dimension. Recall that ξ_j in $\boldsymbol{d}_j^\ell = \mathbf{U}^d\tilde{\boldsymbol{\pi}}_j + \boldsymbol{d}_0^\ell + \xi_j$ controls the degree to which the feature vector $\boldsymbol{d}_j^\ell$ associated with legislator j deviates from the characteristics of her constituents, reflected by $\tilde{\boldsymbol{\pi}}_j$. Moreover, a shrinkage prior was imposed on ξ_j, and therefore large $|\xi_j|$ is reflective of

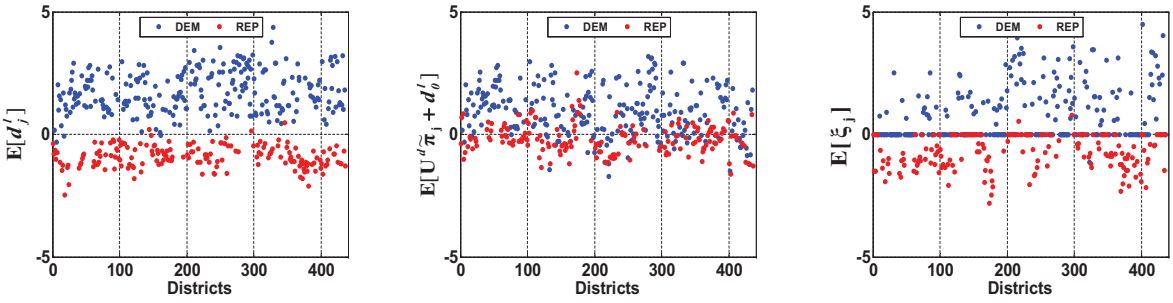


Figure 6: (a) : Principal dimension of $\mathbb{E}[d_j^\ell]$. The horizontal axis is the index of districts (alphabetically ordered). (b): Principal dimension of $\mathbb{E}[\mathrm{U}^d\tilde{\pi}_j + d_0^\ell]$. (c): Principal dimension of $\mathbb{E}[\xi_j]$.

legislators who may be voting in a manner that is not well linked to the people who live in their district (from the standpoint of the Catalist data). Note that ξ_j tends to be sparse, implying that representatives typically vote in line with their constituents, but there are also often significant non-zero ξ_j.

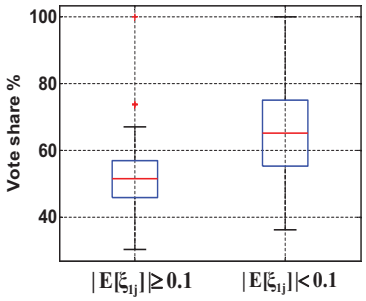


Figure 7: Vote share received for two groups of Democratic congressmen: those with $|\mathbb{E}[\xi_{1j}]| \geq 0.1$ and those with $|\mathbb{E}[\xi_{1j}]| < 0.1$.

We further examine the relationship between principal dimension of ξ_j (denoted ξ_{1j}) and the fraction of voter share for the jth legislator in the 2010 election. We focus on Democratic House members, as these were the ones for which there was significant turnover in that election. In Figure 7, we use box plots for two groups of Democratic representatives: those with $|\mathbb{E}[\xi_{1j}]| \geq 0.1$ and those with $|\mathbb{E}[\xi_{1j}]| < 0.1$. The 0.1 threshold is illustrative, and many related small thresholds yield similar results. Members who voted in a way that the model infers as aligned with the interests of their constituents (small $|\mathbb{E}[\xi_{1j}]|$) on average received a 15% larger share of the election vote than those legislators with relatively large $|\mathbb{E}[\xi_{1j}]|$. Notice a small number of legislators with high value of $\mathbb{E}[\xi_{1j}]$ also receive high vote share. These representatives are mainly from the less competitive districts. For example, Nydia Velazquez (NY-12), one of the two outliers, was challenged only by a third party candidate.

4.3 Analysis of the legislative topics in latent space

We examine the relationship between the topics of the legislation and the latent space associated with the roll-call vote. Specifically, $\delta_l = \boldsymbol{w}^T\tilde{\boldsymbol{\beta}}_l + w_0$ is a random effect associated with legislation l, and note that it is directly linked to the topic distribution on the legislation $\tilde{\boldsymbol{\beta}}_l$. The feature vector associated with the legislation is $\boldsymbol{s}_l^\ell = \mathbf{U}^s\tilde{\boldsymbol{\beta}}_l + \boldsymbol{s}_0^\ell$, and we here consider $\mathbb{E}[s_l]$ in the dominant (first) dimension, denoted $\mathbb{E}[s_{1l}]$. Based on Figure 6(a), positive values of $\mathbb{E}[s_{1l}]$ imply that the legislation is typically favored by Democrats, and negative values by Republicans.

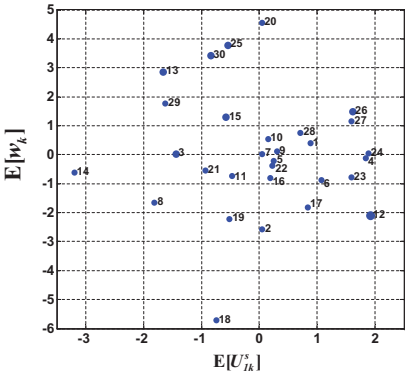
The kth component of the first row of $\mathbf{U}^s$, denoted U_{1k}^s, dictates the degree to which topic k contributes to s_{1l}. Further, component k of $\boldsymbol{w}$, w_k, dictates the degree to which topic k contributes to δ_l. Positive/negative values of U_{1k}^s correspond to topics favored by Democrats/Republicans, and positive/negative w_k correspond to topics that most congressman tend to vote "yes"/"no."

In Figure 8 we show the topics in the space $(\mathbb{E}[U_{1k}^s], \mathbb{E}[w_k])$, and also depict most-probable words associated with six example topics. During this time period, the Iraq and Afghanistan wars, which were started under a Republican president, tended to be aligned with the interest of the Republican Party (negative $(\mathbb{E}[U_{1k}^s])$; see Topic 3. By contrast, Topic 20, about health care, children and military veterans, tended to be favored irrespective of party (large positive $\mathbb{E}[w_k]$).

4.4 Prediction based on legislative text

We consider prediction of the votes of each legislator on held-out legislation, where the votes are predicted entirely by the text of the held-out legislation (the topic model infers $\tilde{\boldsymbol{\beta}}_l$ for new legislation, from which $\boldsymbol{s}_l^\ell$ and δ_l are estimated, and used to predict the probability of a particular vote). This experiment serves as a measure of model fitness.

In [10] the authors developed a model like that in (3), except that they did not have access to district-level constituency characteristics, like the Catalist data considered here. Therefore, in [10] the authors used the model in (3), except that $\boldsymbol{d}_j^\ell$ was drawn i.i.d. from a symmetric multivariate Gaussian distribution, rather than being



TOPIC 3	TOPIC 6	TOPIC 18
war	credit	related
military	loan	measure
force	income	recovery
freedom	tax	expense
international	insurance	requires
TOPIC 19	TOPIC 20	TOPIC 26
financial	child	community
transfer	care	American
property	medical	help
account	health care	opportunity
benefit	veteran	school

Figure 8: Left column: Regression weights of topics. Right column: Selected topics with the top-five most probable words shown.

Table 2: Comparison between proposed method and ideal point probit model from [10]. Shown are the number of votes in each probability bin, and the empirical probability of being correct in the prediction.

	Proposed model		Ideal point probit model	
Probit Confidence Bin	Votes in Confidence Bin	Empirical Probability	Votes in Confidence Bin	Empirical Probability
0.5-0.6	15821	0.54	16231	0.54
0.6-0.7	15241	0.63	17339	0.61
0.7-0.8	15170	0.7	17793	0.72
0.8-0.9	21756	0.81	22998	0.84
0.9-1	258367	0.98	251994	0.98
Pred. log-likelihood	-0.197		-0.204	

related to the constituency information (the latter implemented in the proposed model by relating $\boldsymbol{d}_j^{\ell}$ in (3) to $\boldsymbol{\pi}_j$). As the ideal point model in [8, 11], latent space s_l^l and the random effect δ_l are not associated with the legislative text, thus we cannot evaluate how well these two models predict votes for hold out legislations.

The roll call votes and associated text of legislation of the 111th US House of Representatives are partitioned into 6 folds. We iteratively train the model using five folds, and test on the sixth. The presented result is an aggregation of all six folds. Prediction confidence [20] and accuracy are employed as metrics. Specifically, for each held-out vote by legislator j on legislation l, the model yields a probability of "yes" , $p(r_{jl} = 1|-)$ and a probability of "no", $1 - p(r_{jl} = 1|-)$. We take $max\{p(r_{jl} = 1|-), 1 - p(r_{jl} = 1|-)\}$ for each held out vote, irrespective of whether the actual prediction is "yes" or "no," and place them into the corresponding probability bin, with bins ranging from $[0.5 - 0.6]$ to $[0.9 - 1]$. We wish to examine whether the prediction confidence matches empirical results. For example, if we examine all votes for which the model predicts the vote with confidence in the range $[0.7 - 0.8)$, we would expect the model should be able to correctly predict the vote between 70%-80% of the time. For the test legislations, we also compute the predictive log-likelihood $\log p(r_{test}|r_{train})$, which averaged for all six folds. In Table 2, we compare the prediction confidence and of the proposed model and that in [10] (probit link instead of logistic link). We observe that both models are "correct," in that the predicted confidence of the vote matches the empirical data (*e.g.*, for the proposed model, 258,367 of the held-out votes were predicted with a confidence of 0.9 to 1, and the model was correct in its prediction 98% of the time). In comparing the proposed model and that in [10], note that the former places 6,000 more votes in the 0.9 to 1 confidence bin and the predictive log-likelihood also improved, suggesting that the use of constituency (Catalist) data yields more confident predictions in legislator votes, and that confidence is vindicated experimentally.

Improvement manifested by our model is most prominent for contested legislation and unusual districts. Specifically, most of the 6,000 votes discussed above are for closely contested bills (those receiving less than 400 "yea" votes, corresponding to 267 out of 802 bills). The congressmen for which the model provides most improvement in vote prediction are among Republicans in districts dominated by Democratic constituents, such as Ileana Ros-Lehtinen (FL-18) and Michael Castle (DE). Their district-level characteristics (larger proportion of Democratic voters) adjust the ideal points toward Democratics, yielding more-confident predictions.

5 CONCLUSIONS

Binary matrix factorization is employed for analysis of roll-call data, with latent features associated with legislation informed by a topic model of the legislative text, and the latent features of each legislator informed by a statistical model of the people living in their district. The model is employed in a new manner to uncover insights into the workings of electoral representation, based on large-scale data, here specific to the U.S. Congress. The model is shown to produce improved prediction of votes on held-out legislation based on the text of the legislations, and demonstrates

the electoral consequences of legislators failing to represent the preferences of their constituents.

A APPENDIX

The stochastic variational Bayesian method for the proposed model is summarized in Algorithm 1.

Algorithm 1 Stochastic Variational Bayesian Analysis for Proposed Model

Partition $\mathbf{X}$ and $\mathbf{B}$ into N^* mini-batches.
Define local parameters Θ^l and global parameters Θ^g.
Initialize Θ^g by running model on a mini-batch
for $h = 1$ **to** N^* **do**
  $\omega_h = (a_3 + h)^{-b_3}$
  while stop criterion is not met **do**
    for $j = 1$ **to** J **do**
      for $n = 1$ **to** N_j^* **do**
        Estimate Θ^l (Detailed in Appendix A.1)
      end for
    end for
  end while
  Compute $\tilde{\Theta}^g$ (Detailed in Appendix A.2)
  Update $\Theta^g \leftarrow (1 - \omega_h)\Theta^g + \omega_h \tilde{\Theta}^g$
end for

.

A.1 Updates equations for local parameters

In this model, the local parameters $\Theta^l = \{\boldsymbol{\mu}_{s_{jn}^r}, \Sigma_{s_{jn}^r}, \boldsymbol{\mu}_{s_{jn}^b}, \Sigma_{s_{jn}^b}, \tilde{\boldsymbol{\pi}}_{jn}\}$ are the ones related with $\boldsymbol{s}_{jn}^r, \boldsymbol{s}_{jn}^b, c_{jn}$. Let us denote number of voters in district j within one mini-batch as N_j'.

Update equations for s_{jn}^r

$$q(\boldsymbol{s}_{jn}^r) \sim \mathcal{N}(\boldsymbol{\mu}_{s_{jn}^r}, \Sigma_{s_{jn}^r})$$

where the mean and covariance matrix are as following

$$\Sigma_{s_{jn}^r} = (\mathbb{E}[\sigma_j]\mathbb{E}[\boldsymbol{\Lambda}^r \mathbf{D}^{r\,T} \mathbf{D}^r \boldsymbol{\Lambda}^r] + \mathbf{I})^{-1}$$
$$\boldsymbol{\mu}_{s_{jn}^r}^r = \Sigma_{s_{jn}^r}(\mathbb{E}[\sigma_j]\mathbb{E}[\boldsymbol{\Lambda}^r \mathbf{D}^r]\boldsymbol{x}_{jn} + \sum_{t=1}^{T} \tilde{\pi}_{jnt}\mathbb{E}[\boldsymbol{\theta}_t^*])$$

The related expectation is

$$\mathbb{E}[\boldsymbol{\Lambda}^r \mathbf{D}^r \mathbf{D}^r \boldsymbol{\Lambda}^r] = \sum_{p=1}^{P^r}(\mathbb{E}[\boldsymbol{d}_p^r]\mathbb{E}[\boldsymbol{d}_p^r] + \Sigma_{\boldsymbol{d}_p^r}) \odot (\mathbb{E}[\boldsymbol{\lambda}^r])\mathbb{E}[\boldsymbol{\lambda}^r]^T + Diag(\Sigma_{\lambda_1^r}, ..., \Sigma_{\lambda_K^r}))$$

Update equations for s_{jn}^b

The update equations for $\boldsymbol{s}_{jn}^b$ is similar with $\boldsymbol{s}_{jn}^r$. We can obtain the updates by replacing the superscript r, $\mathbb{E}[\sigma_j]$ and $\boldsymbol{x}_{jn}$ with b,1 and $\mathbb{E}[\tilde{\boldsymbol{b}}_{jn}]$, respectively. The related expectation is as following,

$$\mathbb{E}[\tilde{b}_{jnp}] = \begin{cases} \mathbb{E}[\boldsymbol{d}_p^{b\,T}\boldsymbol{\Lambda}^b \boldsymbol{s}_{jn}^b] + \frac{\phi(\boldsymbol{d}_p^{b\,T}\boldsymbol{\Lambda}^b \boldsymbol{s}_{jn}^b)}{1-\Phi(\boldsymbol{d}_p^{b\,T}\boldsymbol{\Lambda}^b \boldsymbol{s}_{jn}^b)} \text{ if } b_{jnp} = 1 \\ \mathbb{E}[\boldsymbol{d}_p^{b\,T}\boldsymbol{\Lambda}^b \boldsymbol{s}_{jn}^b] - \frac{\phi(\boldsymbol{d}_p^{b\,T}\boldsymbol{\Lambda}^b \boldsymbol{s}_{jn}^b)}{\Phi(\boldsymbol{d}_p^{b\,T}\boldsymbol{\Lambda}^b \boldsymbol{s}_{jn}^b)}] \text{ if } b_{jnp} = 0 \end{cases}$$

Update equations for c_{jn}

See in section 3

A.2 Updates equations for global parameters

The remaining variational parameters are considered as global parameters Θ^g. We list the main update equations to calculate these intermediate global variational parameters $\tilde{\Theta}^g$ as following.

Update equations for d_p^r

$$q(\boldsymbol{d}_p^r) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{d}_p^r}, \Sigma_{\boldsymbol{d}_p^r})$$

where the mean and covariance matrix are as following

$$\boldsymbol{\mu}_{\boldsymbol{d}_p^r} = \mathbb{E}[\sigma_j]\Sigma_{\boldsymbol{d}_p^r}(\sum_{j=1}^{J} \frac{N_j}{N_j'} \sum_{n=1}^{N_j'} \mathbb{E}[\boldsymbol{\Lambda}^r]\mathbb{E}[\boldsymbol{s}_{jn}^r]x_{jnp})$$
$$\Sigma_{\boldsymbol{d}_p^r} = (\sum_{j=1}^{J} \frac{N_j}{N_j'} \sum_{n}^{N_j'} \mathbb{E}[\boldsymbol{\Lambda}^r \boldsymbol{s}_{jn}^r \boldsymbol{s}_{jn}^{r\,T} \boldsymbol{\Lambda}^{r\,T}]\mathbb{E}[\sigma_j])^{-1}$$

The related expectation is

$$\mathbb{E}[\boldsymbol{\Lambda}^r \boldsymbol{s}_{jn}^r \boldsymbol{s}_{jn}^{r\,T} \boldsymbol{\Lambda}^{r\,T}] = (\Sigma_{\boldsymbol{s}_{jn}^r} + \mathbb{E}[\boldsymbol{s}_{jn}^r]\mathbb{E}[\boldsymbol{s}_{jn}^r]) \odot (\mathbb{E}[\boldsymbol{\lambda}^r]\mathbb{E}[\boldsymbol{\lambda}^r]^T + Diag(\Sigma_{\lambda_1^r}, ..., \Sigma_{\lambda_K^r})).$$

where $\odot$ is the Hadamard product and $\boldsymbol{\lambda}^r = [\lambda_1^r, ..., \lambda_K^r]^T$

Update equations for λ_k^r

$$q(\lambda_k^r) \sim (\mu_{\lambda_k^r}, \Sigma_{\lambda_k^r})$$

The mean and variance are as following

$$\Sigma_{\lambda_k^r} = (\sum_{j=1}^{J} \frac{N_j}{N_j'} \sum_{n=1}^{N_j} \mathbb{E}[\sigma_j]\mathbb{E}[\boldsymbol{d}_k^{r\,T} \boldsymbol{d}_k^r s_{jnk}^{r\,2}] + \tau_k^r)^{-1}$$
$$\mu_{\lambda_k^r} = \Sigma_{\lambda_k^r}(\sum_{j=1}^{J} \frac{N_j}{N_j'} \sum_{n=1}^{N_j} \mathbb{E}[s_{jnk}]\mathbb{E}[\boldsymbol{d}_k]^T \hat{\boldsymbol{x}}_{jn})$$

The related expectation and equations are

$$\mathbb{E}[\boldsymbol{d}_k^{r\,T} \boldsymbol{d}_k^r s_{jnk}^{r\,2}] = (\mathbb{E}[\boldsymbol{d}_k^r]^T\mathbb{E}[\boldsymbol{d}_k^r] + tr(\Sigma_{\boldsymbol{d}_k^r}))(\mathbb{E}[s_{jnk}^r]^2 + \Sigma_{s_{jnk}^r})$$
$$\hat{\boldsymbol{x}}_{jn} = \boldsymbol{x}_{jn} - \sum_{\bar{k}=1, \bar{k}\neq k}^{K^r} \mathbb{E}[\boldsymbol{d}_{\bar{k}}^r]\mathbb{E}[s_{jn\bar{k}}]\lambda_{\bar{k}}^r$$

Update equations of φ_h^r

$$q(\varphi_h^r) \sim Gamma(a_{\varphi_h^r}, b_{\varphi_h^r})$$

where the shape and scale parameters are as following

$$a_{\varphi_h^r} = a_1 + \frac{K^r - h + 1}{2}$$
$$b_{\varphi_h^r} = 1 + \sum_{k=h}^{K^r} \frac{\mathbb{E}[\lambda_k^{r\,2}] \prod_{\bar{h}=1, \bar{h}\neq h}^{k} \mathbb{E}[\varphi_{\bar{h}}^r]}{2} I(k \geq h)$$

Update equations for θ_t^*

$$q(\boldsymbol{\theta}_t^*) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}_t^*}, \Sigma_{\boldsymbol{\theta}_t^*})$$

where the mean and covariance matrix are

$$\Sigma_{\boldsymbol{\theta}_t^*} = (\Sigma_0^r + \sum_{j=1}^{J} \frac{N_j}{N_j'} \sum_{n=1}^{N_j'} \tilde{\pi}_{jnt}\mathbf{I})^{-1}$$
$$\mu_{\boldsymbol{\theta}_t^*} = \Sigma_{\boldsymbol{\theta}_t^*}(\mu_0^r + \sum_{j=1}^{J} \frac{N_j}{N_j'} \sum_{n=1}^{N_j'} \tilde{\pi}_{jnt}\mathbb{E}[\boldsymbol{s}_{jn}^r])$$

Update equations for $d_p^b, \lambda_k^b, \lambda_h^b, \psi_t^*$

The update equations for $\boldsymbol{d}_p^b, \lambda_k^b, \lambda_h^b, \boldsymbol{\psi}_t^*$ are similar to $\boldsymbol{d}_p^r, \lambda_k^r, \lambda_h^r, \boldsymbol{\theta}_t^*$, respectively. We can obtain these update equations by replacing the superscript r, $\mathbb{E}[\sigma_j]$ and $\boldsymbol{x}_{jn}$ with b,1 and $\mathbb{E}[\tilde{\boldsymbol{b}}_{jn}]$, respectively.

Update equations for σ_j

$$q(\sigma_j) \sim Gamma(a_{\sigma_j}, b_{\sigma_j})$$
$$a_{\sigma_j} = a_0^r + P^r N_j/2$$
$$b_{\sigma_j} = b_0^r + \frac{N_j}{N_j'} \sum_{n=1}^{N_j'} (\boldsymbol{x}_{jn}^T \boldsymbol{x}_{jn} + \mathbb{E}[\boldsymbol{s}_{jn}^{r\,T} \boldsymbol{\Lambda}^r \mathbf{D}^{r\,T} \mathbf{D}^r \boldsymbol{\Lambda}^r \boldsymbol{s}_{jn}^r] - 2\boldsymbol{x}_{jn}^T \mathbb{E}[\mathbf{D}^r]\mathbb{E}[\boldsymbol{\Lambda}^r]\mathbb{E}[\boldsymbol{s}_{jn}^r])$$

The related expectation is

$$\mathbb{E}[\boldsymbol{s}_{jn}^{r\,T} \boldsymbol{\Lambda}^r \mathbf{D}^{r\,T} \mathbf{D}^r \boldsymbol{\Lambda}^r \boldsymbol{s}_{jn}^r] = tr((\sum_{p=1}^{P^r}(\mathbb{E}[\boldsymbol{d}_p^r]\mathbb{E}[\boldsymbol{d}_p^{r\,T}] + \Sigma_{\boldsymbol{d}_p^r}) \odot \mathbb{E}[\boldsymbol{\lambda}^r \boldsymbol{\lambda}^r)^T])(\mathbb{E}[\boldsymbol{s}_{jn}^r]\mathbb{E}[\boldsymbol{s}_{jn}^{r\,T}] + \Sigma_{\boldsymbol{s}_{jn}^r}))$$

Update equations for π_j

$$q(\boldsymbol{\pi}_j) \sim Dir(\theta_{\boldsymbol{\pi}_j})$$
$$\theta_{\boldsymbol{\pi}_j} = \kappa\boldsymbol{\nu} + \frac{N_j}{N_j'} \sum_{n=1}^{N_j'} \tilde{\boldsymbol{\pi}}_{jn}$$

Update equations for $\boldsymbol{\nu}$

We do a point estimate on $\boldsymbol{\nu}$ and q($\boldsymbol{\nu}$) is a degenerated distribution. The objective function of optimizing $\boldsymbol{\nu}$ is as following.

$$L(\boldsymbol{\nu}) = \log\text{GEM}(\boldsymbol{\nu}; \kappa_0) + \sum_{j=1}^{J} \mathbb{E}[\log Dir(\boldsymbol{\pi}_j|\boldsymbol{\nu})]$$

where $\text{GEM}(\boldsymbol{\nu}; \kappa_0)$ refers to the stick breaking prior. The derivation of the gradient can be found in [14].

Update equations for ξ_k

$$q(\xi_k) \sim \mathcal{N}(\boldsymbol{\mu}_{\xi_k}, \Sigma_{\xi_k})$$

where the mean and covariance is

$$\Sigma_{\xi_k} = (\sum_{l=1}^{L}(\mathbb{E}[\lambda_k^{l\,2}]\mathbb{E}[\bar{\boldsymbol{z}}_l^T \boldsymbol{u}_k^s \boldsymbol{u}_k^{s\,T} \bar{\boldsymbol{z}}_l]) + \mathbb{E}[\boldsymbol{\alpha}_k])^{-1}$$
$$\boldsymbol{\mu}_{\xi_k} = \Sigma_{\xi_k}(\sum_{l=1}^{L} \mathbb{E}[\lambda_k^l \bar{\boldsymbol{z}}_l^T \boldsymbol{u}_k^s \hat{\boldsymbol{r}_l}])$$

where related equations are

$$\hat{\boldsymbol{r}_l} = \tilde{\boldsymbol{r}}_l - \sum_{k=1}^{K^\ell} \lambda_k^l (\mathbf{C}\boldsymbol{u}_k^d + \xi_k)\boldsymbol{u}_k^{s\,T}\bar{\boldsymbol{z}}_l + \lambda_k^l \xi_k \boldsymbol{u}_k^{s\,T}\bar{\boldsymbol{z}}_l$$
$$\mathbf{C} = [\bar{\boldsymbol{c}}_1, ..., \bar{\boldsymbol{c}}_j]^T. \mathbb{E}[\bar{\boldsymbol{c}}_j] = \frac{1}{N_j'} \sum_{n=1}^{N_j'} \tilde{\boldsymbol{\pi}}_{jn}$$

Update equations for $\boldsymbol{u}_k^d$

$$q(\boldsymbol{u}_k^d) = \mathcal{N}(\mu_{\boldsymbol{u}_k^d}, \Sigma_{\boldsymbol{u}_k^d})$$

where the mean and covariance are

$$\Sigma_{\boldsymbol{u}_k^d} = (\sum_{l=1}^{L} \mathbb{E}[\mathbf{C}^T \mathbf{C} \lambda_k^{l\,2} \bar{\boldsymbol{z}}_l^T \boldsymbol{u}_k^s \boldsymbol{u}_k^{T\,s} \bar{\boldsymbol{z}}_l] + \alpha_d \mathbf{I})^{-1}$$
$$\mu_{\boldsymbol{u}_k^d} = \Sigma_{\boldsymbol{u}_k^d}(\sum_{l=1}^{L} \mathbb{E}[\lambda_k^l \bar{\boldsymbol{z}}_l^T \boldsymbol{u}_k^s \mathbf{C}^T \hat{\boldsymbol{r}_l}])$$

where the related equations are

$$\hat{\boldsymbol{r}_l} = \tilde{\boldsymbol{r}}_l - \sum_{k=1}^{K} \lambda_k^l (\mathbf{C}\boldsymbol{u}_k^d + \xi_k)\boldsymbol{u}_k^T \boldsymbol{z}_l + \lambda_k^l (\mathbf{C}\boldsymbol{u}_k^d)\boldsymbol{u}_k^{s\,T}\boldsymbol{z}_l$$
$$\mathbb{E}[\bar{\boldsymbol{c}}_j \bar{\boldsymbol{c}}_j^T] = \frac{1}{N_j'^2}(\sum_{n=1}^{N_j'} \sum_{m \neq n} \boldsymbol{\pi}_{jn}\boldsymbol{\pi}_{jm}^T + \sum_{n=1}^{N_j'} \text{diag}(\boldsymbol{\pi}_{jn}))$$
$$\mathbb{E}[\bar{\boldsymbol{z}}_l \bar{\boldsymbol{z}}_l^T] = \frac{1}{W_l^2}(\sum_{i=1}^{W_l} \sum_{m \neq i} \tilde{\boldsymbol{\beta}}_{il} \tilde{\boldsymbol{\beta}}_{ml}^T + \sum_{i=1}^{W_l} \text{diag}(\tilde{\boldsymbol{\beta}}_{il}))$$

Update equations for $\boldsymbol{u}_k^s$

$$q(\boldsymbol{u}_k) = \mathcal{N}(\mu_{\boldsymbol{u}_k^s}, \Sigma_{\boldsymbol{u}_k^s})$$

where the mean and covariance is

$$\Sigma_{\boldsymbol{u}_k^s} = (\sum_{l=1}^{L} \mathbb{E}[\lambda_k^{l\,2} \bar{\boldsymbol{z}}_l (\mathbf{C}\boldsymbol{u}_k^d + \xi_k)^T (\mathbf{C}\boldsymbol{u}_k^d + \xi_k) \bar{\boldsymbol{z}}_l^T] + \alpha_s \mathbf{I})^{-1}$$
$$\mu_{\boldsymbol{u}_k^s} = \Sigma_{\boldsymbol{u}_k^s}(\sum_{l=1}^{L} \mathbb{E}[\lambda_k^l \bar{\boldsymbol{z}}_l (\mathbf{C}\boldsymbol{u}_k^d + \xi_k)^T \hat{\boldsymbol{r}_l}])$$

$$\hat{\boldsymbol{r}_l} = \tilde{\boldsymbol{r}}_l - \sum_{k=1}^{K} \lambda_k^l (\mathbf{C}\boldsymbol{u}_k^d + \xi_k)\boldsymbol{u}_k^{s\,T}\bar{\boldsymbol{z}}_l + \lambda_k^l (\mathbf{C}\boldsymbol{u}_k^d + \xi_k)\boldsymbol{u}_k^{s\,T}\bar{\boldsymbol{z}}_l$$

Update equations for λ_k^l

$$q(\lambda_k^l) = \mathcal{N}(\mu_{\lambda_k^l}, \Sigma_{\lambda_k^l})$$

where the mean and variance are

$$\Sigma_{\lambda_k^l} = (\sum_{l=1}^{L} \mathbb{E}[\bar{\boldsymbol{z}}_l^T \boldsymbol{u}_k^s (\mathbf{C}\boldsymbol{u}_k^d + \xi_k)^T (\mathbf{C}\boldsymbol{u}_k^d + \xi_k)\boldsymbol{u}_k^{s\,T} \bar{\boldsymbol{z}}_l] + \tau_k^l)$$
$$\mu_{\lambda_k^l} = \Sigma_{\lambda_k^l}(\sum_{l=1}^{L} \hat{\boldsymbol{r}}_l^T (\mathbf{C}\boldsymbol{u}_k^d + \xi_k)\boldsymbol{u}_k^T \bar{\boldsymbol{z}}_l)$$

where $\hat{\boldsymbol{r}_l} = \tilde{\boldsymbol{r}}_l - \sum_{k=1}^{K} \lambda_k^l (\mathbf{C}\boldsymbol{u}_k^d + \xi_k)\boldsymbol{u}_k^{s\,T}\bar{\boldsymbol{z}}_l + \lambda_k^l (\mathbf{C}\boldsymbol{u}_k^d + \xi_k)\boldsymbol{u}_k^{s\,T}\bar{\boldsymbol{z}}_l$

Update equations for z_{il}

See in Section 3

Update equations for $\boldsymbol{\beta}_l$, $\boldsymbol{\phi}_k$

The update equation for $\boldsymbol{\beta}_l$ and $\boldsymbol{\phi}_k$ are same as the related parameter updates of latent Dirichlet allocation (LDA) and are omit for brevity.

REFERENCES

[1] James H Albert and Siddhartha Chib. 1993. Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association* 88, 422 (1993), 669–679.

[2] Joseph Bafumi and Michael C Herron. 2010. Leapfrog representation and extremism: A study of American voters and their members in Congress. *American Political Science Review* 104, 03 (2010), 519–542.

[3] Matthew James Beal. 2003. *Variational algorithms for approximate Bayesian inference*. University of London United Kingdom.

[4] Anirban Bhattacharya, David B Dunson, and others. 2011. Sparse Bayesian infinite factor models. *Biometrika* 98, 2 (2011), 291.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[6] Michael Bryant and Erik B Sudderth. 2012. Truly nonparametric online variational inference for hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems*. 2699–2707.

[7] Brandice Canes-Wrone, David W Brady, and John F Cogan. 2002. Out of step, out of office: Electoral accountability and House members' voting. *American Political Science Review* 96, 01 (2002), 127–140.

[8] Joshua Clinton, Simon Jackman, and Douglas Rivers. 2004. The statistical analysis of roll call data. *American Political Science Review* 98, 02 (2004), 355–370.

[9] Joshua D Clinton. 2006. Representation in Congress: constituents and roll calls in the 106th House. *Journal of Politics* 68, 2 (2006), 397–409.

[10] Sean Gerrish and David M Blei. 2011. Predicting legislative roll calls from text. In *Proceedings of the 28th international conference on machine learning (icml-11)*. 489–496.

[11] Sean Gerrish and David M Blei. 2012. How they vote: Issue-adjusted models of legislative behavior. In *Advances in Neural Information Processing Systems*. 2753–2761.

[12] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. 2013. Stochastic variational inference. *The Journal of Machine Learning Research* 14, 1 (2013), 1303–1347.

[13] Jeffrey R Lax and Justin H Phillips. 2009. How should we estimate public opinion in the states? *American Journal of Political Science* 53, 1 (2009), 107–121.

[14] Percy Liang, Slav Petrov, Michael I Jordan, and Dan Klein. 2007. The Infinite PCFG Using Hierarchical Dirichlet Processes.. In *EMNLP-CoNLL*. 688–697.

[15] Jon D Mcauliffe and David M Blei. 2008. Supervised topic models. In *Advances in neural information processing systems*. 121–128.

[16] Keith T Poole and Howard Rosenthal. 1985. A spatial model for legislative roll call analysis. *American Journal of Political Science* (1985), 357–384.

[17] Rajesh Ranganath, Chong Wang, Blei David, and Eric Xing. 2013. An adaptive learning rate for stochastic variational inference. In *International Conference on Machine Learning*. 298–306.

[18] Ruslan Salakhutdinov and Andriy Mnih. 2008. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the 25th international conference on Machine learning*. ACM, 880–887.

[19] E. Salazar, M. Cain, E. Darling, S. Mitroff, and L. Carin. 2012. Inferring Latent Structure From Mixed Real and Categorical Relational Data. In *ICML*.

[20] Esther Salazar, David B Dunson, and Lawrence Carin. 2013. Analysis of space–time relational data with application to legislative voting. *Computational Statistics & Data Analysis* 68 (2013), 141–154.

[21] Jayaram Sethuraman. 1994. A constructive definition of Dirichlet priors. *Statistica sinica* (1994), 639–650.

[22] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2004. Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes.. In *NIPS*. 1385–1392.

[23] Chong Wang, John William Paisley, and David M Blei. 2011. Online Variational Inference for the Hierarchical Dirichlet Process.. In *AISTATS*, Vol. 2. 4.

[24] Eric Wang, Dehong Liu, Jorge Silva, Lawrence Carin, and David B Dunson. 2010. Joint analysis of time-evolving binary matrices and associated documents. In *Advances in Neural Information Processing Systems*. 2370–2378.

[25] Eric Wang, Esther Salazar, David Dunson, Lawrence Carin, and others. 2013. Spatio-temporal modeling of legislation and votes. *Bayesian Analysis* 8, 1 (2013), 233–268.

[26] XianXing Zhang and Lawrence Carin. 2012. Joint modeling of a matrix with associated text via latent binary features. In *Advances in Neural Information Processing Systems*. 1556–1564.