To appear in the *Journal of Applied Statistics*
Vol. 00, No. 00, Month 20XX, 1–23

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

Zhengming Xing [a*] Sunshine Hillygus[b]

and Lawrence Carin[a]

[a] *Department of Electrical and Computer Engineering, Duke University, Durham, NC*; [b] *Political Science Department, Duke University, Durham, NC*

(*Received 00 Month 20XX; accepted 00 Month 20XX*)

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.

Keywords: stochastic variational inference, ideal point, topic model, hierarchical Dirichlet process, matrix factorization, multiplicative gamma process

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 often 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 [10, 21]. 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, scholars have offered approaches for incorporating information beyond roll-call votes [13, 25, 26, 31–33]. For example, [13] and [31] propose a latent factor model to jointly analyze the congressional votes and the legislative text. [14] improve the model by allowing the ideological position of legislators to vary on specific issues. [25] and [32] propose a spatio-temporal model that accounts for the time of the votes and the spatial location of the legislators' districts. Unfortunately, none of these approaches consider the preferences of the electorate.

Estimating the ideological preferences of a legislative district is far more difficult. Previous research tends to rely on crude proxies of district preferences, such as presidential vote share [9]. More recently, scholars have turned to public opinion polls—often pooling

*Corresponding author. Email: xingzhengming@gmail.com

across many different national surveys to increase sample sizes [2, 11, 18]. For example, [18] aggregates about 100 surveys to be able to get reliable estimates of the ideological preferences of states. [20] introduced a method to estimate public opinion using multilevel regression and post-stratification (MRP). At the state-level, this approach has found considerable success compared to simply disaggregating the data [18]. However, estimates of citizen preferences at a finer geographic scale—the congressional district level—are hindered by the number of available responses in public opinion surveys, resulting in imprecision and bias in the parameter estimates.

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 [24] is integrated with the hierarchical Dirichlet process (HDP) [29], 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 inferred value of this parameter is examined in the context of the success of the legislator in the next election, yielding a new approach for evaluating the relationship between legislative behavior, constituent preferences, and electoral outcomes. To address the massive scale of the constituency data, stochastic variational Bayesian inference is utilized [8, 16, 30].

The remainder of the paper is organized as follows. In Section 2, previous work using roll-call analyses is reviewed. Section 3 presents the proposed model, and details its individual components. A stochastic inference algorithm is discussed in Section 4. Section 5 presents experimental results, and conclusions are provided in Section 6.

2. Related work

2.1 *Ideal point model*

Most modern statistical analyses of roll call data rely on ideal point models to characterize legislators' ideological preferences. In such models, the voting results are represented as a binary matrix $\mathbf{R} \in \{0,1\}^{J\times L}$, where J is the number of legislators and L is the number of legislations they voted on. Binary matrix R is assumed to be connected with a latent real matrix $\tilde{\mathbf{R}} \in \mathbb{R}^{J\times L}$ through link function. [21] used logistic link function while [10] adopted the probit link; probit link is used in this paper. Let r_{jl} denotes the vote of jth legislator on lth legislation and $\tilde{r}_{jl}$ is the corresponding real value. The mathematic formulation is as follows

$$r_{jl} = \begin{cases} 1 & \text{if } \tilde{r}_{jl} \geq 0 \\ 0 & \text{if } \tilde{r}_{jl} < 0 \end{cases} \tag{1}$$

$\tilde{r}_{jl}$ is further modeled as

$$\tilde{r}_{jl} = \boldsymbol{d}_j^{\ell\,T} \boldsymbol{s}_l^{\ell} + \delta_l + \epsilon_{jl} \tag{2}$$

where $\boldsymbol{d}_j^{\ell} \in \mathbb{R}^{K^{\ell}}$ represents ideal point associated with legislator j and $\boldsymbol{s}_l^{\ell} \in \mathbb{R}^{K^{\ell}}$ represents the discrimination feature associated with legislation l. δ_l is the random effect term measuring the difficulty of passing legislation l. In contrast to many ideal point models that must preprocess bills to select a subset of "important" bills, the use of a random effect term allows us to utilize the full set of legislation. δ_l will be a large positive value for bills receiving a large portion of unanimous "yea" votes, such as ceremonial bills. Although some research argues that legislators' ideological preferences are multidimensional [12, 17], standard ideal point models tend to assume a single ideological dimension—-the latent dimension K^l is usually set to 1—for both computational consideration and interpretative purposes [10, 13]. In this paper, we employ the multiplicative gamma process prior to infer the latent dimension automatically. Details are discussed in Section 3. An additional advantage of our approach is that the incorporation of legislative text in the model allows for prediction of votes on future legislation, in contrast to the traditional ideal point models that rely on roll call data alone.

2.2 *Joint analysis of roll call and text*

Recent research has attempted to connect the voting patterns of legislators with the legislative text [13, 31]. In these models, a basic topic model, latent Dirichlet allocation [6], is employed to model the legislative text. Specifically, given a corpus with L piece of legislations and each piece of legislation is represented with a mixture of latent topics $\boldsymbol{\beta}_l$, where each of the K topics ϕ_k is a distribution over corpus wide defined dictionary of size V. For the ith term of legislation l, we first draw topic indicator z_{il} from mult($\boldsymbol{\beta}_l$) and then draw word v_{il} from mult($\phi_{z_{il}}$). One can also extend the topic modeling part with a hierarchical Dirichlett process [29].

[31] connected the roll call analysis with topic modeling by employing a mixture model to jointly cluster the legislation latent feature $\boldsymbol{s}_l$ and document-dependent topic usage $\boldsymbol{\beta}_l$. [13] improved the model by replacing mixture model with text regression which is closely related with supervised latent Dirichlet allocation [5]. Specifically, [13] assume a single latent dimension for the ideal point model, and therefore each piece of legislation l is attached with two scalar response variables s_l^{ℓ} and δ_l. The empirical distribution of the topics $\bar{\boldsymbol{z}}_l$ serves as the covariate. The kth element of $\boldsymbol{z}_l$ is defined as follows.

$$\bar{z}_{kl} = \sum_{i=1}^{N} \mathbf{1}(z_{il} = k) \tag{3}$$

The latent feature of legislation s_l^{ℓ} and random effect term δ_l are modeled in terms of $\bar{\boldsymbol{z}}_l$.

$$\begin{aligned} s_l^l &= \mathbf{U}^s \bar{\boldsymbol{z}}_l + s_0^{\ell} \\ \delta_l^l &= \boldsymbol{\omega} \bar{\boldsymbol{z}}_l + \omega_0 \end{aligned} \tag{4}$$

where $\mathbf{U}^s$ and $\boldsymbol{\omega}$ are the regression coefficients. Gaussian priors are placed on them. Under this setting, one may obtain the empirical topic distribution $\bar{\boldsymbol{z}}_l$ given the text of a new piece of legislation via topic models, and predict the congressional votes with the learned regression coefficients $\mathbf{U}^s$ and $\boldsymbol{\omega}$.

Notice all these works focus on explaining the discrimination feature $\boldsymbol{s}_l^{\ell}$ with legislative text and none of them try to further interpret the ideal points $\boldsymbol{d}_j$. In this paper, we estimate the latent ideological preferences of legislators $\boldsymbol{d}_j$ while accounting for individual-level constituency information from a random 300 million person sample of the American electorate. This allows for comparison of the relationship between the ideological prefer-

ences of the district and the legislator within the same measurement space, enabling an evaluation of representation and electoral accountability. Similar to [13], where the text of each bill is summarized as a mixture of corpus wide topics, we model the constituency of each district as a mixture of subgroups or clusters within the broader population. Model details are presented in the next section.

3. Model construction

3.1 *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 (`www.catalist.us`), a political data management vendor that compiles, checks, and standardizes voter registration files and then appends data from government and commercial sources. An academic subscription provided a 1% random sample (3 million cases)of their database in 2012, and includes a wide range of demographic, political, and commercial characteristics about each individual. Catalist samples have been used in a number of previous studies aimed at examining the American electorate [15, 23]. 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. Legislative votes include all U.S. House of Representative roll-call floor votes (6% are missing) on legislation from 2009-2011 and are denoted as $\mathbf{R} \in \{0,1\}^{J \times L}$. For bills with multiple floor votes, we rely on the final vote. 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.

3.2 *Matrix factorization of constituent data*

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

$$\mathbf{X}_j = \mathbf{D}^r \mathbf{\Lambda}^r \mathbf{S}_j^r + \mathbf{E}_j^r, \tag{5}$$

where $\mathbf{D}^r \in \mathbb{R}^{P^r \times K^r}$, $\mathbf{S}_j^r \in \mathbb{R}^{K^r \times N_j}$, $\mathbf{\Lambda}^r = \text{diag}(\lambda_1^r, \ldots, \lambda_{K^r}^r)$, 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 $\mathbf{\Lambda}^r$ 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]:

$$\lambda_k^r \sim \mathcal{N}(0, 1/\tau_k^r), \tau_k^r \sim \prod_{h=1}^{k} \varphi_h^r, \varphi_h^r \sim \text{Ga}(a_1, 1). \tag{6}$$

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:

$$b_{jpn} = \begin{cases} 1 & \text{if } \tilde{b}_{jpn} + \epsilon^b_{jpn} \geq 0 \\ 0 & \text{if } \tilde{b}_{jpn} + \epsilon^b_{jpn} < 0 \end{cases} \tag{7}$$

where $\epsilon^b_{jpn} \sim \mathcal{N}(0,1)$. We factorize the latent matrix as

$$\tilde{\mathbf{B}}_j = \mathbf{D}^b \mathbf{\Lambda}^b \mathbf{S}^b_j \tag{8}$$

, where $\mathbf{D}^b \in \mathbb{R}^{P^b \times K^b}$ and $\mathbf{S}^b_j \in \mathbb{R}^{K^b \times N_j}$. 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 $\mathbf{\Lambda}^b = \text{diag}(\lambda^b_1, \ldots, \lambda^b_{K^b})$.

3.3 *Clustering the constituency latent features*

Individual n sampled from district j is characterized by the nth column of $\mathbf{S}^r_j$ and $\mathbf{S}^b_j$. Assuming that people are clustered with respect to the attributes included in the Catalist database, we develop a joint mixture model for the columns of $\mathbf{S}^r_j$ and $\mathbf{S}^b_j$. Let $\boldsymbol{s}^r_{jn}$ and $\boldsymbol{s}^b_{jn}$ denote the nth columns of $\mathbf{S}^r_j$ and $\mathbf{S}^b_j$, respectively. We impose the following hierarchical Dirichlet process (HDP) model [29]:

$$\begin{aligned} \boldsymbol{s}^r_{jn} &\sim f(\boldsymbol{\theta}_{jn}),\ \boldsymbol{s}^b_{jn} \sim f(\boldsymbol{\psi}_{jn}), \\ \{\boldsymbol{\theta}_{jn}, \boldsymbol{\psi}_{jn}\} &\sim G_j, G_j \sim \text{DP}(\kappa, G_0), \\ G_0 &\sim \text{DP}(\kappa_0, H) \end{aligned} \tag{9}$$

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 [7, 27] of the HDP developed in [29] and a point estimate of $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots)^T$ [8, 19] to simplify the variational derivations (discussed in Section 4). 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 $\boldsymbol{\pi}_j = (\pi_{j1}, \ldots, \pi_{jT})^T$.

3.4 *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. Similar to the models discussed in Section 2.2, we employ latent Dirichlet allocation (LDA) [6] 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$(1, \boldsymbol{\phi}_k)$ [6].

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 $\boldsymbol{\beta}_l$ [5]. 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.

3.5 *Coupling constituency characteristics and legislative text: Roll-call analysis*

Like the ideal point model discussed in Section 2.1, for the binary roll-call votes we assume a latent matrix $\tilde{\mathbf{R}} \in \mathbb{R}^{J \times L}$ which we factorize as

$$\tilde{\mathbf{R}} = \mathbf{D}^\ell \boldsymbol{\Lambda}^\ell \mathbf{S}^\ell + \mathbf{E}^\ell \tag{10}$$

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

$$\boldsymbol{d}^\ell_j = \mathbf{U}^d \tilde{\boldsymbol{\pi}}_j + \boldsymbol{d}^\ell_0 + \boldsymbol{\xi}_j \ , \quad \boldsymbol{s}^\ell_l = \mathbf{U}^s \tilde{\boldsymbol{\beta}}_l + \boldsymbol{s}^\ell_0 \ , \tag{11}$$

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 $\xi_{jk} \sim \mathcal{N}(0, \alpha_{jk}^{-1})$, $\alpha_{jk} \sim \text{InvGa}(1, \gamma_{jk}/2)$, $\gamma_{jk} \sim \text{Ga}(10^{-6}, 10^{-6})$.

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.

3.6 *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.

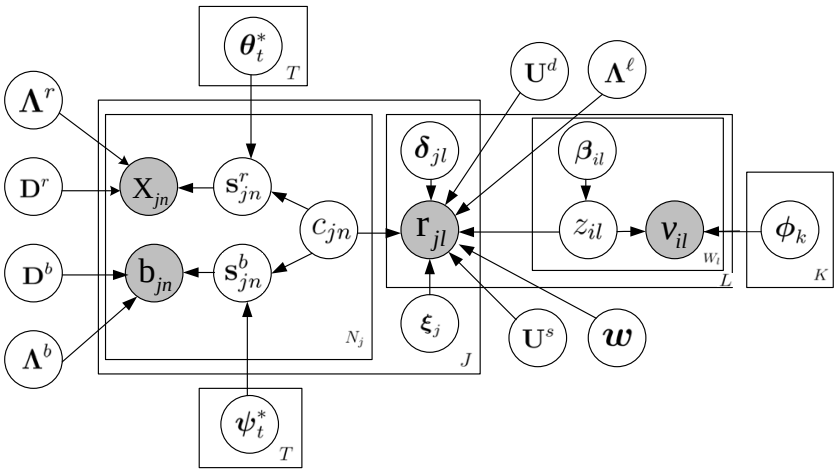


Figure 1. Graphical representation of the model.

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

We assume that the people (columns of $\mathbf{S}_j^r$ and $\mathbf{S}_j^b$) 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, $\boldsymbol{d}_j^\ell$, and a feature vector for legislation l, $\boldsymbol{s}_l^\ell$. To infer the relationship between the legislative votes of the congressman from district j relative to the interests of her/his constituents, we relate $\boldsymbol{d}_j^\ell$ 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.

4. Scaling Up: Variational Bayes and Stochastic Gradient Descent Inference

Our Catalist sample includes 2,969,925 people, and to handle data of this size we employ a mini-batch-based inference algorithm, stochastic variational Bayesian (VB) analysis [8, 16, 28, 30]. Unlike traditional VB inference [3], which includes the full 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 $\boldsymbol{\Theta}^g$. Under this setting, the evidence lower bound (ELBO) can be expanded as

$$\mathcal{L}(q) = E_q[\log p(\boldsymbol{\Theta}^{\boldsymbol{g}})] - E_q[\log Q(\boldsymbol{\Theta}^{\boldsymbol{g}})] + \sum_{n=1}^{N} E_q[\log p(\mathbf{x}_n, \boldsymbol{\theta}_{\boldsymbol{n}}^{\boldsymbol{l}})] - E_q[\log Q(\boldsymbol{\theta}_{\boldsymbol{n}}^{\boldsymbol{l}})] \tag{12}$$

where the first two terms characterize the ELBO for global parameters and the last two terms characterize the local parameters. The update equations for both global and local parameters can be derived via coordinate ascent. 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 $\tilde{\boldsymbol{\Theta}}^g$.

$$\boldsymbol{\Theta}^g \leftarrow (1 - \omega_h)\boldsymbol{\Theta}^g + \omega_h \tilde{\boldsymbol{\Theta}}^g \tag{13}$$

where $\omega_h \in (0, 1)$ is the weight given to each new batch, and also called the learning rate. Following [16], 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 [22] to adapt the learning step.

Details of the VB update equations are presented in the appendix C. 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 cat-

egorical distribution with parameter $\tilde{\boldsymbol{\pi}}_{jn}$, the components of which satisfy

$$\tilde{\pi}_{jnt} \propto \exp\{\mathbb{E}[\log p(\boldsymbol{s}^r_{jn}|\boldsymbol{\theta}^*_t)] + \mathbb{E}[\log p(\boldsymbol{s}^b_{jn}|\boldsymbol{\psi}^*_t)] + \mathbb{E}[\log(\pi_{jt})] + \sum_{l=1}^{L} \mathbb{E}[\log p(\tilde{r}_{jl}|c_{jn} = t, -)]\}. \tag{14}$$

The term $\mathbb{E}[\log(\pi_{jt})]$ characterizes the clustering characteristics of district j, where $p(\boldsymbol{s}^r_{jn}|\boldsymbol{\theta}^*_t)$ and $p(\boldsymbol{s}^b_{jn}|\boldsymbol{\psi}^*_t)$ 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

$$\tilde{\beta}_{ilk} \propto \phi_{v_{il},k} \exp\{\mathbb{E}[\log \beta_{lk}] + \sum_{j=1}^{J} \mathbb{E}[\log p(\tilde{r}_{jl}|z_{il} = k, -)]\} \tag{15}$$

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).

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

Algorithm 1 Stochastic Variational Bayesian Analysis for Proposed Model

```
Partition X and 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
  ω_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 C.2)
      end for
    end for
  end while
  Compute Θ̃^g (Detailed in Appendix C.3)
  Update Θ^g ← (1 − ω_h)Θ^g + ω_h Θ̃^g
end for
```

5. Experimental Results

To evaluate the relationship between the ideological preferences of House members and their constituents, 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 binary valued and real valued attributes employed in the experiment are summarized in the Appendix B. The roll-call data are from the 111th US Congress (January 3, 2009 - January 3, 2011) to match the time period of the Catalist sample. 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 [13], and obtain a bag of words with vocabulary size $V = 4743$. The election

for the 112th Congress took place on November 2, 2010, thus we use votes on bills in the 111th Congress that occurred before that date in order to evaluate the extent of electoral accountability and representation.

For model initialization, we first consider each data source separately. For example, we take a subset of the Catalist data, to infer $\mathbf{D}^r$, $\mathbf{D}^b$, $\mathbf{\Lambda}^r$ and $\mathbf{\Lambda}^b$. 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.

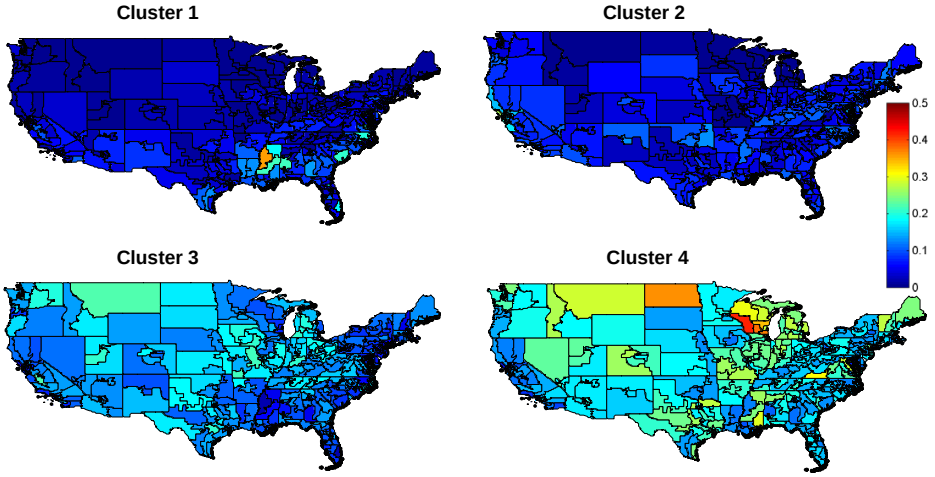


Figure 2. The expected probability of demographic clusters $\mathbb{E}[\pi_{tj}]$ $(t = 1, 2, 3, 4)$ for the 432 congressional districts across US (excluding Alaska and Hawaii).

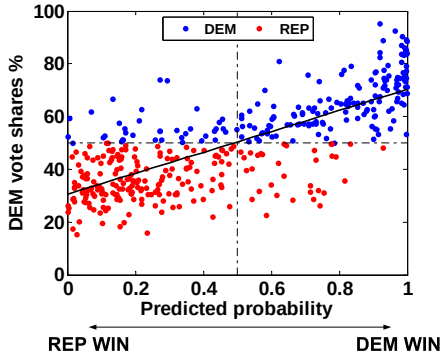
Table 1. Center of sample clusters in original space $\{\Phi(\mathbb{E}[\mathbf{D}^b\Lambda^b\boldsymbol{\theta}^{\mu*}]), \mathbb{E}[\mathbf{D}^r\Lambda^r\boldsymbol{\psi}^{\mu*}]\}$. First 7 columns are the probability of answer "yes" for the corresponding attributes.

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

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.

5.1 *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 (of 14) for 432 congressional districts (excluding Alaska and Hawaii) to show the geographic distribution of these



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 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.

sample clusters. 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 small subset of Catalist parameters). A comparison across a subset of features in Table 1 helps interpret the substantive meaning of the clusters. We see that individuals in Clusters 1 and 2 are more likely to be Democrats, with Cluster 1 capturing low-income Black and Hispanic Democrats with poor turnout in the past election and Cluster 2 capturing high-income Democrats with high turnout in previous elections. Figure 2 shows that Cluster 2 is geographically concentrated in metropolitan areas like San Francisco, Los Angeles, DC and New York. Cluster 1 is geographically concentrated in the South, especially near the border. Clusters 3 and 4, by contrast, are more likely to include whites and Republicans (or undeclared voters) with age and purchasing power (in U.S. dollars) distinguishing Clusters 3 and 4.

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 $\boldsymbol{\xi}_j$. This analysis offers a validation that the estimation of constituent preferences using the Catalist data.

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. In other words, these results offer reassurance that that the Catalist data offer a reasonable characterization of constituent preferences in each district. This relationship is also key to explaining our subsequent findings showing that Catalist data are useful for inferring more-confident prediction of roll call votes based on held-out text of the legislation (see Table 3) and that legislators tend to perform poorly in the next election when her voting record is inconsistent with the district-level preferences (reflected by large $\boldsymbol{\xi}_j$, as depicted in Figure 6(d)).

As a further evaluation of the model, we 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

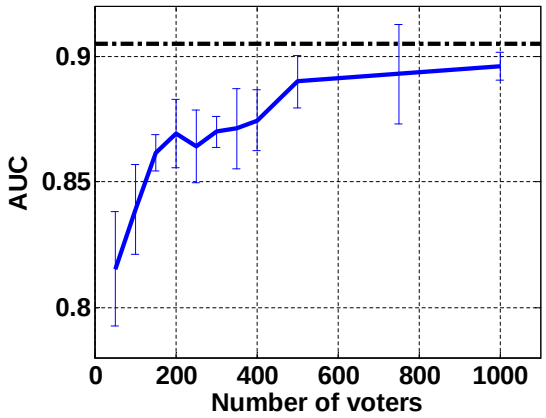


Figure 4. AUC versus number of voters in each districts. Black dash line corresponds to using all the data.

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. Here we see strong improvements once we sample size per district is at least 500 cases, highlighting the advantage of the Catalist database compared to data sources with fewer cases per district. For example, even large national surveys like the 50,000 person CCES has data support of fewer than 50 cases in a handful of congressional districts.

5.2 *Insights on relationships between constituents and representatives*

In the political science literature [e.g., 10] and in recent machine learning research [e.g., 13], 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 can *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.

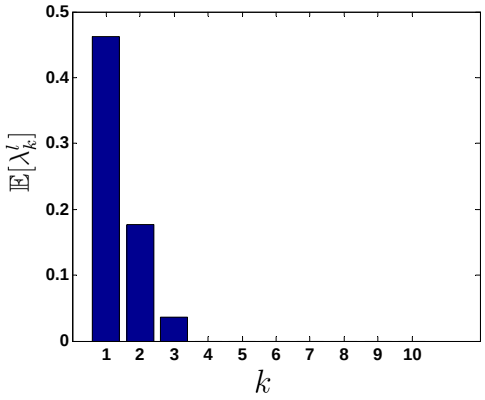


Figure 5. ($\mathbb{E}[\text{diag}(\boldsymbol{\Lambda}^\ell)]$.

To examine the relationship between the ideological preferences of House members and their constituents, we turn to a series of result graphs in Figure 6. In Figure 6(a), we plot the principal dimension of $\boldsymbol{d}_j^\ell$ for each legislator, noting that Democrats tend to be positive in this dimension and Republicans negative. This result agrees with the ideal

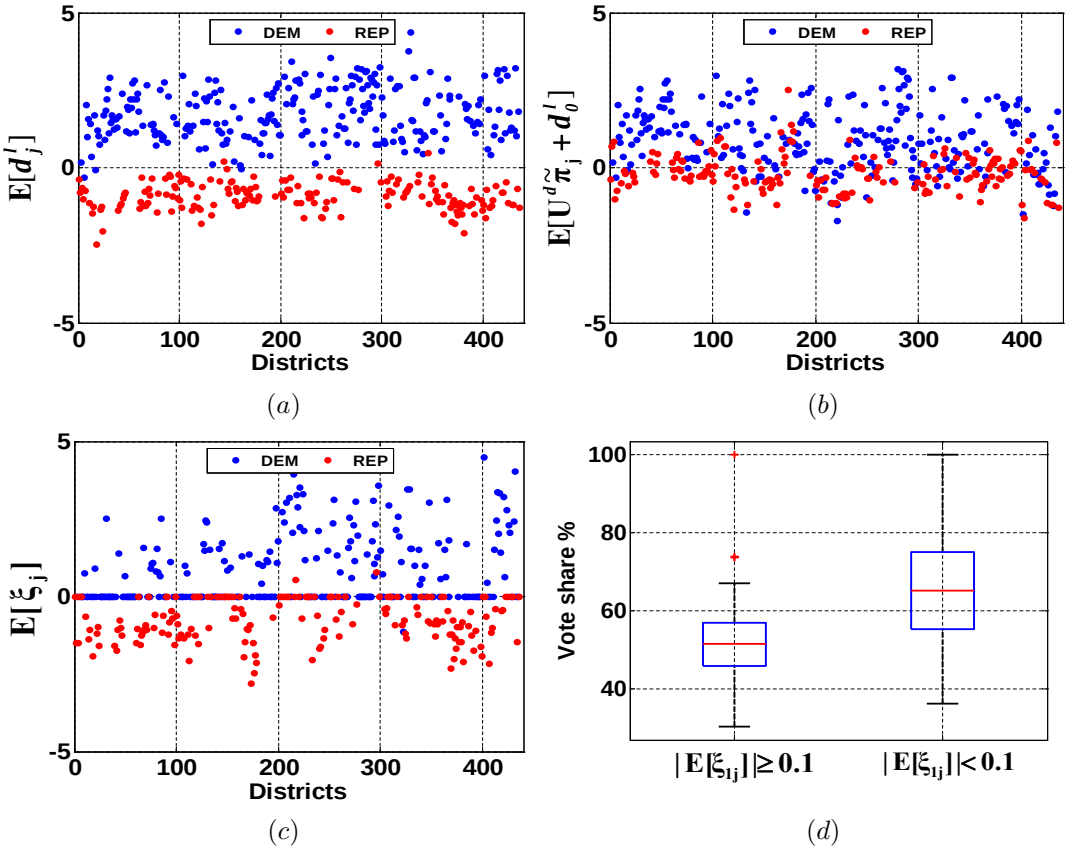


Figure 6. (a) : Principal dimension of $\mathbb{E}[\boldsymbol{d}_j^\ell]$. The horizontal axis is the index of districts (alphabetically ordered). (b): Principal dimension of $\mathbb{E}[\mathbf{U}^d\tilde{\boldsymbol{\pi}}_j + \boldsymbol{d}_0^\ell]$. (c): Principal dimension of $\mathbb{E}[\boldsymbol{\xi}_j]$. (d): Vote share received for two groups of Democratic congressmen: those with $|\mathbb{E}[\xi_{1j}]| \geq 0.1$ and those with $|\mathbb{E}[\xi_{1j}]| < 0.1$

point obtained with the model in [10, 13]. 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 $\mathbb{E}[\boldsymbol{\xi}_j]$ in the principal dimension. Recall that $\boldsymbol{\xi}_j$ in $\boldsymbol{d}_j^\ell = \mathbf{U}^d\tilde{\boldsymbol{\pi}}_j + \boldsymbol{d}_0^\ell + \boldsymbol{\xi}_j$ 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 $\boldsymbol{\xi}_j$, and therefore large $|\boldsymbol{\xi}_j|$ is reflective of 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 $\boldsymbol{\xi}_j$ tends to be sparse, implying that representatives typically vote in line with their constituents, but there are also often significant non-zero $\boldsymbol{\xi}_j$.

Finally, we evaluate the extent to which constituents hold their elected members accountable in the subsequent election by examining the relationship between the principal dimension of $\boldsymbol{\xi}_j$ (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 6(d), 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.

5.3 *Analysis of the legislative topics in latent space*

The inclusion of legislative text in the model is a novel addition to the literature on representation, so we next 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 $\mathbb{E}[\boldsymbol{s}_l]$ 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^s_{1k}, 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^s_{1k} correspond to topics favored by Democrats/Republicans, and positive/negative w_k correspond to topics that most congressman tend to vote "yes"/"no."

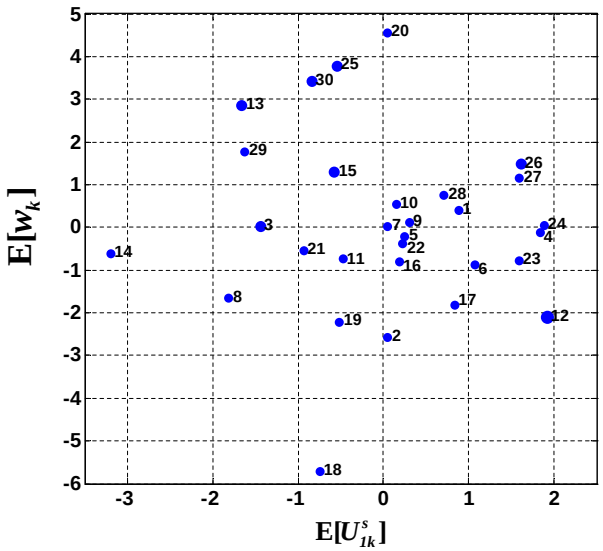


Figure 7. Regression weights of topics.

Table 2. Selected topics with the top-five most probable words shown.

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

In Figure 7 we show the topics in the space $(\mathbb{E}[U^s_{1k}], \mathbb{E}[w_k])$, and also depict most-probable words associated with six example topics in Table 2. 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 $(\mathbb{E}[U^s_{1k}])$; see Topic 3. By contrast, Topic 20, about health care, children and military veterans, tended to be

	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	

Table 3. Comparison between proposed method and ideal point probit model from [13]. Shown are the number of votes in each probability bin, and the empirical probability of being correct in the prediction.

favored irrespective of party (large positive $\mathbb{E}[w_k]$).

5.4 *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.

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 [25] 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 3, we compare the prediction confidence and accuracy of the proposed model and that in [13](probit link instead of logistic link). In their model, $\boldsymbol{d}_j^\ell$ was drawn i.i.d. from a symmetric multivariate Gaussian distribution rather than being related to the constituency information (the latter implemented in the proposed model by relating $\boldsymbol{d}_j^\ell$ in (11) to $\boldsymbol{\pi}_j$). 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 [13], 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.

Our model shows the most prominent improvements 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 legislators 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 Democrats, yielding more-confident predictions.

6. 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, specific to the 111th House of Representatives. 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.

The proposed model also offers significant potential for future substantive research evaluating the nature of representation in the United States. Specifically, we have not fully exploited the model's ability to analyze the political preferences of subgroups. In the paper, we focused on evaluating the basic extent of convergence or divergence between representatives and their constituents, and the electoral consequences for representatives who are diverging from their district preferences. With the information of particular constituents (e.g., voters, primary voters, partisans, wealthy constituents), we may be able to examine whether elected officials better represent some constituents more than others. The current model employs only limited information about individual legislators, as we have focused on incorporating constituent information. However, another model extension could involve taking account of additional factors such as legislator seniority, campaign spending, or legislator background. This offers the opportunity to further understand the relationship between representatives and their constituents ($\mathbb{E}[\xi_j]$).

Finally, while the data considered here are associated with politics, the basic model setup is more general with a variety of potential applications. For example, one could envision trying to assess whether specific individuals, from a region or group with particular demographics, will like/dislike given commercial 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 could predict whether it will be liked/disliked by particular people (here, whether legislators will vote yes/no on a new piece of legislation).

References

[1] J.H. Albert and S. Chib, *Bayesian analysis of binary and polychotomous response data*, J. Am. Stat. Assoc. (1993). 5

[2] J. Bafumi and M.C. Herron, *Leapfrog representation and extremism: A study of American voters and their members in congress*, Am. Polit. Sci. Rev. (2010). 2

[3] M.J. Beal, *Variational algorithms for approximate Bayesian inference*, Ph.D. thesis, University of London, 2003. 8

[4] A. Bhattacharya and D.B. Dunson, *Sparse Bayesian infinite factor models*, Biometrika (2011). 4

[5] D.M. Blei and J.D. McAuliffe, *Supervised Topic Models.*, in *NIPS*, 2007. 3, 6

[6] D.M. Blei, A.Y. Ng, and M.I. .Jordan, *Latent Dirichlet allocation*, J. Mach. Learn. Res. (2003). 3, 5

[7] T. Broderick, M.I. Jordan, and J. Pitman, *Beta processes, stick-breaking and power laws*, Bayesian Anal. 7 (2012), pp. 439–476. 5

[8] M. Bryant and E.B. Sudderth, *Truly Nonparametric Online Variational Inference for Hierarchical Dirichlet Processes.*, in *NIPS*, 2012. 2, 5, 8

[9] B. Canes-Wrone, D.W. Brady, and J.F. Cogan, *Out of step, out of office: Electoral accountability and house members' voting*, Am. Polit. Sci. Rev. 96 (2002), pp. 127–140. 1

[10] J. Clinton, S. Jackman, and D. Rivers, *The statistical analysis of roll call data*, Am. Polit. Sci. Rev. (2004). 1, 2, 3, 12, 13

[11] J.D. Clinton, *Representation in Congress: constituents and roll calls in the 106th House*, J. Polit.

(2006). 2
[12] J.M. Enelow and M.J. Hinich, *The spatial theory of voting: An introduction*, CUP Archive, 1984. 3
[13] S. Gerrish and D.M. Blei, *Predicting legislative roll calls from text*, in *ICML*, 2011. 1, 3, 4, 9, 12, 13, 15
[14] S.M. Gerrish and D.M. Blei, *How they vote: Issue-adjusted models of legislative behavior*, Polit. Sci. (2012). 1
[15] E.D. Hersh and C. Nall, *The primacy of race in the geography of income-based voting: New evidence from public voting records*, American Journal of Political Science (2015). 4
[16] M.D. Hoffman, D.M. Blei, C. Wang, and J. Paisley, *Stochastic variational inference*, J. Mach. Learn. Res. (2013). 2, 8
[17] S. Jackman, *Multidimensional analysis of roll call data via bayesian simulation: identification, estimation, inference, and model checking*, Polit. An. 9 (2001), pp. 227–241. 3
[18] J.R. Lax and J.H. Phillips, *How should we estimate public opinion in the states?*, Am. J. Polit. Sci. (2009). 2
[19] P. Liang, S. Petrov, M.I. Jordan, and D. Klein, *The Infinite PCFG Using Hierarchical Dirichlet Processes.*, in *EMNLP-CoNLL*, 2007. 5, 22
[20] D.K. Park, A. Gelman, and J. Bafumi, *Bayesian multilevel estimation with poststratification: state-level estimates from national polls*, Polit. An. 12 (2004), pp. 375–385. 2
[21] K.T. Poole and H. Rosenthal, *A spatial model for legislative roll call analysis*, Am. J. Polit. Sci. (1985). 1, 2
[22] R. Ranganath, C. Wang, D.M. Blei, and E. Xing, *An adaptive learning rate for stochastic variational inference*, in *ICML*, 2013. 8
[23] T. Rogers and M. Aida, *Vote self-prediction hardly predicts who will vote, and is (misleadingly) unbiased*, American Politics Research 42 (2014), pp. 503–528. 4
[24] R. Salakhutdinov and A. Mnih, *Bayesian probabilistic matrix factorization using Markov chain Monte Carlo*, in *ICML*, 2008. 2
[25] E. Salazar, D.B. Dunson, and L. Carin, *Analysis of space–time relational data with application to legislative voting*, Comput. Stat. Data An. (2013). 1, 15
[26] E. Salazar, M. Cain, E. Darling, S. Mitroff, and L. Carin, *Inferring Latent Structure From Mixed Real and Categorical Relational Data*, in *ICML*, 2012. 1
[27] J. Sethuraman, *A constructive definition of Dirichlet priors*, Tech. Rep., DTIC Document, 1991. 5
[28] L. Tan and D. Nott, *A stochastic variational framework for fitting and diagnosing generalized linear mixed models*, Bayesian Anal. 9 (2014), pp. 963–1004. 8
[29] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei, *Hierarchical Dirichlet processes*, J. Am. Stat. Assoc. (2004). 2, 3, 5
[30] C. Wang, J. Paisley, and D.M. Blei, *Online variational inference for the hierarchical Dirichlet process*, in *AISTATS*, 2011. 2, 8
[31] E. Wang, D. Liu, J. Silva, L. Carin, and D.B. Dunson, *Joint analysis of time-evolving binary matrices and associated documents*, in *NIPS*, 2010. 1, 3
[32] E. Wang, E. Salazar, D.B. Dunson, and L. Carin, *Spatio-temporal modeling of legislation and votes*, Bayesian Anal. (2013). 1
[33] X. Zhang and L. Carin, *Joint Modeling of a Matrix with Associated Text via Latent Binary Features.*, in *NIPS*, 2012. 1

Appendix A. Hierarchical representation of the model

Model Catalist data

$$\boldsymbol{x}_{jn} = \mathbf{D}^r \boldsymbol{\Lambda}^r \boldsymbol{s}^r_{jn} + \boldsymbol{\epsilon}^r_{jn} \qquad \boldsymbol{d}^r_p \sim \mathcal{N}(0, \mathbf{I}) \qquad \boldsymbol{s}^r_{jn} \sim \mathcal{N}(\boldsymbol{\theta}^*_{c_{jn}}, \mathbf{I})$$

$$\tilde{\boldsymbol{b}}_j = \mathbf{D}^b \boldsymbol{\Lambda}^b \boldsymbol{s}^b_{jn} + \boldsymbol{\epsilon}^b_{jn} \qquad b_{jnp} = 0, \text{ if } \tilde{b}_{jnp} > 0 \qquad b_{jnp} = 1, \text{ if } \tilde{b}_{jnp} < 0$$

$$\boldsymbol{d}^b_p \sim \mathcal{N}(0, \mathbf{I}) \qquad \boldsymbol{s}^b_{jn} \sim \mathcal{N}(\boldsymbol{\psi}^*_{c_{jn}}, \mathbf{I}) \qquad c_{jn} \sim Cat(\boldsymbol{\pi}_j)$$

$$\boldsymbol{\pi}_j \sim DP(\kappa \boldsymbol{\nu}) \qquad \nu_t = \nu'_t \prod_{i=1}^{t} (1 - \nu'_t) \qquad \nu'_t \sim beta(1, \kappa_0)$$

$$\boldsymbol{\psi}^*_t \sim \mathcal{N}(\mu^b_0, \boldsymbol{\Sigma}^b_0) \qquad \boldsymbol{\theta}^*_t \sim \mathcal{N}(\boldsymbol{\mu}^r_0, \boldsymbol{\Sigma}^r_0) \qquad \boldsymbol{\epsilon}^r_{jn} \sim \mathcal{N}(0, \sigma_j^{-1}\mathbf{I})$$

$$\boldsymbol{\epsilon}^b_{jn} \sim \mathcal{N}(0, \mathbf{I}) \qquad \bar{c}_{jt} = \frac{1}{N_j} \sum_{n=1}^{N_j} I(c_{jnt} = t)$$

Model legislative text

$$z_{il} \sim Cat(\boldsymbol{\beta}_l) \qquad \boldsymbol{\beta}_l \sim Dir(\eta_1) \qquad v_{il} \sim Cat(\boldsymbol{\phi}_{z_{il}})$$

$$\boldsymbol{\phi}_k \sim Dir(\eta_2) \qquad \bar{z}_{lk} = \frac{1}{W_l} \sum_{i=1}^{W_l} I(z_{il} = k)$$

Model rollcall votes

$$\tilde{r}_{jl} = \boldsymbol{d}^{\ell\,T}_j \boldsymbol{\Lambda}^l \boldsymbol{s}^\ell_l + \delta_l + \delta_{jl} + \epsilon^\ell_{jl} \qquad r_{jl} = 1, \text{ if } \tilde{r}_{jl} > 0 \qquad r_{jl} = 0, \text{ if } \tilde{r}_{jl} \leq 0$$

$$\boldsymbol{d}^\ell_j = \mathbf{U}^d \bar{\mathbf{c}}_j + \boldsymbol{d}^\ell_0 + \boldsymbol{\xi}_j \qquad \boldsymbol{s}^\ell_l = \mathbf{U}^s \bar{\mathbf{z}}_l + s^\ell_0 \qquad \delta_l = \bar{\mathbf{z}}_l \mathbf{w} + w_0$$

$$\boldsymbol{u}^d_k \sim \mathcal{N}(0, \alpha_d^{-1}\mathbf{I}) \qquad \mathbf{u}^s_k \sim \mathcal{N}(0, \alpha_s^{-1}\mathbf{I}) \qquad \mathbf{w} \sim \mathcal{N}(0, \alpha_w^{-1}\mathbf{I})$$

$$\xi_{jk} \sim \mathcal{N}(0, \alpha^{-1}_{jk}) \qquad \alpha_{jk} \sim \text{InvG}(1, \gamma_{jk}/2) \qquad \gamma_{jk} \sim \text{Ga}(10^{-6}, 10^{-6})$$

$$\delta_{il} \sim \mathcal{N}(0, \alpha'^{-1}_{jl}) \qquad \alpha'^{-1}_{jl} \sim \text{InvG}(1, \gamma'_{jl}/2) \qquad \gamma'_{jl} \sim \text{Ga}(10^{-6}, 10^{-6})$$

$$\epsilon^l_{jl} \sim \mathcal{N}(0, 1)$$

Multiplicative gamma prior

$$\boldsymbol{\Lambda}^{\{r,b,l\}} = diag(\lambda_1^{\{r,b,l\}}, ..., \lambda_K^{\{r,b,l\}}) \qquad \lambda_k^{\{r,b,l\}} \sim \mathcal{N}(0, \tau_k^{\{r,b,l\}\,-1}) \qquad \tau_k^{\{r,b,l\}} = \prod_{h=1}^{k} \varphi_h^{\{r,b,l\}}$$

$$\varphi_h^{\{r,b,l\}} \sim Gamma(a_1, 1)$$

$I(.)$ denotes the indicator function. $I(.) = 1$ if the inside condition holds, 0 otherwise. Diffuse gamma priors are placed on σ_j.

Appendix B. Catalist attributes

We summarize the Catalist attributes used in the model in Table B1.

Categories	Number of binary attributes	Number of real attributes	Description
Gender	1	0	male or female
Age	0	3	age, mean and standard deviation of age among house members
Finance	3	3	income; household value; information related with investment, bonds purchasing, credit card
Race	7	5	race includes Black, Caucasian, Hispanic, Asian etc.
Turnout	2	1	turnout in 2006 and 2008 general election; turnout rate of the household members
Party affiliation	9	2	Democrat, Republican and other party
Behavior	1	1	play golf or not; Internet usage
Children	4	1	have child between certain age or not
Religion	10	0	include Catholic, Protestant, Hindu, Muslim, Buddist etc.
Home	0	1	own or rent
Donation	3	0	donate to political, religious and environmental issues

Table B1. Summary of Catalist attributes

Appendix C. Stochastic variational inference update equations

C.1 *Variational evidence lower bound*

The posterior inference of the model is performed via Stochastic Variational Bayesian. Let $\mathcal{X} = \{\mathbf{X}_j, \mathbf{B}_j, \mathbf{R}\}$ denotes the training data, $\mathbf{\Gamma}$ denote all the hyper parameters and $\mathbf{\Theta}$ denotes all the latent variables. We use the following fully factorized variational distributions to approximate the posterior distribution.

$$q(\Theta) = \prod_{p=1}^{P^r} q(\boldsymbol{d}_p^r) \prod_{p=1}^{P^b} q(\boldsymbol{d}_p^b) \prod_{j=1}^{J} \prod_{n=1}^{N_j} q(\boldsymbol{s}_n^r) q(\boldsymbol{s}_n^b) q(c_{jn}) \prod_{k=1}^{K^r} q(\lambda_k^r) q(\varphi_k^r) \prod_{k=1}^{K^b} q(\lambda_k^b) q(\varphi_k^b) \prod_{j=1}^{J} q(\sigma_j)$$
$$q(\boldsymbol{\nu}) \prod_{j=1}^{J} q(\boldsymbol{\pi}_j) \prod_{t=1}^{T} q(\boldsymbol{\theta}_t^*) q(\boldsymbol{\psi}_t^*) \prod_{k=1}^{K^\ell} q(\varphi_k^l) q(\lambda_k^l) q(\boldsymbol{u}_k^s) q(\boldsymbol{u}_k^d) \prod_{j=1}^{J} \prod_{k=1}^{K^\ell} q(\xi_{jk}) q(\alpha_{jk}) q(\gamma_{jk})$$
$$q(\boldsymbol{w}) \prod_{l=1}^{L} \prod_{i=1}^{W^l} q(z_{il}) \prod_{g=1}^{K} q(\boldsymbol{\phi}_g) \prod_{l}^{L} q(\boldsymbol{\beta}_l) \prod_{j}^{J} \prod_{l=1}^{L} q(\delta_{jl}) q(\alpha'_{jl}) q(\gamma'_{jl})$$

The evidence lower bound is as following.

$$\begin{aligned} \log(p(\mathcal{X}|\mathbf{\Gamma})) &\geq \mathbb{E}[\log(p(\mathcal{X}, \mathbf{\Theta}, \mathbf{\Gamma}))] - \mathbb{E}[\log(q(\mathbf{\Theta}))] \\ &= \mathcal{L}(q) \end{aligned}$$

We partition the Catalist data into N^* mini-batches. The roll call data and related legislative votes are considered as a whole. Each mini-batch contains voters from all the 435 districts. The variational parameters specific to each batch are local parameters $\mathbf{\Theta}^l$ and the remaining are globe parameters $\mathbf{\Theta}^g$. The evidence lower bound can be expanded to

$$\mathcal{L}(q) = E_q[\log p(\mathbf{\Theta}^{\boldsymbol{g}})] - E_q[\log Q(\mathbf{\Theta}^{\boldsymbol{g}})] + \sum_{n=1}^{N} E_q[\log p(\mathbf{x}_n, \boldsymbol{\theta}_{\boldsymbol{n}}^{\boldsymbol{l}})] - E_q[\log Q(\boldsymbol{\theta}_{\boldsymbol{n}}^{\boldsymbol{l}})]$$

We can then derive the stochastic coordinate ascent update equations for both global and local parameters.

C.2 *Updates equations for local parameters*

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

Update equations for $\boldsymbol{s}^r_{jn}$

$$q(\boldsymbol{s}^r_{jn}) \sim \mathcal{N}(\boldsymbol{\mu}_{s^r_{jn}}, \mathbf{\Sigma}_{s^r_{jn}})$$

where the mean and covariance matrix are as following

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

The related expectation is

$$\mathbb{E}[\mathbf{\Lambda}^r \mathbf{D}^r \mathbf{D}^r \mathbf{\Lambda}^r] = \textstyle\sum_{p=1}^{P^r}(\mathbb{E}[\boldsymbol{d}^r_p]\mathbb{E}[\boldsymbol{d}^r_p] + \mathbf{\Sigma}_{\boldsymbol{d}^r_p}) \odot (\mathbb{E}[\boldsymbol{\lambda}^r])\mathbb{E}[\boldsymbol{\lambda}^r]^T + Diag(\Sigma_{\lambda^r_1}, ..., \Sigma_{\lambda^r_K}))$$

Update equations for $\boldsymbol{s}^b_{jn}$

The update equations for $\boldsymbol{s}^b_{jn}$ is similar with $\boldsymbol{s}^r_{jn}$. 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,

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

Update equations for c_{jn}

$$q(c_{jn}) \sim Cat(\tilde{\boldsymbol{\pi}}_{jn})$$

The t dimension parameter $\tilde{\boldsymbol{\pi}}_{jn}$ is

$$\tilde{\pi}_{jnt} \propto$$
$$\exp(\mathbb{E}[log(\mathcal{N}(\boldsymbol{s}^r_{jn}|\boldsymbol{\theta}^*_t, \mathbf{I}))] + \mathbb{E}[log(\mathcal{N}(\boldsymbol{s}^b_{jn}|\boldsymbol{\mu}^b_t, \mathbf{I}))] + \mathbb{E}[log(\pi_{jt})] + \textstyle\sum_{l=1}^{L} \frac{\mathbb{E}[\bar{r}^{(1)}_{jlnt} u^{(1)}_{tl}]}{N'_j} - \frac{\mathbb{E}[u^{(1)\,2}_{tl}]}{2N'^{\,2}_j})$$

The corresponding expectation and equations are as following

$$\bar{r}^{(1)}_{jlnt} = \tilde{r}_{jl} - \delta_l - \delta_{jl} - (\boldsymbol{d}^\ell_0 + \boldsymbol{\xi}_j)^T \mathbf{\Lambda}^\ell \boldsymbol{s}^\ell_l - \textstyle\sum_{t' \neq t}\{[\boldsymbol{u}^{d\,T}_t \mathbf{\Lambda}^\ell \boldsymbol{s}^\ell_l][\frac{\sum_{n' \neq n} I(c_{jn'}=t)}{N'_j}]\}$$
$$u^{(1)}_{tl} = [\boldsymbol{u}^{d\,T}_t \mathbf{\Lambda}^\ell \boldsymbol{s}^\ell_l].$$
$$\mathbb{E}[\log(\mathcal{N}(\boldsymbol{s}^r_{jn}|\boldsymbol{\mu}^r_t, \mathbf{I}))] = -\frac{tr(\mathbf{\Sigma}_{\boldsymbol{s}^r_{jn}}) + \mathbb{E}[\boldsymbol{s}^{r\,T}_{jn}]\mathbb{E}[\boldsymbol{s}^r_{jn}] + tr(\mathbf{\Sigma}_{\boldsymbol{\mu}^r_t}) + \boldsymbol{\mu}^{r\,T}_t \boldsymbol{\mu}^r_t - 2\mathbb{E}[\boldsymbol{s}^r_{jn}]\mathbb{E}[\boldsymbol{\mu}^r_r]}{2}$$
$$\mathbb{E}[\log(\mathcal{N}(\boldsymbol{s}^b_{jn}|\boldsymbol{\mu}^b_t, \mathbf{I}))] = -\frac{tr(\mathbf{\Sigma}_{\boldsymbol{s}^b_{jn}}) + \mathbb{E}[\boldsymbol{s}^{b\,T}_{jn}]\mathbb{E}[\boldsymbol{s}^b_{jn}] + tr(\mathbf{\Sigma}_{\boldsymbol{\mu}^b_t}) + \boldsymbol{\mu}^{b\,T}_t \boldsymbol{\mu}^b_t - 2\mathbb{E}[\boldsymbol{s}^b_{jn}]\mathbb{E}[\boldsymbol{\mu}^b_r]}{2}$$
$$\mathbb{E}[\log(\pi_{jt})] = \Psi(\theta_{\pi t}) - \Psi(\textstyle\sum_t \theta_{\pi t})$$

$\Psi(.)$ denotes the digamma function.

C.3 *Updates equations for global parameters*

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

Update equations for $\boldsymbol{d}^r_p$

$$q(\boldsymbol{d}^r_p) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{d}^r_p}, \mathbf{\Sigma}_{\boldsymbol{d}^r_p})$$

where the mean and covariance matrix are as following

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

The related expectation is

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

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

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

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 $\boldsymbol{\theta}_t^*$

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

where the mean and covariance matrix are

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}_t^*} = (\boldsymbol{\Sigma}_0^r + \sum_{j=1}^{J}\frac{N_j}{N_j'}\sum_{n=1}^{N_j'}\tilde{\pi}_{jnt}\mathbf{I})^{-1}$$
$$\mu_{\boldsymbol{\theta}_t^*} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_t^*}(\mu_0^r + \sum_{j=1}^{J}\frac{N_j}{N_j'}\sum_{n=1}^{N_j'}\tilde{\pi}_{jnt}\mathbb{E}[\boldsymbol{s}_{jn}^r])$$

Update equations for $\boldsymbol{d}_p^b, \lambda_k^b, \lambda_h^b, \boldsymbol{\psi}_t^*$

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

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

Update equations for $\boldsymbol{\pi}_j$

$$q(\boldsymbol{\pi}_j) \sim Dir(\theta_{\boldsymbol{\pi}_j})$$

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

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.

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

where GEM($\boldsymbol{\nu};\kappa_0$) refers to the stick breaking prior. The derivation of the gradient can be found in [19].

Update equations for $\boldsymbol{\xi}_k$

$$q(\boldsymbol{\xi}_k) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\xi}_k}, \boldsymbol{\Sigma}_{\boldsymbol{\xi}_k})$$

where the mean and covariance is

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

where related equations are

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

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

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

where the mean and covariance are

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

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 $\boldsymbol{u}_k^s$

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

where the mean and covariance is

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

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

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

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

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

Update equations for z_{il}

$$q(z_{il}) = Cat(\tilde{\boldsymbol{\beta}}_{il})$$

The parameter $\tilde{\boldsymbol{\beta}}_{il}$ is as following,

$$\tilde{\beta}_{ilk} \propto exp\{\mathbb{E}[log\, Cat(v_{il}|\boldsymbol{\phi}_k)] + \mathbb{E}[log\ \beta_{lk}] + \sum_{j=1}^{J}(\frac{\mathbb{E}[\bar{r}_{jkl}^{(2)}u_{jk}^{(2)}]}{W_l} - \frac{\mathbb{E}[u_{jk}^{(2)\ 2}]}{2W_l^2})$$

The related equations are as following,

$$\bar{r}_{jkil}^{(2)} = \tilde{r}_{jl} - \delta_l - \delta_{jl} - \sum_{k'\neq k}\{[\boldsymbol{d}_j^{\ell T}\boldsymbol{\Lambda}^{\ell}\boldsymbol{u}_{k'}^{s} + w_k][\frac{\sum_{i'\neq i} I(z_{li'}=k')}{W_l}]\}$$
$$u_{jk}^{(2)} = [\boldsymbol{d}_j^{\ell T}\boldsymbol{\Lambda}^{\ell}\boldsymbol{u}_k^{s} + w_k].$$

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 omitted for brevity.

After obtaining all the intermediate global parameters $\tilde{\boldsymbol{\Theta}}^g$, we update the global parameters $\boldsymbol{\Theta}^g$ as following.

$$\boldsymbol{\Theta}^g \leftarrow (1-\omega_h)\boldsymbol{\Theta}^g + \omega_h\tilde{\boldsymbol{\Theta}}^g$$