Predicting Legislative Roll Calls from Text

Sean M. Gerrish SGERRISH@CS.PRINCETON.EDU
David M. Blei BLEI@CS.PRINCETON.EDU
Department of Computer Science, Princeton University. 35 Olden St., Princeton, NJ 08540

Abstract

We develop several predictive models linking legislative sentiment to legislative text. Our models, which draw on ideas from ideal point estimation and topic models, predict voting patterns based on the contents of bills and infer the political leanings of legislators. With supervised topics, we provide an exploratory window into how the language of the law is correlated with political support. We also derive approximate posterior inference algorithms based on variational methods. Across 12 years of legislative data, we predict specific voting patterns with high accuracy.

1. Introduction

Quantitative political scientists analyze patterns in legislative data to better understand how governments behave. One focus of quantitative political scientists is *roll call data*, historical records of legislators' votes on a set of issues. Roll call data can reveal information about the members of a government; for example, we can analyze roll call data from the United States Congress or the British Parliament to uncover the political leanings of their members (Clinton et al., 2004).

Roll call data is essential for understanding government because it represents atomic and concrete actions of its members. But this data is only one part of a richer record which includes bill texts, speeches, press releases, public plans, and other items. In this work, we extend methods for roll call data to include other information. Specifically, we integrate bill texts into contemporary models of roll call data. This gives a new way of exploring and analyzing the government record and, further, gives a useful predictor of government. While traditional methods can only fill in missing votes, we develop tools that can predict how legislators will vote on a new bill.

We will extend the *ideal point model*. The ideal point model is a mainstay in quantitative political science for analyzing roll call data (Clinton et al., 2004). It posits a latent "political space" along the real line and places each legislator and bill in that space. The legislator's position is called an *ideal point*, because a bill at this position maximizes her utility. The model assumes that whether a legislator votes `yea` on a bill depends on a function of the ideal point and the bill's location.[1] With these assumptions, we can use observed roll call data to infer ideal points and bills' locations. Figure 1 illustrates ideal points from the 111th Senate.

Political scientists examine ideal points to understand legislators' preferences. But as predictive models, ideal point models suffer from a fundamental limitation: they are models of the votes alone. Consequently, they can be used to fill in missing votes but cannot predict how legislators will vote on future legislation. Further, they provide no insight into what drives voting patterns—the political activity of the legislature is summarized with two columns of real numbers.

To these ends, we describe several models that connect the voting patterns of legislators to the original

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

[1] These assumptions stem from a particular utility model, and this methodology is an instance of the *item-response* model from psychometrics and educational testing (Lord, 1980). Each bill also has another variable, the *difficulty*, which is described below but omitted in the introduction.

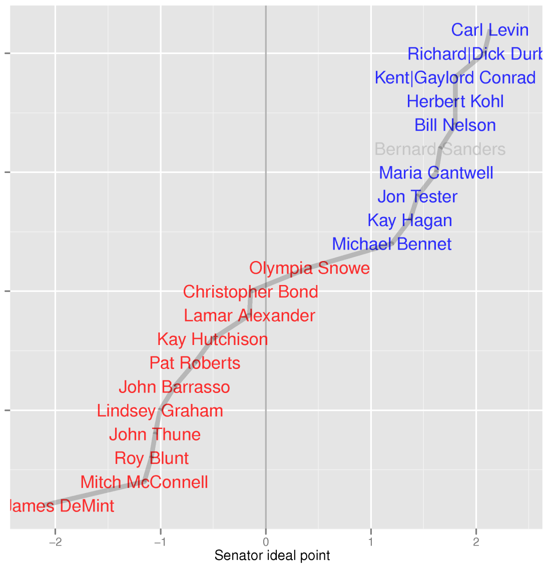

Figure 1. Sample Senator ideal points in the 111th Congress. Ideal points tend to separate the U.S. political parties: Democrat are blue, and Republicans are red. A plot of all legislators is included in the supplementary materials.

text of bills. One of these models embeds the statistical assumptions of *supervised topic modeling* (Blei & McAuliffe, 2008) into the ideal point model, where the locations of the bills are predicted from the latent topics in their texts. This model—the ideal point topic model—can predict complete votes on pending bills and provides a new way of exploring how legislative language is correlated with political support. The other models predict inferred ideal points using different forms of regression on phrase counts.

In the following sections, we review the details of ideal point estimation and develop several models for predicting votes from legislative text. We derive an approximate posterior inference algorithm for ideal point models based on variational methods and analyze six Congresses (12 years) of legislative data from the United States Congress. Given a legislative history, these models can accurately predict votes on future legislation. One of these models, the ideal point topic model, can help summarize and visualize the political landscape of a government body based both on the voting patterns of its members and the language of its issues.

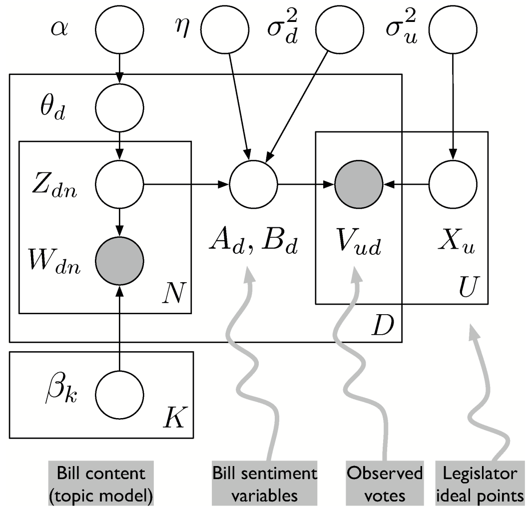

Figure 2. The ideal point topic model. Priors over the multinomials θ_d and β are both symmetric Dirichlet distributions.

2. Text and ideal point models

In this section we develop models of both roll call data and legislative text.

The Bayesian ideal point model. The ideal point model is a generative model of choices (votes) and issues (bills). Each legislator u is associated with an *ideal point* X_u (see Figure 1), and each bill is associated with a *discrimination* B_d and *difficulty* A_d. The vote v_{ud} is assumed drawn from the linear model

$$p(v_{ud} = \text{yea}) = \sigma(x_u b_d + a_d), \quad (1)$$

where $\sigma(t) = \frac{\exp(t)}{1+\exp(t)}$. This is a logistic regression with random effects.[2]

Notice the roles of the per-legislator and per-bill latent variables in Equation 1. The difficulty parameter a_d explains bills that all legislators will vote for or against, e.g., "A bill to congratulate the winner of the World Series." For the remaining bills, i.e., those with some political division, the discrimination parameter b_d and ideal point interact. When estimated from roll call data, these latent variables capture the political leanings of the legislators and the political tone of the legislation.

We assume Gaussian priors for X_u, A_d, and B_d. The prior for X_u has mean zero; the means of A_d, B_d are

[2]Some ideal point models use a probit model; we have found both approaches to ideal points to yield similar results

inferred from text. The ideal point variance is σ_u^2; the variance for both discrimination and difficulty is σ_d^2.

We now develop models relating the text of a bill to the variables a_d and b_d. Associating text to bill variables has a predictive advantage because new bills can be situated in the space of ideal points. It also has an interpretive advantage because language becomes associated with political sentiment.

Modeling ideal points with text regression. We developed two predictive ideal-point models which use text regression (Kogan et al., 2009). For these, we first fit an ideal-point model to a training set of bills and all legislators using the variational algorithm described in Section 2. We then fit ridge regression[3] (`LARS`) and Lasso [4] (`L2`) to these bills' parameters a_d, b_d using a vector of their n-gram[5] counts $\boldsymbol{w}_d$ as covariates.

Modeling ideal points with supervised topics. The text regression models link individual words or phrases to bill sentiment. In this section, we connect textual *themes* with bill sentiment. We refer to this model as an ideal point topic model (IPTM).

To model themes, we use the assumptions of supervised Latent Dirichlet Allocation (sLDA) (Blei & McAuliffe, 2008). As in Latent Dirichlet Allocation (Blei et al., 2003), each bill is represented as a mixture of latent topics θ_d, where each of K topics β_k is a multinomial probability distribution over terms. For the n^{th} term of bill d, we draw topic z_{dn} from $\text{Mult}(\theta_d)$, and then draw word w_{dn} from the topic β_{z_n}.

Like sLDA, the ideal point topic model further assumes each bill d is attached to a response variable. In this case, the response variable is the 2-component vector of bill variables (a_d, b_d). The distribution of the response is a linear model whose covariates are the empirical distribution of the topics $\boldsymbol{z}_d$ for the bill,

$$\begin{aligned} a_d &\sim \mathcal{N}(\boldsymbol{\eta}_a^\top \bar{z}_d, \sigma_d^2) \\ b_d &\sim \mathcal{N}(\boldsymbol{\eta}_b^\top \bar{z}_d, \sigma_d^2), \end{aligned}$$

where $\bar{z}_d = (1/N)\sum_n z_{dn}$. This setting is more complex than the original sLDA model: the response variables are *hidden*—they are not observed directly, but are used downstream in the voting model.

Finally, we add a Gaussian prior to $\boldsymbol{\eta}$. The full model is represented as a graphical model in Figure 2.

The only observed variables in the model are the bill texts and votes. Our goal in fitting this model is to uncover the posterior

$$p(A_d, B_d, X_u, \boldsymbol{\eta}, \beta, z, \theta | \boldsymbol{W}, \boldsymbol{V}), \tag{2}$$

which can then be used in exploratory or predictive tasks. Conditioned on these variables, our analysis proceeds with the posterior distribution of the ideal points, discriminations and difficuties, topics, and coefficients. Computing the posterior exactly is intractable, so we use variational inference to approximate it. We describe this in further detail in Section 4.

This posterior allows us to explore the connection between language and political tone. For example, the coefficients $\boldsymbol{\eta}$ are a direct connection between bills' topics and the political tone of these bills. Examples of this are provided in Section 5. The topics β, learned from both text and votes, provide a lexical window into legislative issues. The parameters $\boldsymbol{\eta}, \beta$ together also allow us to predict votes using the text of new bills; Section 4 provides detail about this.

Anchoring legislators. Note that a fit of the ideal point model has multiple modes. In one mode, Democrats tend to have positive ideal points, while Republicans are negative; in another, Republicans are positive, while Democrats are negative. To keep fits of the different models identifiable, several researchers have applied nonzero priors over specific legislators to encourage the model to prefer one of these modes (Jackman, 2001; Clinton et al., 2004; Martin & Quinn, 2002).

In the study in Section 5, we anchor four legislators with strong priors ($\sigma_d = 10^{-3}$) at ideal points ± 4. We select two congresspersons from each chamber and two from each party: Kennedy (S-Dem) and Waxman (H-Dem) are centered at +4 and Enzi (S-Rep) and Donald Young (H-Rep) are centered at -4.[6] We selected these Senators for consistency with previous work (Clinton et al., 2004). We selected the Representatives because they have held long offices in the

[3]Implemented in the "penalized" package for R

[4]implemented with the "lars" package for R

[5]See Section 5 for details.

[6]This value was selected to be large yet not completely out of the ordinary.

House. Without these sharp priors, the model still discovers ideal points which cleanly separate political parties but may converge on "opposite" modes in different fits. With the priors, we obtain consistent ideal points at the expense of predictive performance.

3. Related work

Ideal point models, a form of spatial voting model, have roots as far back as the 1920s (Enelow & Hinich, 1984). They are fit by both frequentist (Poole & Rosenthal, 1985; Heckman & Snyder, 1996) and Bayesian methods (Jackman, 2001; Martin & Quinn, 2002; Clinton et al., 2004), have been embedded in a time series (Martin & Quinn, 2002; Wang et al., 2010), and have been developed for higher dimensional political spaces (Jackman, 2001; Heckman & Snyder, 1996).

Topic models have been applied to Senate speeches, such as to discern "the substantive structure of the rhetorical [legislative] agenda" (Quinn et al., 2006). They have also been used with legislative speeches to gauge legislators' sentiment toward legislation using roll-calls (Thomas et al., 2006). Modeling sentiment in text is more generally discussed in the field of sentiment analysis; see Pang and Lee (2008) for a review.

The ideal point topic model relates closely to user-recommendation models based on matrix factorization (Salakhutdinov & Mnih, 2008). Matrix factorization methods for recommendation are akin to large-scale spatial behavior models (though usually with no "difficulty" term, which acts as an intercept). Many of these matrix factorization models for user recommendation do not provide a method of predicting one user's item preference without other users' preferences on the same item.

Two works stand out as closely related to this work. One of these is fLDA, which models binary or continuous ratings with user affinity to topics (Agarwal & Chen, 2010). Another is Wang et al. (2010), who describe a similar application by combinating topic models and matrix completion. Their work also draws on ideal point models, models transitions over time, and is designed to learn the dimensionality of the latent factors. Under the generative assumptions of their model, bills and matrix cells (e.g., votes) are conditioned on a shared mixture; in our model, votes are conditioned on words' topics.

This is the first work to study predictive accuracy of votes on new bills, where we use a spatial voting model as a "cold" prediction mechanism.

4. Posterior Inference

Computing the posterior in Equation 2 is intractable. Posterior inference for traditional Bayesian ideal point models is traditionally implemented with MCMC methods such as Gibbs sampling (Johnson & Albert, 1999; Jackman, 2001; Martin & Quinn, 2002; Clinton et al., 2004). We introduce an alternative algorithm – which can be applied to both the standard ideal point model and the ideal point topic model – which uses variational methods (Jordan et al., 1999). Variational methods provide a deterministic alternative to Gibbs sampling that is amenable to optimization in large-scale datasets. They have been successfully applied to many kinds of topic models, where corpus size and vocabulary dimension are large. Furthermore, in the ideal point topic model, fast Gibbs samplers are unavailable because the conditionals needed are not analytically computable. An MCMC strategy would require a more complicated sampling scheme.

Variational methods posit a family of distributions over the latent variables. That family is indexed by free parameters, called *variational parameters*, which are fit to minimize the KL divergence between the variational family and the true posterior. The family is chosen to be simpler than the posterior, which allows for efficient optimization. Though simpler, the fitted variational distributions are found to be good proxies for the true posterior (Jordan et al., 1999).

We begin by specifying a fully-factorized variational distribution for the model posterior. First, word assignments z_{dn} and topic proportions are governed by multinomial parameters ϕ_d and Dirichlet parameters γ_d, as in LDA (Blei et al., 2003). The variational distribution for legislators' ideal points X_u; bills' parameters A_d, B_d; and coefficients $\boldsymbol{\eta}$ are Gaussian with respective means τ_u, κ_d, $\hat{\eta}$ and variances σ^2_τ, σ^2_κ, σ^2_η.

The variational distribution is

$$q(\tau,\sigma_\tau,\kappa,\sigma_\kappa,\phi,\theta) = \prod_u q(X_u|\tau_u,\sigma^2_\tau) \times \prod_D q(A_d,B_d|\kappa_d,\sigma^2_\kappa) \times \prod_D q(\theta_d|\gamma_d) \prod_{N_d} p(z_n|\phi_n) \times q(\boldsymbol{\eta}|\hat{\eta},\sigma_{\hat{\eta}}). \quad (3)$$

Inference proceeds by minimizing the KL between Equation 3 and the true posterior 2, which is equivalent to maximizing a lower bound on the marginal probability of the observations. We use coordinate and gradient ascent to maximize this bound. The supplementary materials give further details of the variational inference algorithm.

Prediction After they are fit to legislators' votes and bill text, the variational parameters τ, $\hat{\eta}$, and β can be used to estimate the vote of each legislator on a *new* bill d using its text. To predict whether legislator u votes `yea` on d, the per-word parameters ϕ_n of d are estimated using the topics β. Once ϕ has been estimated, the probability of a `yea` vote is given by $p(v_{ud} = \texttt{yea}) = \sigma(\tau_u(\bar{\phi}_d\hat{\eta}_b) + \bar{\phi}_d\hat{\eta}_a)$ [7], where $\bar{\phi}_d$ is $\frac{1}{N_d}\sum_{N_d}\phi_n$. In practice, we fit $\hat{\eta}$ with no regularization after the model has converged. This gives slightly better results which are more robust to parameter selection.

5. Analyzing the U.S. House and Senate

We studied the performance of these models on 12 years of data from the United States House of Representatives and Senate. We first demonstrate how the ideal point topic model can be used to explore legislative data; then we evaluate the models' generalization performance in predicting votes from bill texts.

We collected roll-coll votes for Congressional sessions 106 through 111 (January 1997 to January 2011). We used votes about bills and resolutions, and only votes regarding the legislation as a whole (as opposed to, e.g., amendments of the legislation). We downloaded the data from Govtrack, an independent Website which provides comprehensive legislative information to the public. Our collection contains 4,447 bills, 1,269 unique legislators, and 1,837,033 `yea` or `nay` roll-call votes.

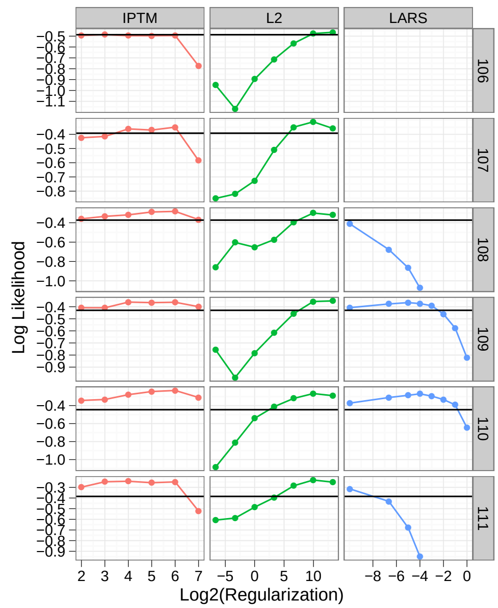

Figure 3. Predictive log likelihood on heldout votes. Models are shown by color for different regularizations (x axis), for Congresses 106 to 111. For LARS and L2, the regularization is the complexity parameter; for the ITPM, the regularization is the the number of topics. The *yea* baseline is the horizontal black line. LARS is below the fold for 106-107. The ideal point topic model performs with less variance across its regularization parameter (the number of topics).

To select the vocabulary, we lemmatized the bills with Treetagger (Schmid, 1994). Then we retained a vocabulary of statistically significant n-grams ($1 \leq n \leq 5$) using likelihood ratios. These n-grams were treated as terms.[8] We removed n-grams occurring in fewer than 0.2% of all bills and more than 15% of bills. We also removed an n-gram if it accounted for more than 0.2% of all tokens or fewer than than 0.001% of all tokens. After this process, our vocabulary contained 4,743 unique n-grams.

We used the anchor legislators described in Section 2. We ran variational inference until the change in increase in the objective function was less than 0.01%.

5.1. Exploring topics and bills

In this section, we examine a fit of the ideal point topic model for all the bills and votes of a session. This demonstrates the model's use as an exploratory tool of political data. For this analysis, we used dispersion

[7] The estimate $\mathbb{E}_q[\sigma(X_u(\bar{z}_d\eta_b)+\bar{z}_d\eta_a)]$ can be more theoretically justified, but results from the two estimates are (in practice) identical.

[8] When one n-gram subsumes another, we chose to observe the longer of the two

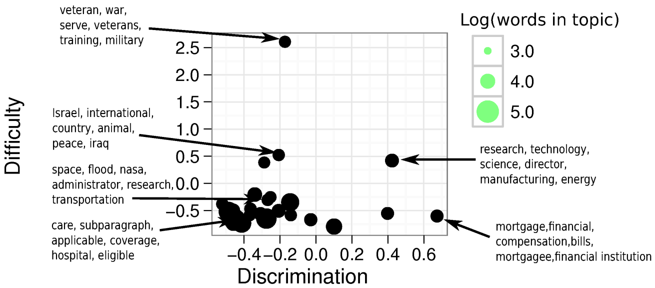

Figure 4. Topics can be visualized in the same latent political space as legislators and bills. This plot shows selected topics by coefficients $\hat{\eta}$, for a 64-topic model ($\hat{\eta}$s are normalized by mean and variance). Two topics (*people, month, recognize, ...* and *clause, motion, chair, ...*) with difficulty 4.68 and discrimination 7.4 (respectively) are not shown.

$\sigma_d = \sigma_u = 1.0$ and 64 topics. We focus on the 111^{th} session (January 2009 to January 2011).

Exploring topics with $\hat{\eta}$. As noted in Section 2, the coefficients $\hat{\eta}$ relate each topic's weight in a bill with the bill's difficulty and discrimination parameters. Figure 4 shows some example topics and their corresponding coefficients $\hat{\eta}$. Below we describe some of these topics in more detail and connect them to the data.

One popular topic in the 111th Congress focused on national recognition: *people, month, recognize, history, week, woman*. In contrast, the *least*-supported topic was more procedural, frequently appearing in bills under consideration or with many amendments (*clause, motion, chair, print, offer, read*). In this case, such legislation is sometimes summarily rejected before further consideration; the language of amendments is a signal that legislation is contentious.

While these topics often explained overwhelming support or rejection of legislation, much legislation was considerably more partisan.

Health Care. One contentious topic was about qualification for public health care: *care, subparagraph, applicable, coverage, hospital, eligible*. This topic was among the most-Democratic 10% of topics, in large part because it helped to explain the *Patient Protection and Affordable Care Act*, i.e. the "Health Care Bill" of 2009. Although this 906-page bill was barely passed: of the 311 Democrats voting on it, 276 voted in favor; of the 217 Republicans voting on it, none voted in favor. The model was moderately accurate on this bill: it correctly predicted 93.8% of votes. The two other topics highly expressed in this bill were about different aspects of public health, including one about government health options (medicare and social security) and one about health insurance coverage; both were slightly Democratic.

NASA Authorization. Another contentious topic was about spaceflight: *space, flood, NASA, administrator, research, transportation*. This topic was expressed in one of the most-poorly predicted bills of the 111^{th} Congress. This bill, the *NASA Authorization Act of 2010*, was a "compromise between the Obama administration, which wants... a commercial space industry in which private companies would transport astronauts, and House lawmakers, who wanted... one government-owned rocket" (Herszenhorn, 2010). In the house vote (a Senate record was not kept), of 249 Democrats voting on the bill, 185 voted in favor; of the 173 Republicans, 119 voted in favor. Because this bill had mixed but nonpartisan support, the model could not represent it well, with only 72% of votes correctly predicted.

5.2. Checking the ideal points

We can also use the in-sample fit to assess the quality of the ideal points of the legislators. In classical ideal point modeling, this is done via in-sample accuracy: How well does the model explain the observed votes?

The average per-legislator accuracy in the in-sample fit was 96% (only 10% of legislators had accuracy lower than 90%). As expected, accuracy increases with more votes ($\rho = 0.51$). Among legislators with over 100 votes, only two stand out. Donald Young (713 votes; accuracy 0.83) had a pre-defined ideal point (see Section 2). Ron Paul, a Republican in the 111^{th} Congress, was also poorly predicted (761 votes; accuracy 0.84). Paul is known for his Libertarian beliefs, even having run for President for the Libertarian party in 1988.

The poor prediction of Paul points to a limitation of the 1-dimensional ideal-point model, which can only capture the two main parties, instead of a limitation of the supervised prediction: fitting votes to the classical ideal point model (ignoring bill text), Paul's in-sample accuracy was consistently poor across sessions.

5.3. Predicting votes from text

Prediction on heldout bills. We measured predictive accuracy and log likelihood for these models under a variety of regularization settings (LARS is parameterized by $0 < f \leq 1$, L2 is parameterized by $\Lambda \geq 0$, and IPTM is parameterized by topics K).

We also devised two baselines for comparison with the three models described so far. The first of these provides a lower bound: assume all votes are yea. Because the majority (85%) of votes in our corpus were yea votes, this presents a more reasonable overall baseline than random guessing (at 50%). We call this model the yea model. The second baseline fit a logistic regression trained for members of each party (with a separate one for mixed or independent legislators), with terms as covariates. This baseline (implemented with the R glm library) used too much memory to use more than 800 terms and therefore led to results worse than the yea baseline.

For each 2-year period (called a Congress), the bills were partitioned into 6 folds. For each model, we iteratively (1) remove a fold, (2) fit the model to the remaining folds (by Congress), and (3) form predictions on the bills in the removed fold. Across folds, we thus obtain a complete data set of held-out votes.

Across all sessions, the yea baseline predicts votes correctly 85% of the time. The ideal point topic model is better, correctly predicting 89% of votes with 64 topics (this means that 62,000 more votes are correctly predicted). Overall performance for L2 was best for $\Lambda = 1000$ (90%), and LARS was best at $f = 0.01$ (82%). While the ideal point topic model had lower accuracy than L2, its log-likelihood was nearly the same. These results are summarized in Figure 4, and further details are in the supplementary materials.

Sequential prediction. Our final study examined the performance of these models on predicting future votes from past votes. To do this, we fit a 64-topic IPTM and L2 predictive models on the first $3, 6, 9, \ldots, 21$ months of a Congress.[9] We then tested these each of these fits on the following three months of unseen votes. The ideal-point topic model correctly predicted 87.0% of votes, and L2 correctly predicted 88.1% of votes; their log-likelihood was identical.

With these models, one could predict 31,000 to 55,000 votes above the baseline, *based only on the text of the bills*. The simpler of the two models, L2, performs better at prediction.

6. Future directions and summary

The text-regression models and the ideal point topic model have incorporated bill texts into the simplest kind of ideal point model of roll call data, although these have only scratched the surface of this field. We suggest several avenues for further work.

Here we have studied multiple topics with a one-dimensional political space. As noted in Section 5, this is a predictive bottleneck.[10] Solutions include increasing the dimension of the legislator and bill variables, a mixture model as in Wang et al. (2010), or modeling individual users' affinities to topics, as in Agarwal & Chen (2010).

One of the central advantages of generative probabilistic models is their modularity. Another avenue of future work is to incorporate other elements of the legislative process, such as speech transcripts (Quinn et al., 2006; Thomas et al., 2006) and auxiliary features such as bill sponsor, into this model's supervision, to improve both the predictive power and exploratory capabilities of the ideal point topic model.

We have developed several models associating the text of legislation to legislators' voting patterns. These models provide a way of exploring large collections of legislative data and predicting the votes of new bills. Though we were motivated by (and focused on) political science data, we note that these models are among several (as, e.g., (Agarwal & Chen, 2010)) that can be applied in a variety of collaborative filtering settings. They provide a way to model a collection of users and their decisions about collections of textual items.

Acknowledgments

We thank the reviewers for their helpful comments. David M. Blei is supported by ONR 175-6343, NSF CAREER 0745520, AFOSR 09NL202, the Alfred P.

[9] A bug prevented LARS from completing in most runs of this setting

[10] The "true" number of dimensions is debatable. See Heckman & Snyder (1996) and Jackman (2001).

Sloan foundation, and a grant from Google.

References

Agarwal, Deepak and Chen, Bee-Chung. flda: matrix factorization through latent dirichlet allocation. *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pp. 91–100, 2010.

Blei, David M. and McAuliffe, Jon D. Supervised topic models. 2008.

Blei, David M., Ng, Andrew Y., and Jordan, Michael I. Latent Dirichlet allocation. *JMLR*, 2003.

Clinton, J., Jackman, S., and Rivers, D. The statistical analysis of roll call data,. *American Political Science Review*, 98(2), 2004.

Enelow, J. and Hinich, M. *The Spatial Theory of Voting: an Introduction*. Cambridge University Press, New York, 1984.

Heckman, J. and Snyder, J. Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators. *RAND Journal of Economics*, 27(0), 1996.

Herszenhorn, David M. Congress wraps up session early as midterm races loom. In *The New York Times*. September 29 2010. Accessed 17 Jan 2011 at http://www.nytimes.com/2010/09/30/us/politics/-30cong.html.

Jackman, Simon. Multidimensional analysis of roll call data via bayesian simulation: Identification, estimation, inference, and model checking. *Political Analysis*, 9(3), 2001.

Johnson, Valen E. and Albert, James H. *Ordinal Data Modeling*. Springer-Verlag, New York, 1999.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models. *Learning in Graphical Models*, 1999.

Kogan, S., Levin, D., Routledge, B.R., Sagi, J.S., and Smith, N.A. Predicting risk from financial reports with regression. In *ACL Human Language Technologies*, pp. 272–280. Association for Computational Linguistics, 2009.

Lord, Frederic M. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Eribaum Associates, Inc., 1980.

Martin, Andrew D. and Quinn, Kevin M. Dynamic ideal point estimation via markov chain monte carlo for the u.s. supreme court, 1953-1999. *Political Analysis*, 10:134–153, 2002.

Pang, Bo and Lee, Lillian. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), 2008. doi: 10.1561/1500000001.

Poole, K. T. and Rosenthal, H. A spatial model for legislative roll call analysis. *American Journal of Political Science*, 1985.

Quinn, Kevin M., Monroe, Burt L., Colaresi, Michael, Crespin, Michael H., and Radev, Dragomir R. An automated method of topic-coding legislative speech over time with application to the 105th-108th u.s. senate. In *In Midwest Political Science Association Meeting*, 2006.

Salakhutdinov, R. and Mnih, A. Probabilistic matrix factorization. *Advances in Neural Information Processing Systems*, 20:1257–1264, 2008.

Schmid, Helmut. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, September 1994. URL `http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf`.

Thomas, Matt, Pang, Bo, and Lee, Lillian. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. *Proceedings of EMNLP*, pp. 327–335, 2006.

Wang, Eric, Liu, Dehong, Silva, Jorge, Dunson, David, and Carin, Lawrence. Joint analysis of time-evolving binary matrices and associated documents. In Lafferty, J., Williams, C. K. I., Shawe-Taylor, J., Zemel, R.S., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 23*, pp. 2370–2378. 2010.

Predicting Legislative Roll Calls from Text

1. Supplementary materials

A Variational inference

Inference for the ideal point topic model requires variational updates (see (Jordan et al., 1999) for more details about variational inference). Minimizing the KL between the variational distribution and the true posterior is equivalent to maximizing the following lower bound on the model evidence (called the "evidence lower bound", or ELBO):

$$\begin{aligned}
&\log p(\boldsymbol{W}, \boldsymbol{V}) = \\
&\quad \int p(\boldsymbol{W}, \boldsymbol{V} | \beta, \boldsymbol{\eta}, I, X, z, \theta) p(\beta, \boldsymbol{\eta}, I, X, z, \theta) \\
&\quad \geq \mathbb{E}_q \left[\sum_D \sum_N \log p(w_n | z_n, \beta) + \log p(z_n | \theta_d) \right] \\
&\quad\quad + \mathbb{E}_q \left[\sum_D \log p(A_d, B_d | z_{d,1:n}, \boldsymbol{\eta}) + \log p(\boldsymbol{\eta}) \right] \\
&\quad\quad + \mathbb{E}_q \left[\sum_U \log p(x_u) + \sum_D \log p(v_{ud} | x_u, A_d, B_d) \right] \\
&\quad\quad + \mathbb{E}_q \left[\sum_D \log p(\theta_d | \alpha) \right] + H(q) \\
&\quad\quad =: \mathcal{L}(\hat{\eta}, \kappa, \tau, \phi, \gamma),
\end{aligned} \tag{1}$$

where the expectations are taken with respect to the variational distribution q. This bound is optimized by block coordinate ascent. We repeatedly optimize each variational parameter until the relative increase in the lower bound is below a specified threshold.

One important detail in this equation is that $\mathbb{E}_q [\log p(v_{ud} | x_u, A_d, B_d)]$ is not available in closed form under the variational distribution. We approximate the expectation in Equation 1 by applying the second-order multivariate Delta method (Bickel & Doksum, 2007), also applied to the logit distribution in (Chang & Blei, 2009; Braun & McAuliffe, 2010). This Taylor approximation no longer guarantees that our objective is a lower bound; however, (Braun & McAuliffe, 2010) have found it to work better than a first-order approximation (which does maintain the lower bound).

We now turn to the coordinate updates.

Updates for $\boldsymbol{\eta}$ The variational update for $\hat{\eta}$ can be found by collecting terms in the evidence lower bound, taking the derivative with respect to $\hat{\eta}$, setting this to zero, and solving for $\hat{\eta}$. Letting $\boldsymbol{\kappa}_{\text{disc}}$ be a bill's discrimination parameters, we have the the exact update for the vector $\hat{\eta}_{\text{disc}}$:

$$\hat{\eta}_{\text{disc}} \leftarrow \left(\mathbb{E}_q \left[\bar{Z}^T \bar{Z} \right] + \frac{\sigma_d^2}{\sigma_\eta^2} \right)^{-1} \mathbb{E}_q \left[\bar{Z} \right]^T \boldsymbol{\kappa}_{\text{disc}}.$$

The update for $\hat{\eta}_{\text{diff}}$, controlling a bill's difficulty parameter $\boldsymbol{\kappa}_{\text{diff}}$, is analogous.

Updates for β, ϕ, and γ The updates for β and γ are exactly as in LDA (Blei et al., 2003), and the update for ϕ is exactly as in sLDA (Blei & McAuliffe, 2008); we omit details here.

Updates for κ_d and τ_u We cannot solve for κ and τ exactly, so they must be optimized via gradient ascent. For bill d, the gradient with respect to κ is

$$\begin{aligned}
&\nabla_{\kappa_{d,i}} \mathcal{L}(\kappa_{d,i}) = \\
&\quad \textstyle\sum_D - \frac{\kappa_{d,i} - \eta_i \bar{\phi}}{\sigma_d^2} + \sum_{v \in V(u)} 1_v \lambda_{u_v,i} - \lambda_{u_v,i} \rho_{ud} \\
&\quad - \textstyle\sum_{v \in V(d)} \frac{1}{2} \Big(\big(\sigma_\kappa^2 (\lambda_{u_v}^T \lambda_{u_v}) + \sigma_\lambda^2 (\kappa_d^T \kappa_d) \big) \\
&\quad\quad \times \lambda_{u_v,i} \left(\rho_{ud} - 2\rho_{ud}^2 + 2\rho_{ud}^3 \right) \Big) \\
&\quad - \textstyle\sum_{v \in V(d)} \frac{1}{2} \sigma_\lambda^2 \Big(\kappa_{d,i} \circ \left(\rho_{ud} - \rho_{ud}^2 \right) \Big),
\end{aligned} \tag{2}$$

where $\rho_{ud} = \frac{\exp(\tau_u^T \kappa_d - a_d)}{\exp(\tau_u^T \kappa_d - a_d) + 1}$ and 1_v is an indicator describing whether vote v was a `yea`-vote.

To optimize this, we apply second-order gradient ascent to the sum $\sum_d \frac{\partial \mathcal{L}}{\partial \kappa_d}$, repeating the updates

$$\kappa_d^n = \kappa_d^{n-1} - \frac{1000}{1000 + n^{0.6}} H^{-1} \left(\nabla_{\kappa_d} \mathcal{L}(\kappa_d) \right)$$

until convergence. In implementation, we constructed the Hessian H numerically by evaluating the above gradient with coordinates perturbed by 10^{-5}. For the data we used, this was sufficiently fast; if a bill has enough votes, an alternative implementation might use more frequent updates and fewer iterations through the votes.

The gradient for the user-ideal parameter τ_u is nearly identical to that for κ:

$$\begin{aligned} \nabla \tau_{u,i} \mathcal{L}(\tau_{u,i}) = & \\ & \textstyle\sum_U -\frac{\tau_{u,i}}{\sigma_u^2} + \sum_{v \in V(u)} 1_v \kappa_{d_v,i} - \kappa_{d_v,i} \rho_{ud} \\ & - \textstyle\sum_{v \in V(u)} \frac{1}{2} \Big(\left(\sigma_\tau^2 (\kappa_{d_v}^T \kappa_{d_v}) + \sigma_\kappa^2 (\tau_u^T \tau_u) \right) \\ & \quad \times \kappa_{d_v,i} \left(\rho_{ud} - 2\rho_{ud}^2 + 2\rho_{ud}^3 \right) \Big) \\ & - \textstyle\sum_{v \in V(u)} \frac{1}{2} \sigma_\kappa^2 \Big(\tau_{u,i} \circ \left(\rho_{ud} - \rho_{ud}^2 \right) \Big). \end{aligned} \tag{3}$$

Again, we update this via second-order gradient ascent.

Updates for σ_κ and σ_λ. Once per iteration, we update the the variances σ_κ and σ_λ. As with $\hat{\eta}$, these updates are exact:

$$\begin{aligned} \sigma_\kappa^2 & \leftarrow \frac{ND}{\sum_{D, v \in V(d)} \tau_u^T \tau_u (\rho_{u_v d} - \rho_{u_v d}^2)_n + ND / \sigma_d^2} \\ \sigma_\tau^2 & \leftarrow \frac{NU}{\sum_{U, v \in V(u)} \kappa_d^T \kappa_d (\rho_{u d_v} - \rho_{u d_v}^2)_n + NU / \sigma_u^2}, \end{aligned}$$

where above we have U users, D bills, and an N-dimensional ideal-point model.

B Implementation details

We provided details of a variational implementation of the ideal point topic model. Here we describe several modifications to improve this algorithm.

Second order updates. Note that the second-order updates for κ and τ may violate the convexity assumption. To mitigate this, and to prevent the parameters from diverging for large σ_d or σ_u, we add a constant to each element of the diagonal (Levenberg, 1944). We add a sufficiently large constant to guarantee that all 1×1 and 2×2 principal minors have positive determinant (this is necessary but not sufficient to guarantee that H is positive definite). We have observed that H only requires this adjustment for early model iterations.

Identifiability. In the modeling section, we discussed using nonzero priors for certain legislators to make the posterior identifiable. These priors may not be sufficient to guarantee that the model finds specific modes. To encourage the model to converge to the desired optimum, we allow the first two iterations of this model one extra dimension for the ideal point. We believe this "blessing of dimensionality" allows the model to rotate ideal points toward the desired mode.

Annealing. We set the model parameters y for σ_d^2 to 1.0 before the first iteration and update it with $y \leftarrow y^{0.9} (\sigma_d^2)^{0.1}$ in a form of "variational annealing". We apply the same annealing to σ_u.

Model	Regularization	Accuracy	Log Likelihood	Expected Correct Probability
lars	0.001	0.819	**-0.855**	0.792
lars	0.01	**0.822**	-0.984	**0.793**
lars	0.03125	0.817	-1.091	0.792
lars	0.0625	0.807	-1.214	0.787
lars	0.125	0.799	-1.337	0.781
lars	0.25	0.786	-1.479	0.770
lars	0.5	0.770	-1.640	0.755
lars	1	0.735	-1.903	0.723
l2	0.01	0.815	-0.914	0.793
l2	0.1	0.832	-0.794	0.811
l2	1	0.850	-0.636	0.829
l2	10	0.876	-0.498	0.853
l2	100	0.891	-0.371	0.866
l2	1000	**0.897**	**-0.302**	**0.868**
l2	10000	0.873	-0.324	0.841
iptm	4	0.871	-0.370	0.849
iptm	8	0.869	-0.348	0.845
iptm	16	0.883	-0.321	0.858
iptm	32	0.883	-0.314	0.856
iptm	64	**0.887**	**-0.306**	**0.858**
iptm	128	0.873	-0.456	0.845
yea		0.853	-0.417	0.749

Figure 1. Prediction metrics for heldout prediction experiments.

2. Experimental Results.

The experimental results for cross-fold validation are presented in Figure 2. Top performers by various metrics are highlighted in bold.

We also display ideal points for all Senators (Figure 4) and all legislators (Senators and House representatives) (Figure 3) in the fit of the 111th Congress.

Model	Accuracy	Log Likelihood	Expected Correct Probability
l2	0.881	-0.346	0.852
iptm	0.870	-0.346	0.824
yea	0.851	-0.422	0.746

Figure 2. Prediction metrics for time-series prediction experiments.

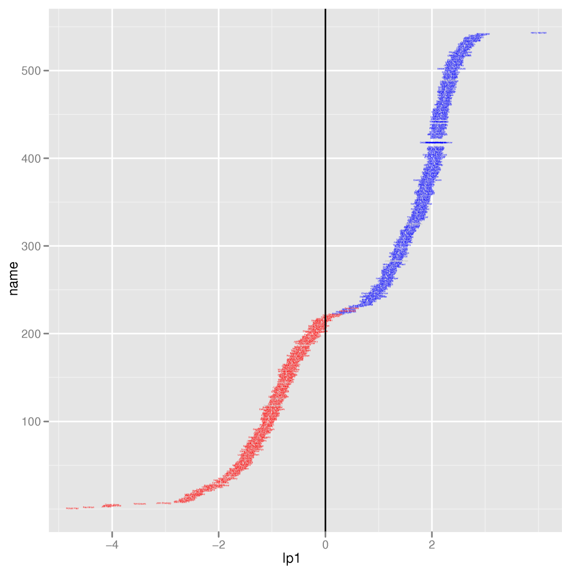

Figure 3. All legislator ideal points in the 111th Congress. Using votes, ideal points can separate the U.S. political parties Democrats (blue) and Republicans (red).

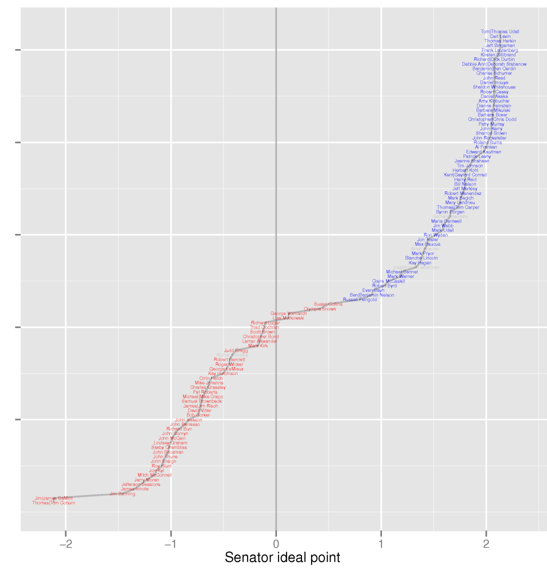

Figure 4. All Senator ideal points in the 111th Congress. Using votes, ideal points can separate the U.S. political parties Democrats (blue) and Republicans (red).

References

Bickel, Peter J. and Doksum, Kjell A. *Mathematical statistics: Basic ideas and selected topics*, volume 1. Pearson Prentice Hall, 2nd ed edition, 2007.

Blei, David M. and McAuliffe, Jon D. Supervised topic models. 2008.

Blei, David M., Ng, Andrew Y., and Jordan, Michael I. Latent Dirichlet allocation. *JMLR*, 2003.

Braun, Michael and McAuliffe, Jon. Variational inference for large-scale models of discrete choice. *Journal of the Americal Statistical Association*, 105 (489):324–334, 2010.

Chang, Jonathan and Blei, David M. Relational topic models for document networks. *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AIStats) 2009*, 5, 2009.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models. *Learning in Graphical Models*, 1999.

Levenberg, K. A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, 2(2):164–168, July 1944.