The Ideal Point Topic Model: Predicting Legislative Roll Calls from Text

Sean Gerrish
Princeton University
Computer Science Department
sgerrish@cs.princeton.edu

David Blei
Princeton University
Computer Science Department
blei@cs.princeton.edu

Abstract

We develop the *ideal point topic model*, a probabilistic model of legislative text. Our model – drawing on ideas from ideal point estimation and topic modeling – predicts voting patterns based on the contents of bills and the inferred political leanings of legislators. It also provides an exploratory window into how legislative language is correlated with political support. Across 14 years of legislative data, we predict specific voting patterns with high accuracy.

Introduction

Legislative behavior has long been scrutinized by lobbyists, citizens, and political scientists. In Western politics, much of this scrutiny has focused on roll-call votes: tables of `yea` or `nay` votes which indicate an individual legislator's sentiment for or against items of legislation.

These models have remained largely observational in nature; legislators are summarized, and their votes on past legislation is analyzed, once all votes are observed. However, a glaring lack of *predictive* models have been developed for automatically inferring legislators' positions given the *text* of legislation. In this work, we create a predictive probabilistic topic model for roll-call votes. Given the text of legislation, this model is able to predict how each legislator will vote on it; at the same time, the model is an exploratory tool which can summarize topical issues.

Model

A latent-space voting model. The *ideal point topic model* (refer to Figure 1) summarizes each legislator's preferences with a real-valued hidden random variable x_i, called an *ideal point* in modern statistical voting literature (see Figure 2 for a sample of senators' ideal points) [3, 5, 4, 8].

Legislative items such as bills and resolutions are each assigned two hidden sentiment parameters: a *discrimination* parameter a_d describing how partisan the bill is; and a *difficulty* parameter b_d, an intercept term describing how likely anyone is to vote yea on each item. The probability

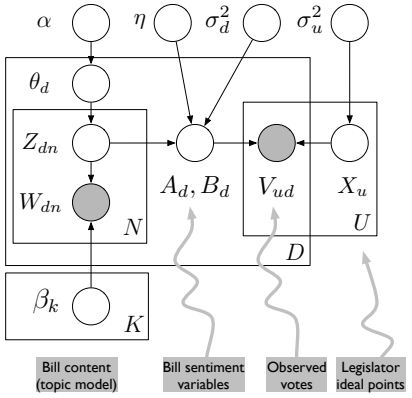

Figure 1: The ideal point topic model.

of a yea-vote by legislator i on a specific piece of legislation d is then given by the logistic function

$$p(yea|x_i, a_d, b_d) \sim \sigma(x_i a_d + b_d) := \frac{\exp(x_i a_d + b_d)}{\exp(x_i a_d + b_d) + 1}$$

This logistic function can be motivated by a model of behavioral choice (the probit function can be similarly motivated) [4]. With this setup, ideal points learned by the model provide a lens into legislators' political leanings. Because the U.S. has two primary political parties which tend to vote in blocks, ideal points indicate party affiliation; see Figure 2.

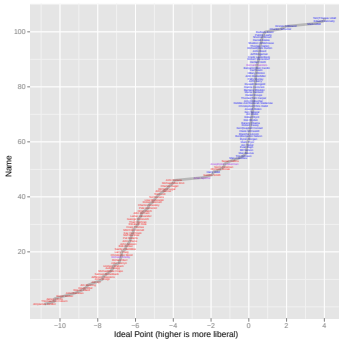

Figure 2: Senators by ideal point, with Republicans (red) and Democrats (blue).

Supervised Topics. We turn this model into a discriminative model with supervised topics [1]. The supervised topic model represents text documents with the generative assumptions of Latent Dirichlet Allocation: each document is a mixture θ_d of *topics*, or distributions over word counts [2]. These documents' topics are then used to predict documents' sentiment parameters a_d, b_d with a linear model. The topics β naturally adjust during inference to improve the model's predictive performance.

Posterior Inference. Note that the only observed variables in the model are the text of each legislative item and an incomplete user-item matrix of votes. To estimate values of the hidden random variables, we use posterior inference. To fit this posterior, we derived fast variational updates [6] [1]. which converge much more quickly than Gibbs sampling (the current state-of-the-art).

Experiments: Analyzing the U.S. House and Senate

We studied the performance of the ideal point topic model on 14 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 its generalization performance as a way to predict votes from bill texts.

We collected roll-coll votes for Congresses 104 through 110 (January 1995 to January 2009). Only votes regarding bills and resolution final passage was included (as opposed to amendments of the legislation). We downloaded the data from Govtrack, an independent Website which provides comprehensive tracking of legislative information to the public. Our collection contains 4,915 documents, 1,253 unique legislators, and 1,802,767 `yea` or `nay` roll-call votes (all other voting information, such as abstentions, were removed from the dataset).

To select tokens, we first lemmatized the documents with Treetagger [10]. Then we retained a vocabulary of statistically significant n-grams ($1 \leq n \leq 5$). After this, our vocabulary contained 4,063 unique n-grams (used as tokens) with 667,648 total n-gram observations.

In all of the fits described below, we ran coordinate ascent variational inference until the objective function increased by no more than 0.001%.

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 $\sigma_d = 0.001$ and 32 topics. We focus on the 109th session (January 2005 to January 2007).

The regression coefficients of the ideal point topic model $\hat{\eta}$ provide a window into how the prevalence of each topic in a bill correlates with its corresponding difficulty and discrimination parameters. When used for exploration, this parameter demonstrates which topics are likely to receive `yea`

[1] Correlation between legislators' ideal points with our variational and existing (MCMC) methods exceeds 0.98.

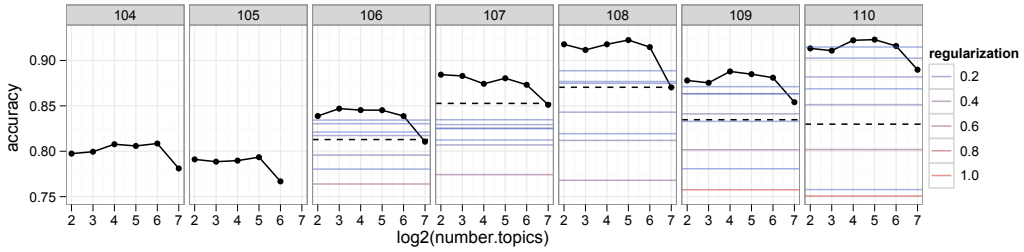

Figure 3: Accuracy by session. The ideal point topic model is shown in solid black for different numbers of topics. The *yea* baseline is the horizontal dotted black line. *LARS* for different regularization parameters is shown in other horizontal lines. Note that for the sessions 104-105, *LARS* and *yea* are below the fold.

votes regardless of political party (difficulty) and which are likely to receive partisan votes (discrimination). Thus, the topics are positioned in the same latent space as the bills. Below we describe some of these topics in more detail and connect them to the data.

Health care. The topic about health care has a low (but positive) difficulty and is on the right wing. A document that contains many words assigned to this topic is the *Help Efficient, Accessible, Low-cost, Timely Healthcare (HEALTH) Act of 2005*. At a time of rising health costs, the HEALTH Act was proposed to "reduce the incidence of 'defensive medicine' and lower the cost of health care liability insurance, all of which contribute to the escalation of health care costs". This bill was only voted on in the House. It passed with 53% of the votes, 93% of which were Republican.

We can consider the twenty bills which most express this topic and examine their votes in aggregate. 69% of these votes were "Yea" (accounting for the positive difficulty) and Democrats accounted for only 20% of those votes (accounting for the far right wing position).

Immigration The topic about immigration has negative difficulty and is slightly on the left wing. One of the most topical documents in this topic was the *Securing America's Borders Act*, which was weakly favored by Republicans but failed a cloture motion (i.e., a filibuster by Democrats). This topic also included the *Comprehensive Immigration Reform Act of 2006*, which was moderately Democratic and passed the House. Looking at the top twenty bills that most express this topic, only 48% of the votes are "Yea" (accounting for the negative difficulty) and Democrats casted 53% of these positive votes (accounting for the slight left leaning).

Banking The topic about banking has a high difficulty. The most topical documents in this topic were a *To provide regulatory relief and improve productivity for insured depository institutions* (passed unanimously) and the *Military Personnel Financial Services Protection Act* (passed overwhelmingly). Among the top twenty bills expressing this topic, 99% of votes were positive.

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 worst-predicted legislator was Ronald Paul, who was consistently predicted poorly across sessions, with average ideal point of 0.89 – slightly more Republican than average (for comparison, *National Journal*'s 2006 Conservative score placed Paul at the 39*th* percentile of Conservatives). Paul is a former member of the Libertarian party, even having run for President for the Libertarian party in 1988.

The poor prediction of Ron Paul appears to be a limitation of the ideal-point model instead of a limitation of the supervised document parameters. Indeed, Paul's accuracy with the classical ideal point model is also poor.

Predicting votes from bills

In addition to providing a new lens for exploring bill texts and votes, an advance of the ideal point topic model is that it can be used to predict roll calls before any votes have been cast. (The classical ideal point model can predict censored votes, but needs to have observed several votes to infer the discrimination and difficulty of the bill in question.)

We devised two baselines as points of comparison. The first of provides a lower bound: assume all votes are `yea`. Because the majority (78%) 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.

For the second baseline, we fit an ideal-point model to training documents and all legislators. Documents' parameters were then estimated on heldout validation documents with *Least-Angle Regression* (LARS), using *n*-gram counts as covariates (the same vocabulary was used). This was implemented using the `lars` package for `R` [9], for a range of regularization parameters $0 < s \leq 1$. (Unregularized linear regression performed extremely poorly and is not reported.) Note this is a form of "text regression" [7].

We evaluated the model using 6-fold cross validation. Across all sessions, the ideal point topic model predicted best with 32 topics, correctly predicting 88% of heldout votes. Overall performance for `lars` is best for $f = 0.03$ (81%). (The best ideal point topic model correctly predicts 126,000 more votes than the best lars setting.) Dividing the bills by session, Figure 3 compares the baselines to the ideal point topic model. Note that in addition to providing overall better performance, the ideal point topic model performs with less variance across its regularization parameter (the number of topics).

Future directions

The ideal point topic model provides a new way of exploring collections of legislative data and predicting votes of new bills. Because of its modularity, new features can easily be incorporated into the model for better prediction.

We finally emphasize the generality of this model. It can be applied out-of-the-box in many collaborative filtering settings: matching listeners to music, matching users to Web advertisements, and recommending books to readers, for example. It is a general model for individuals and their preferences.

References

[1] D. M. Blei and J. D. McAuliffe. Supervised topic models. 2008.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *JMLR*, 2003.

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

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

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

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

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

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

[9] M.-Y. Park and T. Hastie. An l1 regularization-path algorithm for generalized linear models. *JRSSB*, 69:659–677, 2007.

[10] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, September 1994.