Supplementary materials for How They Vote: Issue-Adjusted Ideal Point Models of Legislative Behavior

In this appendix we provide additional experiment and implementation details for *How They Vote*.

A.1. Variational posterior inference

We begin by providing more detail about the inference algorithm summarized in the Inference section of the main paper.

Optimizing the variational objective

Variational bounds are typically optimized by gradient ascent or block coordinate ascent, iterating through the variational parameters and updating them until the relative increase in the lower bound is below a specified threshold. Traditionally this would require symbolic expansion of $\mathbb{E}_q[p(v,x,\mathbf{z},a,b,\boldsymbol{\theta}) - q(x)]$, a process which presupposes familiarity with variational methods.

Instead of expanding this bound symbolically, we update each parameter by making Taylor approximations of the KL objective and performing a series of second-order updates to these parameters, iterating through the parameters and repeating until convergence.

To be concrete, we describe how to perform the mth update on the variational parameter $\tilde{x}$, assuming that we have the most-recent estimate $\tilde{x}_{m-1}$ of this parameter (updates for the other random variables are analogous). Writing the variational objective as $f(\tilde{x}) = \mathrm{KL}(q_{\tilde{x}}||p)$ for notational convenience (where all parameters in η except $\tilde{x}$ are held fixed), we estimate the KL divergence as a function of $\tilde{x}$ around our last estimate $\tilde{x}_{m-1}$ with its Taylor approximation

$$\begin{aligned} f_{m-1}(\tilde{x}) \approx & f(\tilde{x}_{m-1}) + \left(\frac{\partial f}{\partial \tilde{x}}\Big|_{\tilde{x}_{m-1}}\right)^T \Delta\tilde{x} \\ & + \frac{1}{2}\Delta\tilde{x}^T \left(\frac{\partial^2 f}{\partial\tilde{x}\partial\tilde{x}^T}\Big|_{\tilde{x}_{m-1}}\right)\Delta\tilde{x}, \end{aligned} \tag{1}$$

where $\Delta\tilde{x} = \tilde{x} - \tilde{x}_{m-1}$. Once we have estimated the Taylor coefficients (as described in the next section), we can perform the update

$$\tilde{x}_m \leftarrow \tilde{x}_{m-1} - \left(\frac{\partial^2 f_{m-1}}{\partial\tilde{x}\partial\tilde{x}^T}\Big|_{\tilde{x}_{m-1}}\right)^{-1}\left(\frac{\partial f_{m-1}}{\partial\tilde{x}}\Big|_{\tilde{x}_{m-1}}\right). \tag{2}$$

Taylor Coefficient approximation

We approximate the Taylor coefficients in Equation 2 above with Monte Carlo integration, taking samples from $q_{\tilde{x}_{m-1}}$, which is easy to sample from because it has known mean and variance. We approximated the Taylor coefficients by approximating the gradient of $f(\tilde{x}) = \mathrm{KL}(q_{\tilde{x}}||p)$ with samples: We will illustrate this approximate gradient with respect to the variational parameter $\tilde{x}$. Let

$\tilde{x}_0$ be the current estimates of the variational mean, $q_{\tilde{x}_0}(x, \boldsymbol{z}, a, b)$ be the variational posterior at this mean, and define $\mathcal{L}_{\tilde{x}_0} = \mathbb{E}_q\left[p(x_0, \boldsymbol{z}, a, b) - q(x_0, \boldsymbol{z}, a, b)\right]$.

We then approximate the gradient with Monte Carlo samples as

$$\frac{\partial \mathcal{L}_{\tilde{x}_0}}{\partial \tilde{x}}\Big|_{\tilde{x}_0} = \frac{\partial}{\partial \tilde{x}} \int q_{\tilde{x}}(x, \boldsymbol{z}, a, b)(\log p(x, \boldsymbol{z}, a, b, v) \tag{3}$$

$$- \log q_{\tilde{x}}(x, \boldsymbol{z}, a, b))dx d\boldsymbol{z} da db \tag{4}$$

$$= \int \frac{\partial}{\partial \tilde{x}} \left(q_{\tilde{x}}(x)(\log p(x, \boldsymbol{z}, a, b, v) - \log q_{\tilde{x}}(x, \boldsymbol{z}, a, b))\right) d\tilde{x}$$

$$= \int q_{\tilde{x}}(x) \frac{\partial \log q_{\tilde{x}(x)}}{\partial \tilde{x}} (\log p(x, \boldsymbol{z}, a, b, v) - \log q_{\tilde{x}}(x, \boldsymbol{z}, a, b)) d\tilde{x}$$

$$\approx \frac{1}{N}\Bigg(\sum_{n=1}^{N} \frac{\partial \log q_{\tilde{x}}(x_n, z_n, a_n, b_n)}{\partial \tilde{x}}\Big|_{\tilde{x}_0}$$

$$\times \left(\log p(x_n, z_n, a_n b_n, v) - C - \log q_{\tilde{x}_0}(x_n, z_n, a_n, b_n)\right)\Bigg),$$

where we have used N samples from the current estimate of the variational posterior.

$$\frac{\partial f}{\partial \tilde{x}}\Big|_{\tilde{x}_{m-1}} = \frac{\partial}{\partial \tilde{x}} \int q_{\tilde{x}}(x)(\log p(x) - \log q_{\tilde{x}}(x)) d\tilde{x} \tag{5}$$

$$= \int \frac{\partial}{\partial \tilde{x}} \left(q_{\tilde{x}}(x)(\log p(x) - \log q_{\tilde{x}}(x))\right) d\tilde{x}$$

$$= \int q_{\tilde{x}}(x) \frac{\partial \log q_{\tilde{x}(x)}}{\partial \tilde{x}} (\log p(x) - \log q_{\tilde{x}}(x)) d\tilde{x}$$

$$\approx \frac{1}{N}\Bigg(\sum_{n=1}^{N} \frac{\partial \log q_{m-1}(x_{m-1,n})}{\partial \tilde{x}_{m-1}}\Big|_{\tilde{x}_{m-1}}$$

$$\times \left(\log p(x_{m-1,n} | \boldsymbol{z}_{m-1,n}, x_{m-1,n}, a_{m-1,n}, b_{m-1,n}, V)\right.$$

$$\left. - C - \log q_{m-1}(x_{m-1,n})\right)\Bigg),$$

where we have taken the gradient through the integral using Liebniz's rule. The second Taylor coefficient is straightforward to derive with similar algebra:

$$\frac{\partial^2 f}{\partial \tilde{x} \partial \tilde{x}^T}\Big|_{\tilde{x}_{m-1}} \approx \frac{1}{N} \sum_{n=1}^{N} \Bigg(\left(\frac{\partial \log q_{m-1}(x_{m-1,n})}{\partial \tilde{x}}\Big|_{\tilde{x}_{m-1}} \right) \tag{6}$$

$$\times \left(\frac{\partial \log q_{m-1}(x_{m-1,n})}{\partial \tilde{x}}\Big|_{\tilde{x}_{m-1}} \right)^T$$

$$\times \left(\log p(x_{m-1,n} | \boldsymbol{z}_{m-1,n}, a_{m-1,n}, b_{m-1,n}, V)\right.$$

$$\left. - C - \log q_{m-1}(x_{m-1,n}) - 1\right)$$

$$+ \left(\left(\frac{\partial^2 \log q_{m-1}(x_{m-1,n})}{\partial \tilde{x} \partial \tilde{x}^T}\Big|_{\tilde{x}_{m-1}} \right)\right.$$

$$\times \left(\log p(x_{m-1,n} | \boldsymbol{z}_{m-1,n}, a_{m-1,n}, b_{m-1,n}, V)\right.$$

$$\left.\left. - C - \log q_{m-1}(x_{m-1,n})\right)\right)\Bigg),$$

where we increase N as the model converges. Note that C is a free parameter that we can set without changing the final solution. We set C to the average of $\log p(x_{m-1,n}|...) - \log q_{m-1}(x_{m-1,n})$ across the set of N samples.

Instead of taking *iid* samples from the variational distribution q_{M-1}, we used quasi-Monte Carlo sampling [1]. By taking non-*iid* samples from q_{m-1}, we are able to decrease the variance around estimates of the Taylor coefficients. To select these samples, we took N equally-spaced points from the unit interval, passed these through the inverse CDF of the variational Gaussian $q_{m-1}(x)$, and used the resulting values as samples.[1]

We did this for each random variable in the Markov blanket of x_u, permuted each variable's samples, and combined them for N multivariate samples $\{x_{m-1,n}, \ldots, B_{m-1,n}\}_n$ from the current estimate q_{m-1} of the variational distribution.

We estimate the gradients of $\log q$ above based on the distribution of the variational marginals. We have defined the variational distribution to be factorized Gaussians, so these take the form

$$\begin{aligned} \frac{\partial \log q_{m-1}(x_{m-1,n})}{\partial \tilde{x}}\Big|_{\tilde{x}_{m-1}} &= \frac{x_{m-1,n} - \tilde{x}_{m-1}}{\sigma_x^2} \\ \frac{\partial^2 \log q_{m-1}(x_{m-1,n})}{\partial \tilde{x}^2}\Big|_{\tilde{x}_{m-1}} &= -\frac{1}{\sigma_x^2}. \end{aligned} \tag{7}$$

The variance σ_x^2 was fixed to $\exp(-5)$. Allowing σ_x to vary freely provides a better variational bound at the expense of accuracy. This happens because the issue-adjusting model would sometimes fit poor means to some parameters when the posterior variance was large: there is little penalty for this when the variance is large. Low posterior variance σ_x^2 is similar to a non-sparse MAP estimate.

These updates were repeated until the exponential moving average $\Delta_{\text{est},i} \leftarrow 0.8\Delta_{\text{est},i-1} + 0.2\Delta_{\text{obs},i}$ of the change in KL divergence dropped below one and the number N of samples passed 500. If the moving average dropped below one and $N < 500$, we doubled the number of samples.

For all experiments, we began with $M = 21$ samples to estimate the approximate gradient and scaled it by 1.2 each time the Elbo dropped below a threshold, until it passed 500.

Sparsity.

Issue adjustments z_u ranged widely, moving some lawmakers significantly. The variational estimates were not sparse, although a high mass was concentrated around 0. Twenty-nine percent of issue adjustments were within $[-0.01, 0.01]$, and eighty-seven percent of issue adjustments were within $[-0.1, 0.1]$.

Numerical stability and hyperparameter sensitivity

We address practical details of implementing issue-adjusted ideal points.

Hyperparameter settings

The most obvious parameter in the issue voting model is the regularization term λ. The Bayesian treatment described in the Inference section of *How they Vote* demonstrated considerable robustness to overfitting at the expense of precision. With $\lambda = 0.001$, for example, issue adjustments z_{uk} remained on the order of single digits, while the MAP estimate yielded adjustment estimates over 100.

We recommend a modest value of $1 < \lambda < 10$. At this value, the model outperforms ideal points in validation experiments consistently in both the House and Senate.

Implementation.

When performing the second-order updates described in the Inference section, we skipped variable updates when the estimated Hessian was not positive definite (this disappeared when sample sizes grew large enough). We also limited step sizes to 0.1 (another possible reason for smaller coefficients).

[1] Note that these samples produce a biased estimate of Equation 1. This bias decreases as $N \to \infty$.

A.2. Issue labels

In the empirical analysis, we used issue labels obtained from the Congressional Research Service. There were 5,861 labels, ranging from *World Wide Web* to *Age*. We only used issue labels which were applied to at least twenty five bills in the 12 years under consideration. This filter resulted in seventy-four labels which correspond fairly well to political issues. These issues, and the number of documents each label was applied to, is given in Table 1.

Table 1: Issue labels and the number of documents with each label (as assigned by the Congressional Research Service) for Congresses 106 to 111 (1999 to 2010).

Issue label	Bills
Women	25
Military history	25
Civil rights	25
Government buildings; facilities; and property	26
Terrorism	26
Energy	26
Crime and law enforcement	27
Congressional sessions	27
East Asia	28
Appropriations	28
Business	29
Congressional reporting requirements	30
Congressional oversight	30
Special weeks	31
Social services	31
Health	33
Special days	33
California	33
Social work; volunteer service; charitable organizations	33
State and local government	34
Civil liberties	35
Government information and archives	35
Presidents	35
Government employees	35
Executive departments	35
Racial and ethnic relations	36
Sports and recreation	36
Labor	36
Special months	39
Children	40
Veterans	40
Human rights	41
Finance	41
Religion	42
Politics and government	43
Minorities	44
Public lands and natural resources	44

Issue label	Bills
Europe	44
Military personnel and dependents	44
Taxation	47
Government operations and politics	47
Postal facilities	47
Medicine	48
Transportation	48
Emergency management	48
Sports	52
Families	53
Medical care	54
Athletes	56
Land transfers	56
Armed forces and national security	56
Natural resources	58
Law	60
History	61
Names	62
Criminal justice	62
Communications	65
Public lands	68
Legislative rules and procedure	69
Elementary and secondary education	74
Anniversaries	82
Armed forces	83
Defense policy	92
Higher education	103
Foreign policy	104
International affairs	105
Budgets	112
Education	122
House of Representatives	142
Commemorative events and holidays	195
House rules and procedure	329
Commemorations	400
Congressional tributes	541
Congress	693

Corpus preparation

In this section we provide further details of vocabulary selection. We began by considering all phrases with one to five words. From these, we immediately ignored phrases which occurred in more than 10% of bills and fewer than 4 bills, or which occurred as fewer than 0.001% of all phrases. This resulted in a list of 40603 phrases.

Table 2: Features and coefficients used for predicting "good" phrases. Below, test is a test statistic which measures deviation from a model assuming that words appear independently; large values indicate that they occur more often than expected by chance. We define it as test $= \frac{\text{Observed count} - \text{Expected count}}{\sqrt{\text{Expected count under a language model assuming independence}}}$

Coefficient	Summary	Weight
log(count + 1)	Frequency of phrase in corpus	-0.018
log(number.docs + 1)	Number of bills containing phrase	0.793
anchortext.presentTRUE	Occurs as anchortext in Wikipedia	1.730
anchortext	Frequency of appearing as anchortext in Wikipedia	1.752
frequency.sum.div.number.docs	Frequency divided by number of bills	-0.007
doc.sq	Number of bills containing phrase, squared	-0.294
has.secTRUE	Contains the phrase *sec*	-0.469
has.parTRUE	Contains the phrase *paragra*	-0.375
has.strikTRUE	Contains the phrase *strik*	-0.937
has.amendTRUE	Contains the phrase *amend*	-0.484
has.insTRUE	Contains the phrase *insert*	-0.727
has.clauseTRUE	Contains the phrase *clause*	-0.268
has.provisionTRUE	Contains the phrase *provision*	-0.432
has.titleTRUE	Contains the phrase *title*	-0.841
test.pos	$\ln(max(-\text{test}, 0) + 1)$	0.091
test.zeroTRUE	1 if test $= 0$	-1.623
test.neg	$\ln(max(\text{test}, 0) + 1)$	0.060
number.terms1	Number of terms in phrase is 1	-1.623
number.terms2	Number of terms in phrase is 2	2.241
number.terms3	Number of terms in phrase is 3	0.315
number.terms4	Number of terms in phrase is 4	-0.478
number.terms5	Number of terms in phrase is 5	-0.454
log(number.docs + 1) * anchortext	$\ln(\text{Number of bills containing phrase}) \times 1_{\{\text{Appears in Wikipedia anchortext}\}}$	-0.118
log(count + 1) * log(number.docs + 1)	$\ln(\text{Number of bills containing phrase} + 1) \times \ln(\text{Frequency of phrase in corpus} + 1)$	0.246

We then used a set of features characterizing each word to classify whether it was good or bad to use in the vocabulary. Some of these features were based on corpus statistics, such as the number of bills in which a word appeared. Other features used external data sources, including whether, and how frequently, a word appeared as link text in a Wikipedia article. For training data, we used a manually curated list of 458 "bad" phrases which were semantically awkward or meaningles (such as *the follow bill*, *and sec ammend*, *to a study*, and *pr*) as negative examples in a L_2-penalized logistic regression to select a list of 5,000 "good" words.

A.3. Summary of corpus statistics

We studied U.S. Senate and House of Representative roll-call votes from 1999 to 2010. This period spanned Congresses 106 to 111 and covered an historic period in U.S. history, the majority of which Republican President George W. Bush held office. Bush's inauguration and the attacks of September 11th, 2001 marked the first quarter of this period, followed by the wars in Iraq and Afghanistan. Democrats gained a significant share of seats from 2007 to 2011, taking the majority from Republicans in both the House and the Senate, and Democratic President Obama was inaugurated in January 2009. A summary of statistics for our datasets in these Congresses is provided in Table 3.

Table 3: Roll-call data sets used in the experiments. These counts include votes in both the House and Senate. Congress 107 had fewer fewer votes than the remaining congresses in part because this period included large shifts in party power, in addition to the attacks on September 11th, 2001.

Congress	**Years**	**Lawmakers**	**Bills**	**Votes** (Senate)
106	1999-2000	516	391	149,035 (7,612)
107	2001-2002	391	137	23,996 (5,547)
108	2003-2004	539	527	207,984 (7,830)
109	2005-2006	540	487	194,138 (7,071)
110	2007-2008	549	745	296,664 (9,019)
111	2009-2010	552	826	336,892 (5,936)

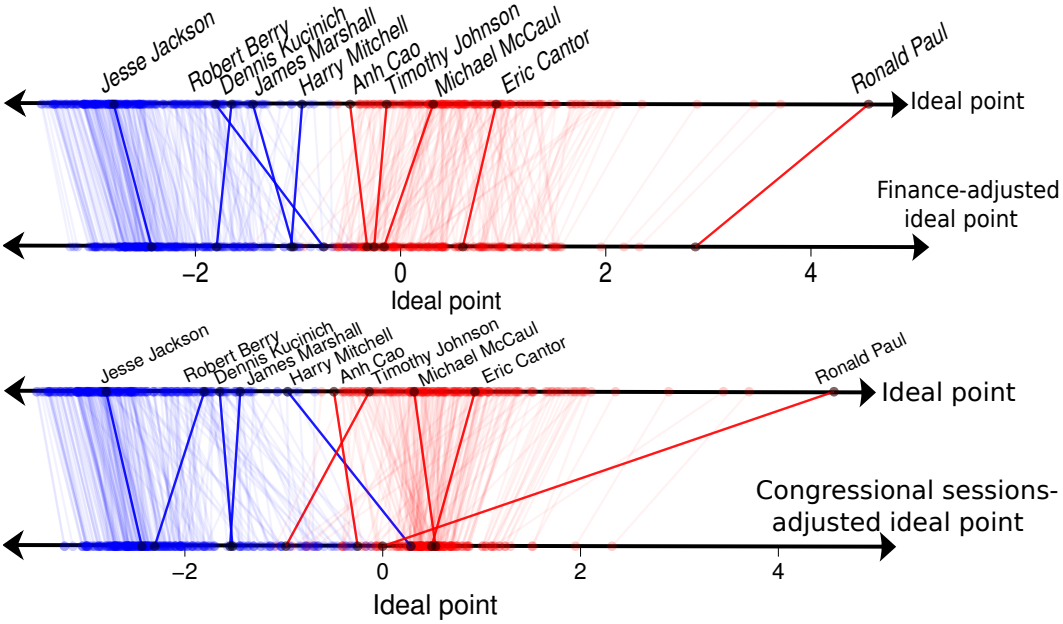

Figure 1: Ideal points x_u and issue-adjusted ideal points $x_u + z_{uk}$ from the 111th House. Democrats are blue and Republicans are red. Votes were most improved for the issue *Congressional sessions*, which focuses on procedural matters such as when to adjourn for a House recess or whether to consider certain legislation. Lawmakers were split into factions: some became further polarized by these bills, but some did not; the resulting mixture was not on party lines. Votes about Finance were also better fit with this model. Democrats were mostly fixed on this issue, but Republicans (who were less-well predicted by ideal points alone) saw more adjustment.

A.4. Additional figures

Figure 1 shows lawmakers offsets for two different issues. This exemplifies how much lawmakers diverge from a one-dimensional ideal point model.

Figure 2 illustrates the extent to which the issue-adjusted ideal point model improves prediction for different issues. These values were computed over a fit of the model to all votes in the 111th House of Representatives.

A.5. Controlling for lawmakers' ideal point x_u in issue adjustments

Controlling for ideal points

The issue-adjusted ideal point model is under-specified in several ways. It is well known that the signs of ideal points x_u and bill polarities a_d are arbitrary, for example, because $x_u a_d = (-x_u)(-a_d)$.

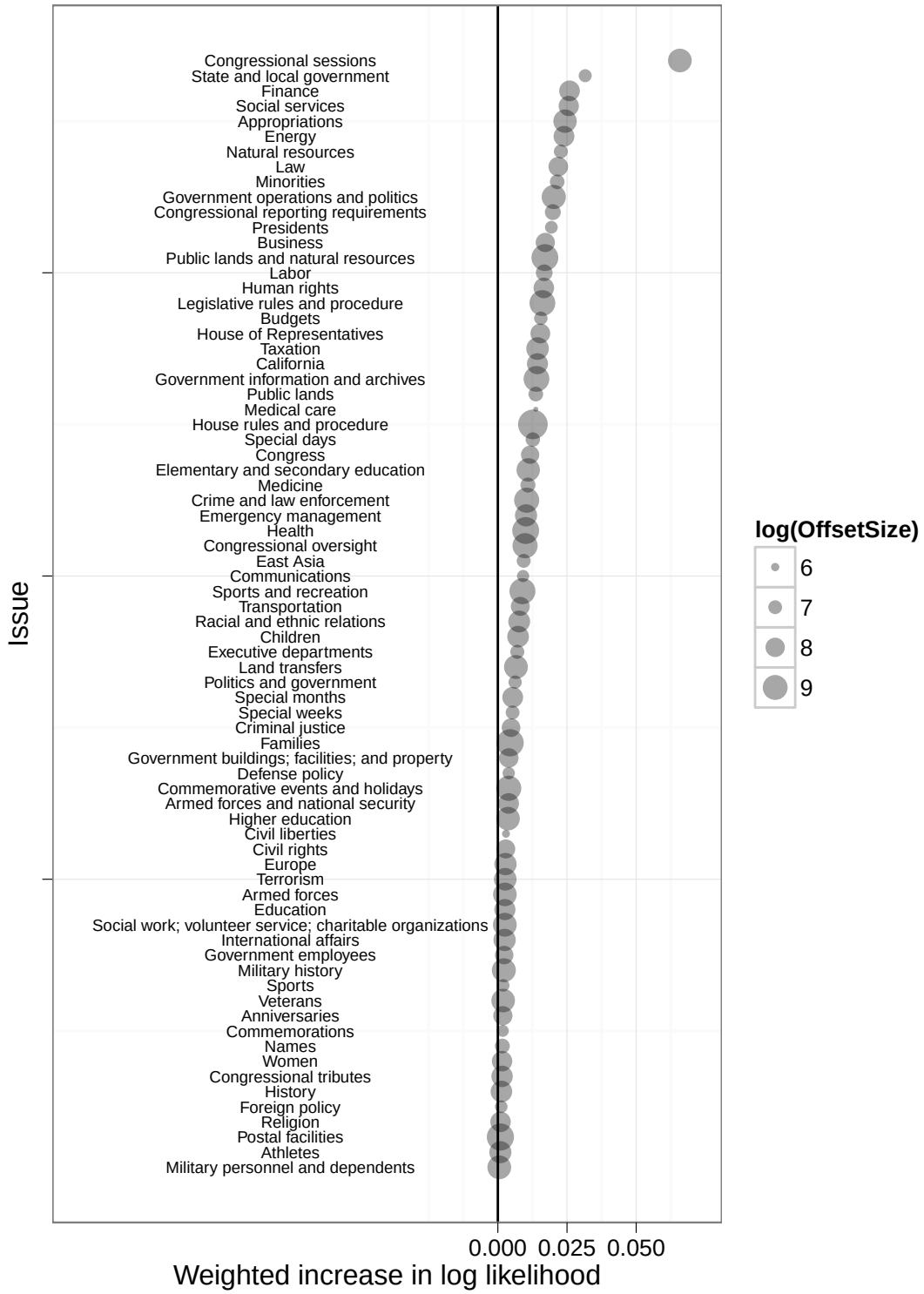

Figure 2: Issue adjustments (defined in Equation 6) for all issues.

This leads to a multimodal posterior [2]. We address this by flipping ideal points (and bill polarities) if necessary to make Republicans positive and Democrats negative.

The model is also underspecified because lawmakers' issue preferences can be explained in part by their ideal points (this is especially true on procedural issues). A typical Republican tends to have a Republican offset on taxation, but this surprises nobody. Instead, we are more interested in understanding when this Republican lawmaker *deviates* from behavior suggested by his ideal point. We therefore fit a regression for each issue k to explain away the effect of a lawmaker's ideal point $\boldsymbol{x}_u$ on her offset $\boldsymbol{z}_{uk}$:

$$\boldsymbol{z}_k = \beta_k \boldsymbol{x} + \boldsymbol{\varepsilon},$$

where $\beta_k \in \mathbb{R}$. Instead of evaluating a lawmaker's observed offsets, we use her residual $\hat{z}_{uk} = \boldsymbol{z}_{uk} - \beta_k \boldsymbol{x}_u$. By doing this, we can evaluate lawmakers in the context of other lawmakers who share the same ideal points: a positive offset $\hat{z}_{uk}$ for a Democrat means she tends to vote more liberally about issue k than Democrats with the same ideal point.[2]

Most issues had only a moderate relationship to ideal points. *House rules and procedure* was the most-correlated with ideal points, moving the adjusted ideal point $\beta_k = 0.26$ right for every unit increase in ideal point. *Public land and natural resources* and *Taxation* followed at a distance, moving an ideal point 0.04 and 0.025 respectively with each unit increase in ideal point. *Health*, on the other hand, moved lawmakers $\beta_k = 0.04$ left for every unit increase in ideal point.

Assessing significance

A handful of lawmakers stood out with the most exceptional issue adjustments. Any reference in this section to lawmakers' issue adjustments refers to lawmakers' residuals $\hat{z}_{uk}$ fit from their variational parameters $\tilde{z}_{uk}$. Lawmakers' issue adjustments are confounded because estimated issue adjustments had high variance, and issue adjustments had fatter tails than expected under a normal distribution. We therefore turned again to the same nonparametric permutation test described in the main experiments section: permute issue vectors' document labels, i.e. $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D) \mapsto (\boldsymbol{\theta}_{\pi_i(1)} \ldots \boldsymbol{\theta}_{\pi_i(D)})$, and refit lawmakers' adjustments using both the original issue vectors and permuted issue vectors. We then compare a normal issue residual $\hat{z}_{uk}$'s absolute value with issue residuals $\hat{z}_{uki}$ estimated with permuted issue vectors $\boldsymbol{\theta}_{\pi_i(d)k}$. By performing this test twenty times, we can say that a lawmaker's offset $\hat{z}_{uk}$ is significant if it is outside of the range of $\{\hat{z}_{uki}\}_i$ for all permutations i.

This provides a nonparametric method for finding issue adjustments which are more extreme than expected by chance: an extreme issue adjustment has a greater absolute value than all of its permuted counterparts.

References

[1] Harald Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. Society for Industrial and Applied Mathematics., 1992.

[2] Simon Jackman. Multidimensional analysis of roll call data via bayesian simulation: Identification, estimation, inference, and model checking. *Political Analysis*, 9(3):227–241, 2001.

[2] We also fit a model with this regression explicitly encoded. That model performed slightly worse in experiments on heldout data.