Appendix for: “Income Inequality and Electoral Theories of Polarization”

A Theoretical Evaluations of Electoral Theories of Polarization

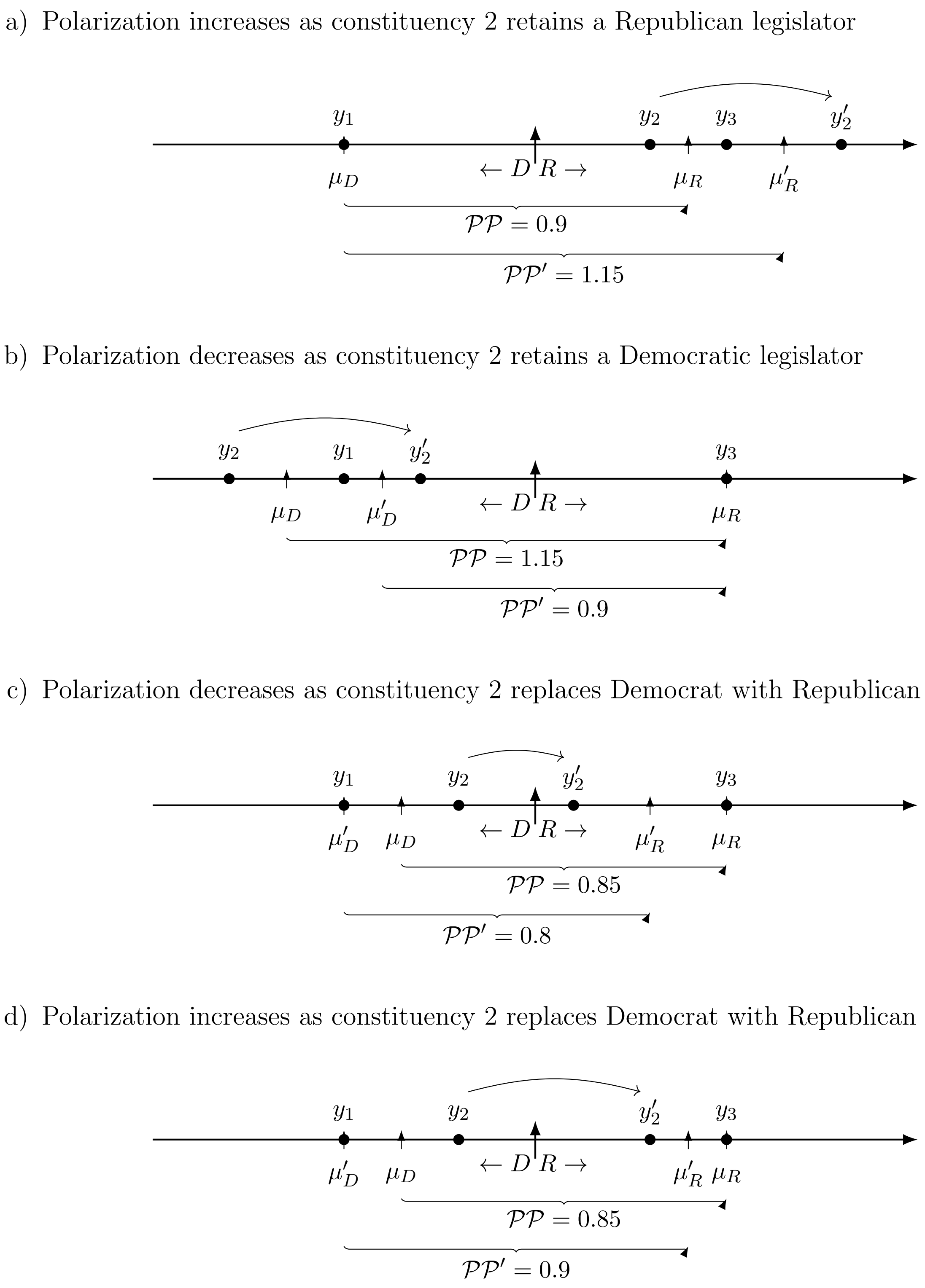
A brief exercise illustrates the difficulty of extrapolating a constituency-level comparative static into a theoretical prediction about polarization, primarily because of partisan selection. Extending the Meltzer-Richard model to speak to partisanship, it may seem innocuous enough to assume that the more conservative the median voter in a district, the more likely the district is to elect a Republican representative. Specifically, letting x_k, for constituencies $k = 1, \ldots, K$, be the ratio of a constituencies median income to the national mean, our electoral theory of polarization would be built on the following flexible assumptions: legislative voting on behalf of districts undergoing an increase in x_k is more conservative, and the partisanship of that district's representative is weakly "more Republican."[1]

Will polarization increase or decrease as a result of a change in the income distribution, i.e., of the x_ks? Consider a highly simplified example of three constituencies (districts), $k = 1, 2, 3$, where the ideology (scored voting patterns) of each constituency is denoted y_k. Figure A-1 displays four scenarios. In each case, we hold constituencies 1 and 3 fixed to consider the effect on polarization of an increase from y_2 to y_2' in constituency 2, a response to an increase in x_2. For simplicity, party affiliation depends on which side of the party dividing line ($\leftarrow D|R \rightarrow$) y_k falls. For instance, if y_2 lies to the left of this line and y_2' lies to the right, the constituency replaced a Democratic representative with a Republican one. We calculate the mean party ideologies under y_2 and y_2', μ_D, μ_R and (if different) μ_D', μ_R'. Subtracting μ_D from μ_R yields political polarization, $\mathcal{PP}$, and we calculate $\mathcal{PP}'$ (polarization under y_2' instead of y_2) analogously.

If we hold the partisanship of constituency 2 fixed (cases 1(a) and 1(b)), it seems that an increase in polarization necessitates that x_k increases in districts affiliated with the Repub-

[1] By weakly more Republican, we mean that if the district was represented by a Republican before the increase in x_k, it will be represented by one after; if the district was represented by a Democrat before the increase in x_k, it may or may not switch to being represented by a Republican.

Figure A-1: A simple comparative static has ambiguous implications for polarization



a) Polarization increases as constituency 2 retains a Republican legislator

b) Polarization decreases as constituency 2 retains a Democratic legislator

c) Polarization decreases as constituency 2 replaces Democrat with Republican

d) Polarization increases as constituency 2 replaces Democrat with Republican

Notes: In all examples, the only change in legislator ideology (voting patterns) occurs for the representative of constituency 2, who becomes weakly more conservative in voting and partisanship.

lican party in both periods and decreases in districts affiliated with the Democratic party in both periods. In scenario 1(a), a Republican legislator's voting becomes more conservative, leading μ_R to rise to μ'_R while leaving μ_D unchanged, resulting in an increase in polarization. In scenario 1(b), a Democratic legislator becomes more conservative, such that the Democratic mean is closer to the Republican mean, resulting in a decrease in polarization.

It would be wrong, however, to conclude that increases in x_2 in a Democratic district always mitigate polarization. In scenario 1(c), constituency 2 replaces its Democratic legislator with a Republican legislator. In this case, y'_2 is farther from y_3 than y_2 was from y_1. Though moderate in both cases, the inward pull that y_2 exerted on the Democratic mean was less than the inward pull y'_2 exerts on the Republican mean. And yet it would also be incorrect to suggest that any Democratic district experiencing an increase in x_k results in a reduction in polarization. In scenario 1(d), constituency 2 replaces a more moderate Democratic legislator with a less moderate Republican legislator (though still to the left of the Republican mean) such that polarization increases.

The situation only grows more complex when considering districts undergoing changes simultaneously. We might wonder, however, whether constraining the changes across constituencies in an exogenous variable to accord with some aggregate, national-level change might narrow down the range of implications for polarization. As discussed in the previous section, national-level trends are something of a red herring when it comes to electoral theories of polarization. Indeed, even if we go beyond consideration of a particular constituency-level comparative static and make additional, distributional assumptions about the exogenous changes that constituencies undergo *en masse*, we still find ourselves facing the ambiguity of competing predictions.

B Detailed Derivations of Estimands for each DAG

Ia. $X \to Y$

$$\begin{aligned}
&\mathbb{E}_Y\left[\frac{1}{\sum_j R_j}\sum_k R_k Y_k \,\middle|\, \vec{X}^C\right] \\
=\ &\mathbb{E}_Y\left[\frac{1}{\sum_j R_j} \,\middle|\, \vec{X}^C\right]\sum_k \left(\mathbb{E}_Y\left[R_k \mid X_k^C\right]\mathbb{E}_Y\left[Y_k \mid X_k^C\right]\right) && \text{by } R \perp\!\!\!\perp Y \\
=\ &\frac{1}{N_R}\sum_k R_k\,\mathbb{E}_Y\left[Y_k \mid X_k^C\right] && \text{by } R \perp\!\!\!\perp X
\end{aligned}$$

Ib. $X \to Y$**,** $X \to R$

$$\begin{aligned}
&\mathbb{E}_R\left[\mathbb{E}_Y\left[\frac{1}{\sum_j R_j}\sum_k R_k Y_k \,\middle|\, \vec{X}^C\right] \,\middle|\, \vec{X}^C\right] \\
=\ &\mathbb{E}_R\left[\mathbb{E}_Y\left[\frac{1}{\sum_j R_j}\middle|\vec{X}^C\right]\sum_k \mathbb{E}_Y[R_k Y_k \mid X_k^C] \,\middle|\, \vec{X}^C\right] && \text{by } Y \perp\!\!\!\perp R \mid X \\
=\ &\mathbb{E}_R\left[\mathbb{E}_Y\left[\frac{1}{\sum_j R_j}\middle|\vec{X}^C\right]\sum_k \mathbb{E}_Y[R_k \mid X_k^C]\,\mathbb{E}_Y[Y_k \mid X_k^C] \,\middle|\, \vec{X}^C\right] && \text{by } Y \perp\!\!\!\perp R \mid X \\
=\ &\mathbb{E}_R\left[\frac{1}{\sum_j R_j}\sum_k R_k\,\mathbb{E}_Y[Y_k \mid X_k^C] \,\middle|\, \vec{X}^C\right] && \text{by } Y \perp\!\!\!\perp R \mid X \\
=\ &\mathbb{E}_R\left[\sum_k \frac{R_k\,\mathbb{E}_Y[Y_k \mid X_k^C]}{\sum_j R_j} \,\middle|\, \vec{X}^C\right] \\
=\ &\sum_k\left(\mathbb{E}_R\left[\frac{R_k}{\sum_j R_j} \,\middle|\, \vec{X}^C\right]\mathbb{E}_R[\mathbb{E}_Y[Y_k \mid X_k^C] \mid X_k^C]\right) && \text{by } Y \perp\!\!\!\perp R \mid X \\
=\ &\sum_k\left(\mathbb{E}_R\left[\frac{R_k}{\sum_j R_j} \,\middle|\, \vec{X}^C\right]\mathbb{E}_Y[Y_k \mid X_k^C]\right) && \text{by L.I.E.}
\end{aligned}$$

IIa. $X \to Y$**,** $Y \to R$

$$\begin{aligned}
&\mathbb{E}_Y\left[\mathbb{E}_R\left[\frac{1}{\sum_j R_j}\sum_k R_k Y_k \,\middle|\, \vec{Y}^C\right] \,\middle|\, \vec{X}^C\right] \\
=\ &\mathbb{E}_Y\left[\mathbb{E}_R\left[\sum_k \frac{R_k}{\sum_j R_j} Y_k \,\middle|\, \vec{Y}^C\right] \,\middle|\, \vec{X}^C\right]
\end{aligned}$$

IIb. $X \to Y$**,** $Y \to R$**,** $X \to R$

$$\begin{aligned} &\mathbb{E}_Y\left[\mathbb{E}_R\left[\frac{1}{\sum_j R_j}\sum_k R_k Y_k \;\middle|\; \vec{X}^C, \vec{Y}^C\right] \;\middle|\; \vec{X}^C\right] \\ = \;&\mathbb{E}_Y\left[\mathbb{E}_R\left[\sum_k \frac{R_k}{\sum_k R_k} Y_k \;\middle|\; \vec{X}^C, \vec{Y}^C\right] \;\middle|\; \vec{X}^C\right] \end{aligned}$$

IIIa. $X \to Y$**,** $R \to Y$

$$\begin{aligned} &\mathbb{E}_Y\left[\frac{1}{\sum_k R_k}\sum_k R_k Y_k \;\middle|\; \vec{X}^C, \vec{R}^C\right] \\ &= \frac{1}{\sum_j R_j}\sum_k R_k \, \mathbb{E}_Y[Y_k \mid \vec{X}^C, \vec{R}^C] \\ &= \sum_k \frac{R_k \, \mathbb{E}_Y[Y_k \mid \vec{X}^C, \vec{R}^C]}{\sum_j R_j} \end{aligned}$$

IIIb. $X \to Y$**,** $R \to Y$**,** $X \to R$

$$\mathbb{E}_R\left[\mathbb{E}_Y\left[\frac{1}{\sum_k R_k}\sum_k R_k Y_k \;\middle|\; \vec{X}^C, \vec{R}^C\right] \;\middle|\; \vec{X}^C\right]$$

C Simulations

We test the validity of our estimation approach with a set of simulations that allow us to observe the true treatment effect of income on polarization, and to evaluate how well our method recovers this effect with data that is partially observed by the analyst. Before proceeding, it is worth recalling the four identification assumptions discussed in Section 4, all of which will be relevant here:

Assumption 1: *Causal identification of the underlying theory of elections*

Assumption 2: *Choosing the right causal structure*

Assumption 3: *Non-interference among constituencies*

Assumption 4: *Good out-of-sample prediction*

We begin by generating two income distributions for one hundred hypothetical constituencies. An unequal treatment income vector, $\vec{X}^T$, has ten constituencies with each integer value from 1 to 10, while a more egalitarian counterfactual vector, $\vec{X}^C$, has a narrower range from 3 to 6 (ten constituencies with 3, and thirty each from 4 to 6). Table C-1 summarizes these distributions, showing the number of constituencies in each income bin under both conditions.

	1	2	3	4	5	6	7	8	9	10	Total
$\vec{X}^T$	10	10	10	10	10	10	10	10	10	10	100
$\vec{X}^C$	0	0	10	30	30	30	0	0	0	0	100

Table C-1: Treatment and Counterfactual Income Distributions

Next, we propose a specific model consistent with each of the DAGs in Figure **??**, and use these functional forms to generate income and partisanship vectors under the treatment and counterfactual income conditions. Each constituency constitutes an independent random draw from the data-generating process (**Assumption 3**). Table C-2 summarizes our models, and shows the true expected polarization in each case.

Then, we assume the analyst can only observe one-half of the data: the income, ideology, and partisanship vectors under the treatment condition. The analyst is therefore tasked with estimating expected polarization under the counterfactual. This setup highlights our reliance on good out-of-sample prediction (**Assumption 4**), though the issue is by no means more pronounced under this than other structures.[2]

DAG	Model	$\Delta EP(\vec{X}^T, \vec{X}^C)$
I(a)	$Y \sim \mathcal{N}\left(\mu = -\frac{3}{2} + \frac{1}{2}X, \sigma = \frac{1}{2}\right)$ $R \sim \text{Binomial}(n = 100, p = \frac{1}{2})$	0.00
I(b)	$Y \sim \mathcal{N}\left(\mu = -\frac{3}{2} + \frac{1}{2}X, \sigma = \frac{1}{2}\right)$ $R \sim \text{Binomial}\left(n = 100, p = \text{logit}^{-1}(-2 + \frac{1}{2}X)\right)$	0.64
II(a)	$Y \sim \mathcal{N}\left(\mu = -\frac{3}{2} + \frac{1}{2}X, \sigma = \frac{1}{2}\right)$ $R \sim \text{Binomial}(n = 100, p = \text{logit}^{-1}(-1 + Y)$	1.30
II(b)	$Y \sim \mathcal{N}\left(\mu = -\frac{3}{2} + \frac{1}{2}X, \sigma = \frac{1}{2})\right)$ $R \sim \text{Binomial}\left(n = 100, p = \text{logit}^{-1}\left(-2 + Y + \frac{1}{2}X\right)\right)$	1.62
III(a)	$R \sim \text{Binomial}(n = 100, p = \frac{1}{2})$ $Y \sim \mathcal{N}\left(\mu = -3 + \frac{1}{2}X + R, \sigma = \frac{1}{2}\right)$	0.00
III(b)	$R \sim \text{Binomial}\left(n = 100, p = \text{logit}^{-1}\left(-2 + \frac{1}{2}X\right)\right)$ $Y \sim \mathcal{N}\left(\mu = -3 + \frac{1}{2}X + R, \sigma = \frac{1}{2}\right)$	1.39

Table C-2: Data-Generating Process and True Treatment Effect of Income on Polarization for Each DAG

Finally, we apply our method to the observed data to generate an estimate of expected polarization under the counterfactual condition. We employ a Bayesian framework to accommodate the simultaneous estimation of all model parameters, as well as the prediction and aggregation steps. **Assumptions 1 and 2** are critical here: in each case, we specify the correct causal model (and functional forms). In Table C-3, we report results from running

[2] Specifically, one could have just as easily constructed the data as a random sample of observations, with each constituency observed under only one of the treatment or control conditions, as in an experimental framework. Another alternative is to generate the data based on yet a third income vector. Regardless, as long as the data are always generated according to the same model — our crucial out-of-sample prediction assumption — this simulation will produce roughly the same results.

1,000 iterations of the same process: generating the data, estimating the model parameters, constructing predictions for ideology and partisanship under the counterfactual, and aggregating those predictions to an estimate of expected polarization. The second column reports the average of true polarization under $\vec{X}^C$ across the 1,000 iterations, the third column reports the average of our estimates, and the last column reports mean squared error (MSE) between the two. While the more complicated causal structures generate somewhat noisier predictions, on average we recover the truth in each case.

DAG	Mean, $EP(\vec{X}^C)$	Mean, $\widehat{EP}(\vec{X}^C)$	MSE
I(a)	0.00	0.00	0.01
I(b)	0.24	0.25	0.02
II(a)	0.45	0.48	0.03
II(b)	0.62	0.67	0.05
III(a)	1.00	1.00	0.02
III(b)	1.24	1.25	0.04

Table C-3: Results of Simulations
100 Legislators, 1,000 Iterations

D Exploring Potential Conditional Independence Assumptions

	(1)	(2)	(3)	(4)
Republican	0.696^{***}	0.696^{***}	0.617^{***}	0.617^{***}
	(0.007)	(0.007)	(0.006)	(0.006)
State median/national mean income				0.167^{*}
				(0.065)
Congress FE	No	Yes	Yes	Yes
State FE	No	No	Yes	Yes
R^2	0.859	0.863	0.932	0.933
Adj. R^2	0.859	0.862	0.930	0.930
Num. obs.	1800	1800	1800	1800

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table D-4: Effect of Partisanship on Ideology in the U.S. Senate, 1984-2018

	(1)	(2)	(3)	(4)
State median/national mean income	-0.831^{***}	-0.878^{***}	0.497	-0.162
	(0.107)	(0.113)	(0.260)	(0.098)
NOMINATE score				1.392^{***}
				(0.014)
Congress FE	No	Yes	Yes	Yes
State FE	No	No	Yes	Yes
R^2	0.032	0.040	0.394	0.914
Adj. R^2	0.032	0.030	0.371	0.911
Num. obs.	1800	1800	1800	1800

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

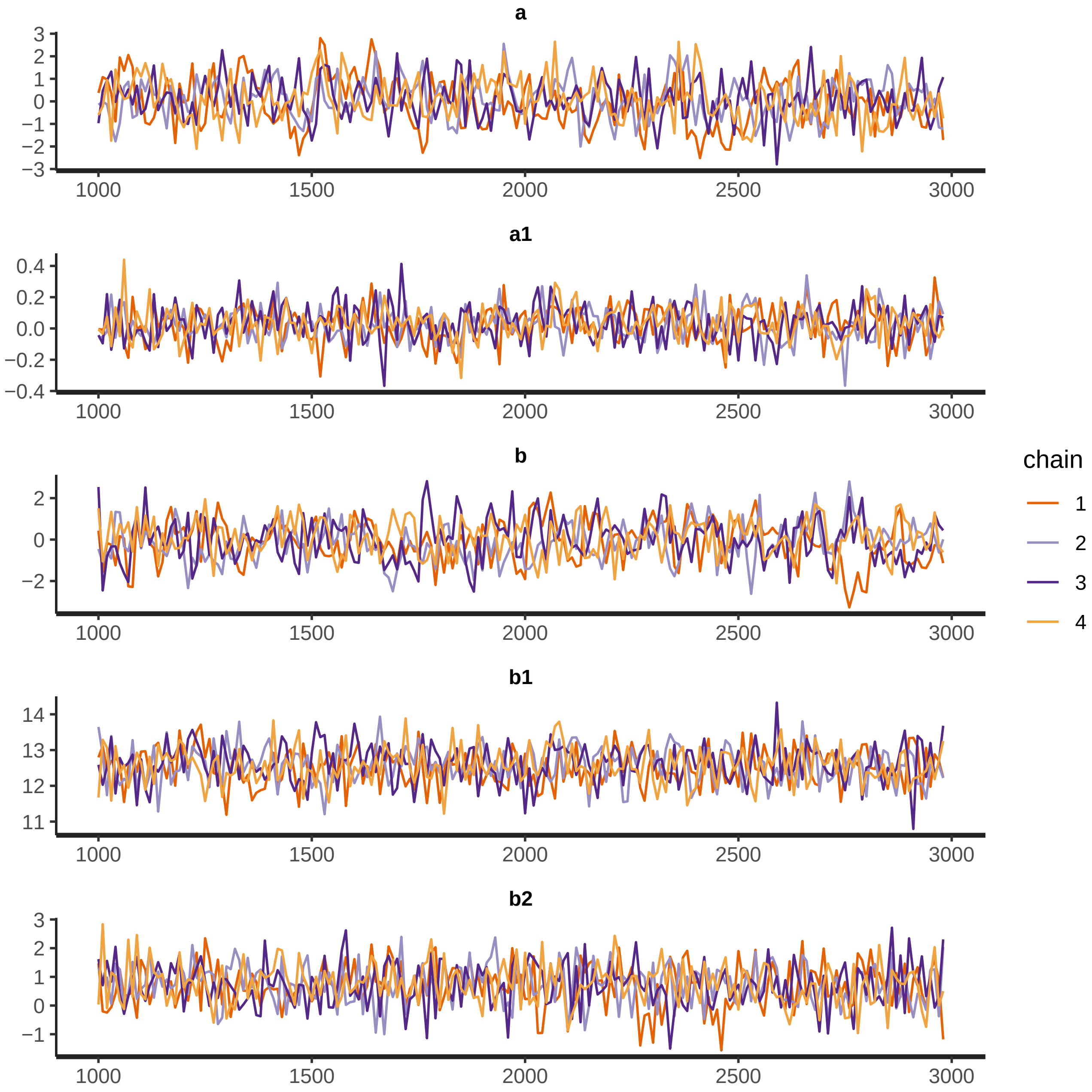
Table D-5: Effect of Income on Partisanship in the U.S. Senate, 1984-2018

E Bayesian Estimation Under DAG II(b)

We estimate the hierarchical model:

$$
\begin{aligned}
Y_{skt} &\sim \mathcal{N}(a + a_{0k} + a_1 X_{kt} + \text{Congress}_t \alpha,\ \sigma_Y) \\
R_{skt} &\sim \text{Binomial}(p = \text{logit}^{-1}(b + b_{0k} + b_1 X_{kt} + b_2 Y_{skt} + \text{Congress}_t \beta)) \\
a_0 &\sim \mathcal{N}(\mu_a, \sigma_a) \\
b_0 &\sim \mathcal{N}(\mu_b, \sigma_b) \\
\mu_a &\sim \mathcal{N}(0, 5) \\
\mu_b &\sim \mathcal{N}(0, 5)
\end{aligned}
$$

where Y is the first-dimension NOMINATE score for senator s in state k and Congress t, X is state median income over the national mean, and R is a binary indicator for being a Republican. Thus our model includes year fixed effects and state random effects, which are drawn from a weakly informative hyperprior (**?**). We use a standard normal prior for the rest of the model parameters. We estimate the model using Hamiltonian Monte Carlo, implemented in Stan (**?**). Figure E-2 shows a traceplot of the estimation for a sample of model parameters, with good mixing over the four chains.



Notes: Estimation is performed over 3,000 iterations (1,000 warm-up) with four chains.

Figure E-2: Traceplot for Estimation of DAG II(b)