Hierarchical Bayesian Aldrich–McKelvey Scaling

Jørgen Bølstad

Department of Political Science, University of Oslo, Oslo, Norway. E-mail: jorgen.bolstad@stv.uio.no

Abstract

Estimating the ideological positions of political actors is an important step toward answering a number of substantive questions in political science. Survey scales provide useful data for such estimation, but also present a challenge, as respondents tend to interpret the scales differently. The Aldrich–McKelvey model addresses this challenge, but the existing implementations of the model still have notable shortcomings. Focusing on the Bayesian version of the model (BAM), the analyses in this article demonstrate that the model is prone to overfitting and yields poor results for a considerable share of respondents. The article addresses these shortcomings by developing a hierarchical Bayesian version of the model (HBAM). The new version treats self-placements as data to be included in the likelihood function while also modifying the likelihood to allow for scale flipping. The resulting model outperforms the existing Bayesian version both on real data and in a Monte Carlo study. An R package implementing the models in Stan is provided to facilitate future use.

Keywords: ideal point estimation, Aldrich–McKelvey scaling, Bayesian estimation, hierarchical modeling

1 Introduction

Information about the ideological positions of different political actors is important for many subfields of political science. Over the last few decades, several approaches have been developed to estimate such positions from various types of data (e.g., Clinton, Jackman, and Rivers 2004; Imai, Lo, and Olmsted 2016; Poole and Rosenthal 1985, 1991). These approaches can be particularly useful when they provide information on both citizens and other political actors, locating them on the same scale (e.g., Bafumi and Herron 2010; Barberá 2015). One way to achieve this is to rely on ideological survey scales that are common in electoral studies (e.g., Aldrich and McKelvey 1977). This is useful because electoral surveys are available for many countries and time points, enabling researchers to conduct wide-ranging, cross-national analyses (e.g., Bakker *et al.* 2014; Lo, Proksch, and Gschwend 2014). Furthermore, electoral studies aim to cover nationally representative samples and they contain additional data on the respondents—in contrast to some of the other potential data sources, like Twitter data (Barberá 2015).

A key challenge with ideological and policy-related survey scales is that respondents tend to interpret them differently—a problem sometimes referred to as differential item functioning. Some respondents tend to place actors that are on their own side of the scale closer to the center while moving others away (Hare *et al.* 2015). Respondents also differ in the degree to which they use the full scale: Some stretch the space, spreading out their responses over wider range, whereas others do the opposite. As a consequence, the respondents' placements are not directly comparable. Aldrich and McKelvey (1977) developed a method to correct for such differences and estimate the positions of both respondents and other political actors on a common scale. Their method—which will be referred to as the AM model—has been developed further by Poole (1998), and later been given a Bayesian implementation by Hare *et al.* (2015).

The Bayesian version of the model has proved useful in a number of studies, but it also has some important shortcomings. One is that the model yields improper marginal posterior distributions for a notable share of the latent respondent positions. Another issue is that the model is prone to overfitting because it is an "unpooled" model, where the individual-level parameters are modeled separately. This is problematic because the model is used in settings with as few as

Political Analysis (2024)
vol. 32: 50–64
DOI: 10.1017/pan.2023.18

Published
8 June 2023

Corresponding author
Jørgen Bølstad

Edited by
Jeff Gill

four observations per individual, even though it entails at least two individual-level parameters. Furthermore, these parameters are among the model's key outputs as they are used to rescale the respondent's self-placements: When these parameter estimates are sensitive to noise, the estimated respondent positions will be as well.

This article addresses these issues by developing a hierarchical version of the Bayesian AM model. The new model treats each set of individual-level parameters as a sample from a common population distribution, which makes the model less prone to overfitting. In addition, respondents' self-placements are treated as data containing error, and their true positions are treated as latent parameters to be estimated. Like the other individual-level parameters, these latent positions are given hierarchical priors. Furthermore, the model entails a new likelihood function, which allows for scale flipping in combination with the new priors. Together, these changes result in a model that yields more accurate estimates and produces proper marginal posterior distributions for all respondent positions. The new model is compared with the existing Bayesian implementation using both real and simulated data, and the new model outperforms the older version on all performance measures. To facilitate future use, an R package implementing the models in Stan is provided along with this article.[1]

2 The Bayesian AM Model

Aldrich and McKelvey (1977) noted that when survey respondents are asked to place political actors (or "stimuli") on an ideological or policy-related scale, they are likely to interpret the scale in different ways. Respondents may *shift* the political space by moving all positions in one or the other direction, and they may *stretch* (or contract) the space by moving all positions outward (or inward). Aldrich and McKelvey therefore developed a least squares procedure yielding point estimates of stimulus positions as well as individual shift and stretch parameters. They further used these parameters to transform respondents' self-reported positions and place them on the same scale as the stimuli. This approach has been implemented in the R package *basicspace* (Poole *et al.* 2016) and has been used in a number of studies since it was first introduced (e.g., Hollibaugh, Rothenberg, and Rulison 2013; Lo *et al.* 2014; Palfrey and Poole 1987; Saiegh 2009).[2]

Hare *et al.* (2015) developed a Bayesian version of the AM model, and their version has also been used in an increasing number of studies (e.g., Alemán *et al.* 2018; Bakker, Jolly, and Polk 2020; Carroll and Kubo 2018; Clay *et al.* 2020; Saiegh 2015; Zakharova and Warwick 2014). The model is intended to capture how respondent $i \in \{1, \ldots, N\}$ reports the position of political actor $j \in \{1, \ldots, J\}$ as a function of this actor's true latent position, θ_j. If we denote a reported stimuli position Y_{ij} and let $\phi(\cdot|\cdot, \cdot)$ be the probability density function of the normal distribution, then the likelihood function introduced by Hare *et al.* (2015) can be written as

$$\prod_{i=1}^{N} \prod_{j=1}^{J} \phi(Y_{ij}|\alpha_i + \beta_i \theta_j, \sigma_{ij}^2), \tag{1}$$

where α_i is a shift parameter, β_i is a stretch parameter, and σ_{ij}^2 is a variance term allowing for heteroskedasticity.[3]

1 The R package, called *hbamr*, is available at https://cran.r-project.org/package=hbamr.

2 Poole (1998) developed a generalized version of the AM model, which allows for missing values and has been used by, for example, Bakker *et al.* (2014).

3 Bølstad (2020) proposed an extended version of this model aiming to capture rationalization. As the model proposed by Bølstad (2020) retains several key features of the original model, the adjustments made in the present article also improve the former. For those who would like to capture rationalization, a revised version of the model in Bølstad (2020) is included in the R package accompanying this article.

2.1 The Unpooled Model (BAM)

A key question is how to specify priors for the individual-level parameters. One option is to let the parameters be estimated separately, as if they were completely unrelated. This is often referred to as an "unpooled" specification—in contrast to a "pooled" one, which would restrict the parameters to be equal for all respondents (see, e.g., Gelman *et al.* 2014). Both Aldrich and McKelvey (1977) and Hare *et al.* (2015) used unpooled specifications, and the latter did so by placing wide, uniform priors on the shift and stretch parameters: $\alpha_i, \beta_i \sim \text{Unif}(-100, 100)$. Such priors serve to emulate a maximum-likelihood-based approach, but they also leave the model sensitive to noise, and one of the contributions of this article is therefore to improve this part of the model.

Another key question is how to estimate each respondent's position, χ_i, on the same scale as the stimuli, while correcting their self-placement, V_i, for shifting and stretching. For each posterior draw, Hare *et al.* (2015) transform the self-placements as follows:

$$\chi_i = \frac{V_i - \alpha_i}{\beta_i}. \tag{2}$$

This transformation is logically consistent with the likelihood function in Equation (1) and similar to the approach of Aldrich and McKelvey (1977). However, the division by β_i is problematic because it may lead to division by values that are arbitrarily close to zero, resulting in draws approaching positive and negative infinity. To address this issue, the authors use the posterior median rather than the mean to obtain point estimates of the respondents' positions. This prevents the most extreme draws from shifting the point estimates around, but it does not resolve the underlying problem. A key contribution of this article is therefore to offer a more satisfactory solution.

A final question is how to identify the parameters. Models of these kinds are typically unidentified, as the latent space can be shifted and stretched while yielding the same likelihood values (see, e.g., Bafumi *et al.* 2005). To address this issue, Hare *et al.* (2015) fixed two stimulus positions on the latent scale while placing standard normal priors on the others: $\theta_j \sim \text{Normal}(0, 1)$. For the sake of comparison, I retain all these specification choices while translating the model from JAGS (Plummer 2003) to Stan (Carpenter *et al.* 2017).[4] I refer to the resulting unpooled Bayesian AM model as the BAM model.[5]

2.2 A Limited Hierarchical Model (HBAM_0)

An alternative to the unpooled or completely pooled specifications is a "partially pooled" one, which typically involves a hierarchical model. Such a specification follows naturally if we view each set of individual-level parameters as a sample from a common population distribution. This leads to a hierarchical prior structure where the individual-level parameters are given common prior distributions that themselves have parameters to be estimated. A key point is that a hierarchical model uses information across individuals rather than treating them as unrelated, which typically allows the individual-level parameters to be estimated more accurately given the finite data at hand (see, e.g., Gelman *et al.* 2014).

In the present case, replacing the uniform priors on α_i and β_i also offers an alternative way to identify the model. We can select one out of the many possible transformations of the latent space by setting specific means or medians for the prior distributions.[6] For the α parameters,

4 Stan uses an automatically tuned form of Hamiltonian Monte Carlo, which is more efficient than simpler algorithms when it comes to sampling from complicated distributions exhibiting posterior correlations (Bølstad 2019; Hoffman and Gelman 2014).

5 There is one aspect of the model I adjust compared with the JAGS version, and that is the specification of the variance term, σ^2_{ij}. I use the same specification of σ^2_{ij} for all models in this article, and this specification is discussed in a later section.

6 An advantage of this approach is that it yields meaningful posterior distributions for all stimuli—in contrast to the approach of keeping two stimulus positions fixed, which yields awkward distributions for the fixed stimuli.

setting the mean to zero is a natural choice (implying no shifting on average): $\alpha_i \sim \text{Normal}(0, \sigma_\alpha^2)$. In this specification, the standard deviation, σ_α, is an aspect of the population distribution to be estimated jointly with the other parameters. For the β parameters, using a distribution with a median of one yields a latent scale with units similar to the observed one, and we can achieve this by using a log-normal distribution: $\beta_i \sim \text{Log-normal}(0, \sigma_\beta^2)$.[7]

The use of a log-normal distribution also has a second motivation: It keeps the β parameters away from zero.[8] As noted above, Hare *et al.* (2015) scaled respondents' self-placements as shown in Equation (2) for each posterior draw, and this may lead to division by β values that are arbitrarily close to zero—which in turn yields extreme and implausible draws for the latent respondent positions. This problem implies that not all relevant information has been incorporated in the model, and using a prior that keeps the β parameters away from zero is one way of including more such information.

Another key point is that the reported self-placements represent additional data that inevitably contain errors. Instead of rescaling the self-placements while assuming they are measured without error, we can model these data probabilistically, making them part of the likelihood function. If we model the self-placements the same way as the stimuli placements, we get the following likelihood for these data:

$$\prod_{i=1}^{N} \phi(V_i | \alpha_i + \beta_i \chi_i, \sigma_i^2), \tag{3}$$

where χ_i is a parameter representing the latent position of respondent i and σ_i^2 is a variance term.[9] The complete likelihood for the model is then the product of Equations (1) and (3).

Finally, we do have some information about what values are plausible for the latent positions: As they come from the same population, there is a limit to how extreme we would expect any respondent to be, relative to the rest. By treating the latent respondent positions as parameters to be estimated, we can model these hierarchically and let the degree of shrinkage be determined by the data: $\chi_i \sim \text{Normal}(0, \sigma_\chi^2)$.[10] This results in a more nuanced posterior distribution where areas of the parameter space that yield extreme and implausible values for χ_i are weighted down compared to the BAM specification. Together, these adjustments should reduce the model's sensitivity to noise and ensure meaningful results for all respondents.

2.3 A Hierarchical Mixture Model (HBAM)

The model outlined above restricts the stretch parameters to be positive, but as Aldrich and McKelvey (1977) noted, some respondents may flip the scale and thus have negative parameters. These respondents would see the ideological space as a mirror image, and report conservative actors as liberal, and vice versa. This possibility is one of the reasons why Aldrich and McKelvey did not attempt to restrict the stretch parameters to be positive, and later applications of their method suggest there is a notable minority of such respondents in a typical survey sample (e.g., Lo *et al.* 2014). I therefore expand the HBAM_0 model to allow for such behavior.

I consider each respondent's potential scale flipping as a latent discrete parameter and modify the HBAM_0 model accordingly. I define the likelihood for each respondent as a mixture of two

7 The hyperparameters are given weakly informative priors, allowing the data to drive the results: $\sigma_\beta \sim \text{Gamma}(3, 10)$, and $\sigma_\alpha \sim \text{Gamma}(2, 5/B)$. The latter implies a mode at $B/5$, where B denotes the upper bound of a centered, symmetrical survey scale and serves to adjust the priors to different scales. For a 7-point scale, B is 3; for an 11-point scale, it is 5.

8 It also restricts the parameters to be positive, and the likelihood function is therefore expanded in the next section to allow for negative β's.

9 χ_i is identified by assuming that each respondent places themselves on the scale with the same accuracy as they place the stimulus with the smallest errors.

10 The standard deviation, σ_χ, is given a hyperprior of $\text{Gamma}(5, 8/B)$, which has a mode at half B while allowing all plausible values.

distributions, representing the flipped and non-flipped states. I further marginalize out the latent discrete flipping parameter, obtaining the following likelihood for Y_{ij}, where I estimate the flipping parameter's expectation, λ_i:

$$\prod_{i=1}^{N}\prod_{j=1}^{J}[\lambda_i\phi(Y_{ij}|\alpha^*_{1i}+\beta^*_{1i}\theta_j,\sigma^2_{ij}) + (1-\lambda_i)\phi(Y_{ij}|\alpha^*_{2i}+\beta^*_{2i}\theta_j,\sigma^2_{ij})]. \quad (4)$$

In this specification, each respondent is given separate α and β parameters for each state, denoted α^* and β^*. The first set of these β^* parameters, β^*_{1i}, takes positive values, whereas the second set, β^*_{2i}, takes negative values. The α^* and β^* parameters are given the same priors and hyperpriors as the α and β parameters in the HBAM_0 model, with the exception that the prior on β^*_{2i} has a negative sign. The mixing proportion, λ_i, is given a beta prior, specified in terms of its expectation, ψ, and concentration, δ: $\lambda_i \sim \text{Beta}(\psi\delta,(1-\psi)\delta)$.[11] The expectation parameter is given the following prior: $\psi \sim \text{Beta}(8.5,1.5)$, which implies a 15% prior probability that a respondent flips the scale.[12] This prior permits all plausible values for ψ while emphasizing those that are more probable in light of existing studies.[13]

Still treating the self-placements as data to be modeled probabilistically and the respondent positions as latent parameters, I also expand the likelihood for V_i in Equation (3) to allow for flipping:

$$\prod_{i=1}^{N}[\lambda_i\phi(V_i|\alpha^*_{1i}+\beta^*_{1i}\chi^*_{1i},\sigma^2_i) + (1-\lambda_i)\phi(V_i|\alpha^*_{2i}+\beta^*_{2i}\chi^*_{2i},\sigma^2_i)]. \quad (5)$$

The latent positions are still modeled as coming from the same population distribution: $\chi^*_{1i},\chi^*_{2i} \sim \text{Normal}(0,\sigma^2_{\chi^*})$, and the prior for $\sigma^2_{\chi^*}$ is the same as for σ^2_χ.

To generate posterior draws for the latent discrete flipping parameters, κ_i, I use their respective expectations, λ_i: $\kappa_i \sim \text{Bernoulli}(\lambda_i)$. I then use these draws to combine χ^*_{1i} and χ^*_{2i} and obtain draws for each respondent's position, χ_i:

$$\chi_i = \kappa_i\chi^*_{1i} + (1-\kappa_i)\chi^*_{2i}. \quad (6)$$

Draws for the α and β parameters are obtained the same way:

$$\alpha_i = \kappa_i\alpha^*_{1i} + (1-\kappa_i)\alpha^*_{2i}, \quad (7)$$

$$\beta_i = \kappa_i\beta^*_{1i} + (1-\kappa_i)\beta^*_{2i}. \quad (8)$$

To see the implications of the HBAM model and some of its key priors, it is helpful to simulate draws from the prior for β_i, integrating over the relevant hyperpriors. Figure 1 summarizes the draws from such a simulation. The prior distribution for β_i is a bimodal mixture, reflecting the two possibilities that respondents either do or do not flip the scale. The medians of the two mixture

11 The concentration parameter is given a prior with a lower limit at 2: $\delta \sim \text{Gamma}(2,.1)+2$. This limit prevents the beta distribution from turning bimodal, which would cause trouble for the sampling algorithm.

12 ψ is also given a lower limit of .5 to ensure that the model is identified. The prior probability of scale flipping is still 15% after applying this limit.

13 As the prior will be dominated by the data, the exact specification will not make a notable difference, but the curvature at the upper limit may be beneficial for the sampling algorithm.

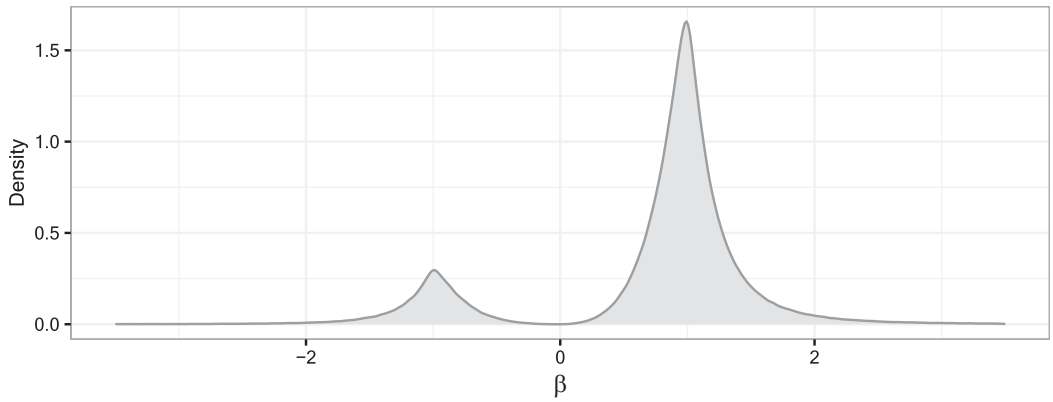

Figure 1. Prior simulation for β_i in the HBAM model, integrating over all relevant hyperpriors.

components are set to 1 and −1 to identify the latent space, whereas the concentration of each component will be determined by the data, due to the hierarchical prior structure. The positive part of the distribution contains more probability mass than the negative part, which reflects the expectation that scale flipping is less likely than non-flipping. Again, the relative weights of the negative and positive components of the mixture will predominantly be determined by the data, given the hierarchical setup. A more comprehensive discussion of the priors is provided in the Supplementary Material (including prior simulations for all parameters).

2.4 Specification of Errors

Implementing the models in Stan makes it natural to model the standard deviation of the errors, σ_{ij}, more directly than one would in JAGS—where the use of non-conjugate priors leads to inefficient Metropolis sampling. Specifically, σ_{ij} is constructed as a function of two parameters: $\sigma_{ij} = \rho_j\sqrt{\eta_i}$, where η_i captures the average variance of respondent i (implicitly increased by a factor of J^2 for more efficient sampling), whereas ρ is a unit J-simplex vector, splitting the variance by stimuli. The simplex vector is given a weakly informative, symmetric Dirichlet prior: $\rho \sim \text{Dirichlet}(5)$, whereas η_i is given a scaled inverse chi-squared prior, $\eta_i \sim \text{Scale-inv-}\chi^2(\nu, J^2\tau^2)$.[14] To ensure that the models are comparable and sampling efficiently, this specification is used for all models—including the BAM model.[15]

3 Empirical Application

To compare and evaluate the models introduced above, I start with an empirical application, before conducting a Monte Carlo study.[16] The empirical application provides a test on real data, showing whether the results from the models differ, whether the posterior distributions for the respondent positions are meaningful, and whether the models differ in their out-out-sample predictive accuracy. The empirical test will also provide input for the simulation study, making the simulated data as realistic as possible. The simulation study will shed light on the models' ability to recover true latent positions and will also provide information on how the models perform in terms of other assessment criteria.

14 The hyperparameter ν, which represents the degree of similarity in error scales across respondents, is given a moderately informative prior to make the models robust to difficult empirical scenarios: $\nu \sim \text{Gamma}(25, 2.5)$. The hyperparameter τ, which represents the general scale of the errors across respondents, is given a weakly informative prior: $\tau \sim \text{Gamma}(2, 5/B)$, which implies a mode at $B/5$ while allowing all plausible values.

15 The *hbamr* package also provides a few simpler homoskedastic HBAM variants that are not discussed in this article.

16 Complete replication materials for all analyses in this article are available at Harvard Dataverse (Bølstad 2023).

Table 1. Estimated ELPDs based on a 10-fold cross-validation. ELPD is the theoretical expected log pointwise predictive density for a new dataset, whereas $SE_{\widehat{ELPD}}$ is the standard error of the ELPD estimate.

	BAM	$HBAM_0$	HBAM
$\widehat{ELPD}$	−28,049.2	−30,350.9	−27,405.3
$SE_{\widehat{ELPD}}$	146.5	120.7	107.6

For the empirical application, I use a dataset that was analyzed by Hare *et al.* (2015) when they introduced the BAM model. The data come from the 2012 American National Election Study (ANES) and consist of 7-point Liberal-Conservative scales.[17] The respondents were asked to place themselves, as well as the following four stimuli: the Democratic Party, Democratic presidential candidate Barack Obama, the Republican Party, and Republican candidate Mitt Romney.[18]

3.1 *K*-fold Cross-Validation

I start by estimating the pointwise out-of-sample prediction accuracy of the models, which will help assess model fit and overfitting. Unpooled models are prone to overfitting, and we would expect the HBAM model to outperform the BAM model. A comparison between the $HBAM_0$ and HBAM models will also provide information on whether allowing for scale flipping is worthwhile—or an unnecessary complication.

To compare the models, I perform a 10-fold cross-validation on the reported stimuli positions.[19] In other words, I randomly partition the data into 10 equally sized subsets (using stratified sampling so that each respondent at most provides one observation to each subset). I then fit the models 10 times, each time holding out one of the subsets, while evaluating the pointwise log likelihood of the held-out data given the estimated parameter values. For each model, I then estimate the expected log pointwise predictive density (ELPD) and its standard error.

The results of the cross-validation are shown in Table 1. We see that the estimated ELPD is significantly lower for the $HBAM_0$ model than for the BAM model, while the HBAM model has a significantly higher estimate than either of the other models. In short, the results suggest that the HBAM model performs better than the other models in terms of predicting new data. The lower performance of the unpooled model is expected and implies that this model is overfitting the data. The difference between $HBAM_0$ and the other models implies that the $HBAM_0$ model is too restrictive and underfits the data: Allowing for scale flipping yields a considerable improvement in predictive accuracy.

3.2 Estimated Shifting

One of the key findings of Hare *et al.* (2015) was that respondents tend to shift the ideological space in a way that may lead us to underestimate polarization: They move the stimuli that they disagree with too far away on the other side of the scale while understating their own ideological extremism. This pattern is reproduced in Figure 2, which shows how the estimated α parameters from the BAM and HBAM models are distributed over respondents' self-placements. Those on the left tend to

17 Hare *et al.* (2015) also analyzed ANES data from 2004 and 2008, as well as the 2010 Cooperative Congressional Election Study. I am not analyzing these additional datasets in this article. However, the vignette for the *hbamr* package includes analyses of the ANES 1980 data that serve to illustrate the original AM model in the R package *basicspace* (Poole *et al.* 2016). The Supplementary Material for this article also includes several analyses of this dataset.

18 Because the data contain only four stimuli and we want to perform cross-validation, I include only those respondents who have complete data. I further require that respondents have used at least two unique positions to place the stimuli. This leaves 4,949 respondents.

19 In other settings, approximate leave-one-out cross-validation using Pareto smoothed importance sampling could be an efficient alternative (Vehtari, Gelman, and Gabry 2017). However, a large share of high Pareto *k*-values for the BAM model rules out this option.

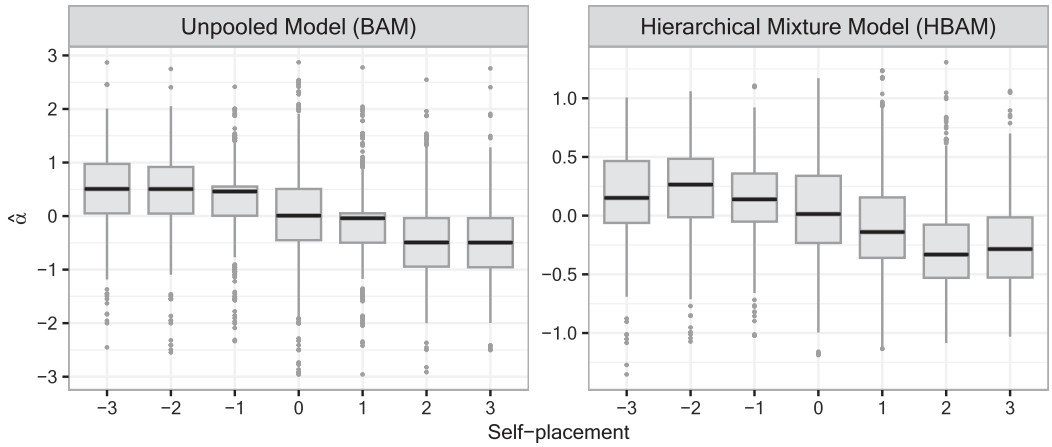

Figure 2. Estimated shifting over self-placements. The plots summarize posterior median α_i estimates.

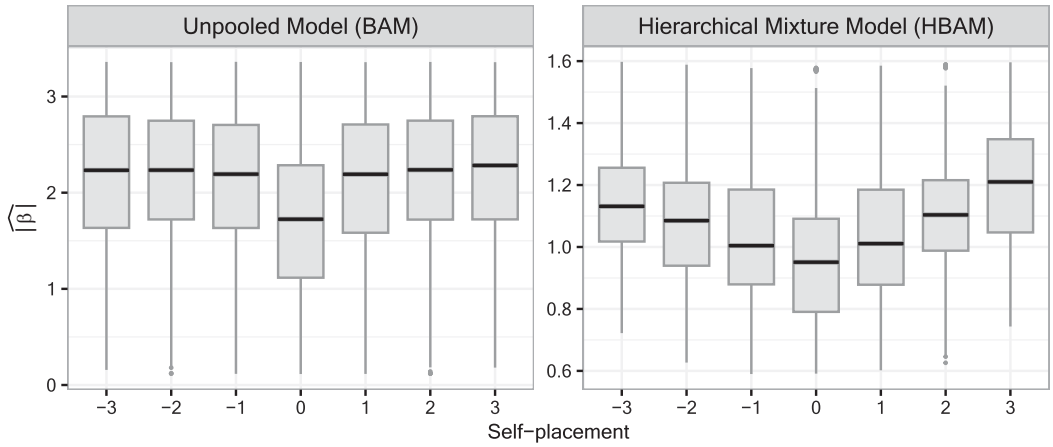

Figure 3. Estimated stretching over self-placements. The plots summarize medians of absolute values of posterior draws for β_i.

shift their placements to the right, whereas those on the right tend to shift their placements to the left. The two models produce a similar picture, but the estimates from the BAM model are larger and contain more outliers.[20]

3.3 Estimated Stretching

The fundamental logic behind these models implies that we would expect to find a particular pattern in the estimated β parameters: A key assumption is that some respondents stretch the space more than others, and thus place both the stimuli and themselves at more extreme positions. In short, we would expect to find that those who place themselves at more extreme positions tend to stretch the space to a greater extent. This is why the models use the stimuli placements to estimate the degree of stretching and rescale the respondents' self-placements accordingly. If the mentioned assumption does not hold, a key part of these models is invalidated.

As the β parameters can take on both positive and negative values in both the BAM and HBAM models, the most straightforward way to gauge the overall degree of stretching regardless of flipping is to use the median of the absolute values of the posterior draws for each respondent. The boxplots in Figure 3 show these median absolute β estimates from each model over respondents'

20 To avoid shrinking these estimates toward a common mean—like the HBAM version presented here does—one could give different priors to α_i for each self-placement, and give these priors their own mean hyperparameter to be estimated (while fixing the mean for the center category to zero to identify the latent space). The *hbamr* package provides such a model (referred to as HBAM_2).

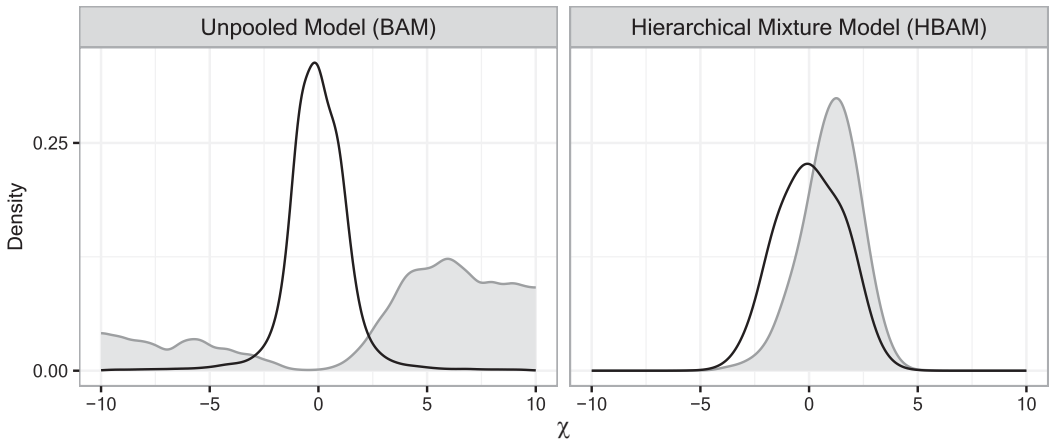

Figure 4. Posterior distribution for the position of a specific respondent compared to the population. The gray areas summarize the draws for one of the respondents for which the BAM model produces a problematic distribution. The black lines show the density of the draws for all respondents. The horizontal axes have been capped at ±10, although the draws produced by the unpooled model range from below −10,000 to above 10,000.

self-placements. Interestingly, the β parameter estimates produced by the BAM model follow a virtually flat line across all but the center category. In contrast, the estimates from the HBAM model follow a V-shape, where the β estimates are more extreme for respondents who place themselves at the extremes of the scale. This is important because it is exactly the kind of pattern we would expect to find if the assumption behind these models is correct. It also illustrates that the models can yield substantively different conclusions, even at an aggregate level.

3.4 Estimated Respondent Positions

A key problem with the current implementation of the BAM model is that it does not yield meaningful marginal posterior distributions for all respondent positions. To rescale a respondent position, the BAM model subtracts the draws for α_i and divides by the draws for β_i, which can result in extreme values if the draws for β_i get close to zero. This problem is illustrated in Figure 4, which focuses on one of the respondents for which the BAM model produces a problematic distribution. Along with the draws for the respondent's position, the plot shows the estimated population distribution (as the density of the draws for all respondents).

From a Bayesian perspective, the population distribution contains relevant information about the positions of individual respondents: If we had no other information, the population distribution would define a broad range within which we would expect a randomly chosen respondent to be located. In Figure 4, most of the BAM model's draws for the selected respondent are in areas that are highly implausible according to the population distribution. At the most extreme, the draws go beyond ±10,000. In contrast, the HBAM_0 and HBAM models avoid this problem by treating the latent respondent positions as parameters and incorporating information about the population distribution in their priors.

To assess the extent of this issue, I compare all individual respondents' marginal posterior distributions to the estimated population distribution. For each model, I calculate the range within which 95% of the draws for all respondent positions lie. This is an approximation of where we would expect most respondents to be located, and if individual respondents have 95% credible intervals (CIs) notably wider than this range, it suggests that their posterior distributions are too wide. I therefore calculate the 95% CI for each respondent's position and compare these to the population distribution. For the BAM model, the share of CIs that are more than 50% wider than the population 95% range is 8%. For the HBAM_0 and HBAM models, the number of such cases is zero.

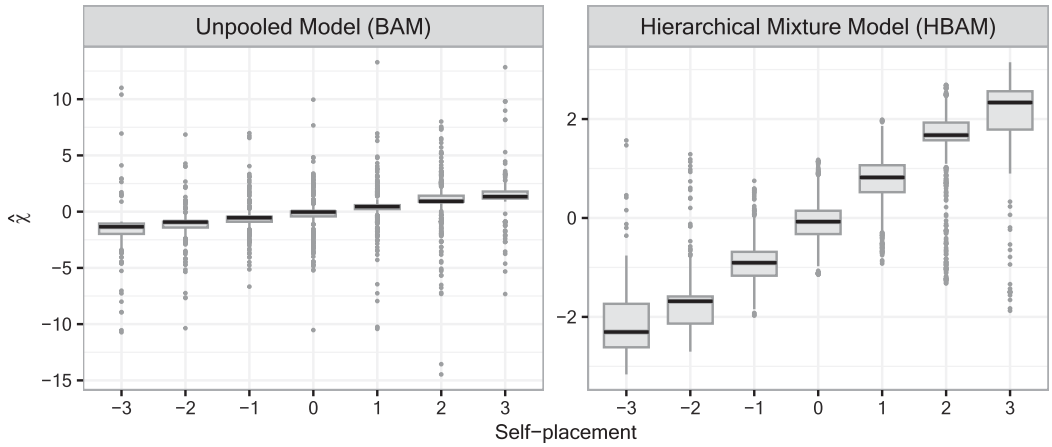

Figure 5. Respondent position estimates over self-placements. The estimates are posterior medians.

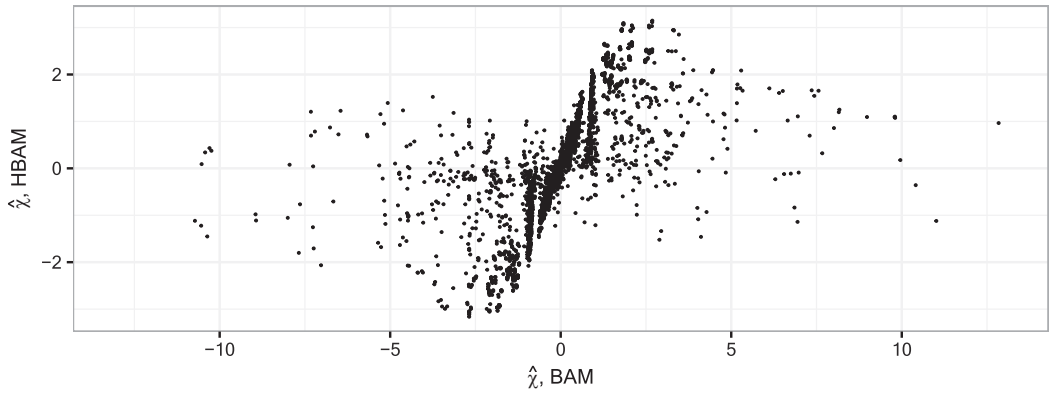

Figure 6. Respondent position estimates from the HBAM model over estimates from the BAM model. The estimates are posterior medians.

In addition to the problem shown in Figure 4, we would expect the overfitting demonstrated by the cross-validation to reduce the accuracy of the BAM model's point estimates of the respondent positions. Figure 5 shows the estimated posterior median respondent positions over respondents' self-placements. The most notable about the results from the BAM model is the amount of extreme respondent position estimates (compared, e.g., to the interquartile range represented by the boxes). Even respondents who place themselves at zero are occasionally moved to extreme positions. In contrast, the HBAM results show a pattern where respondents are rarely moved to much more extreme positions than they place themselves at, although they may be flipped to the other side.

Figure 6 shows the posterior median respondent position estimates from the HBAM model over the estimates from the BAM model. Again, we see that the BAM model produces some very extreme estimates, even for respondents that are considered centrist by the HBAM model. The correlation between the two sets of estimates is .6, which implies a substantive difference in the results from the two models. To see whether this also reflects a difference in the models' abilities to uncover the true latent parameters, I conduct a Monte Carlo study in the next section.

3.5 Estimated Hyperparameters

Table 2 summarizes the marginal posterior distributions of the hyperparameters in HBAM model. Notably, ψ is estimated to be between .88 and .90 with 95% probability. This implies a 10%–12% probability that an individual respondent flips the scale, which is plausible in light of previous studies (e.g., Lo *et al.* 2014). The other estimates are less substantively interesting, but relevant for the Monte Carlo study below.

Table 2. Summary of marginal posterior distributions for the hyperparameters in the HBAM model.

	2.5%	50%	97.5%	N_{eff}	Rhat
ψ	0.88	0.89	0.89	752	1.00
δ	2.00	2.01	2.02	3,836	1.00
σ_α	0.55	0.56	0.58	1,281	1.00
σ_β	0.26	0.27	0.28	1,317	1.00
σ_χ	1.51	1.55	1.59	1,657	1.00
τ	0.55	0.57	0.58	319	1.01
ν	7.15	8.40	10.04	297	1.01

4 Monte Carlo Study

While the empirical application shows that the models yield notably different results when applied to a real dataset, we also want to know how well the models perform more generally. In particular, we would like to know whether they succeed in recovering true latent positions. To test the models in a controlled setting, I conduct a Monte Carlo study covering typical parameter values.

4.1 Setup

Because we would expect model performance to depend on the amount of data available at the individual level, I run simulations for different numbers of stimuli. In Lo *et al.*'s (2014) study of 22 countries, the number of stimuli ranged from 5 to 11, with a mode of 8, and I run simulations for $J \in \{4, 8, 12\}$. Lo *et al.* also found nontrivial shares of negative stretch parameters—typically around 5%–10%. To see how the models perform with and without a moderate degree of scale flipping, I run simulations for $\psi \in \{.85, 1\}$. Finally, to see how robust the models are to random noise, I incrementally increase the scale of the errors, τ, across a wide range covering all commonly observed values.[21]

To make the simulated data as realistic as possible, the remaining parameter values for the data generating process are taken from the empirical application reported above (see Table 2). For each combination of J and ψ values, I simulate 1,000 datasets with $N = 500$. I then fit the three models to each set and calculate relevant performance criteria.[22] Because the models yield differently scaled results, measures like the root-mean-square error are not directly applicable without some normalization, and I focus instead on the correlation (Pearson's r) between the sets of estimated and true respondent positions.

4.2 Results

Figure 7 reports results for the posterior median respondent positions. The figure shows loess smoothed curves, summarizing the correlations over the different error scales for each combination of J and ψ values. The unpooled BAM model proves sensitive to noise, particularly when there are few observations at the individual level ($J = 4$). When there is no flipping, the $HBAM_0$ and HBAM models perform very well, and nearly identically. However, in scenarios where flipping is present ($\psi = .85$), the $HBAM_0$ model performs notably worse. Overall, the HBAM model outperforms (or performs as well as) the other models in every scenario.

21 Specifically, I increase τ from .5 to 1.4 for a survey scale ranging from −3 to +3. (In the analysis reported above, the posterior median for τ is .57, whereas the equivalent estimate from an analysis of ANES 1980 is .74.) When the ratio of τ to the length of the scale is held constant, the results for other scale lengths are essentially identical to those reported here.

22 The number of respondents is not crucial here, as this article mainly aims to improve the individual-level parameter estimates, which are largely unaffected by the number of respondents. An N of 500 is within the range of sample sizes commonly seen in this literature. In Lo *et al.*'s (2014)'s study, N ranged from 284 to 875.

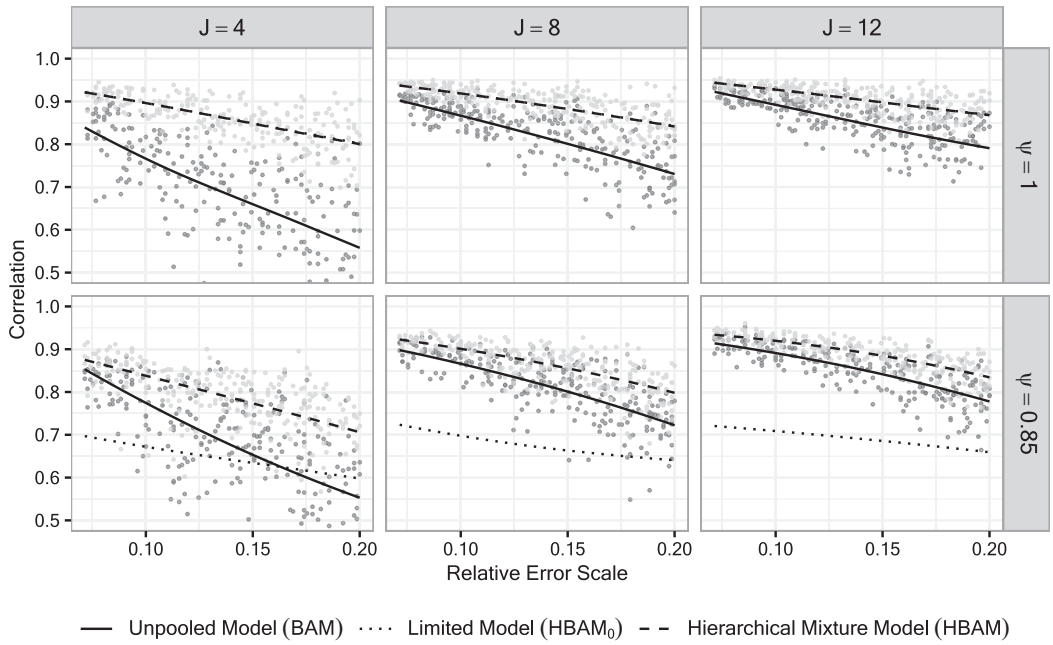

Figure 7. Results for posterior median respondent positions in the Monte Carlo study. The lines show loess smoothed curves, summarizing the correlations between estimates and true respondent positions in 1,000 trials per panel. The dots represent a random sample of 250 results per model (dark gray for the BAM model and light gray for the HBAM model). J is the number of stimuli, ψ is the probability that respondents do *not* flip the scale, and the relative error scale is τ divided by the scale length.

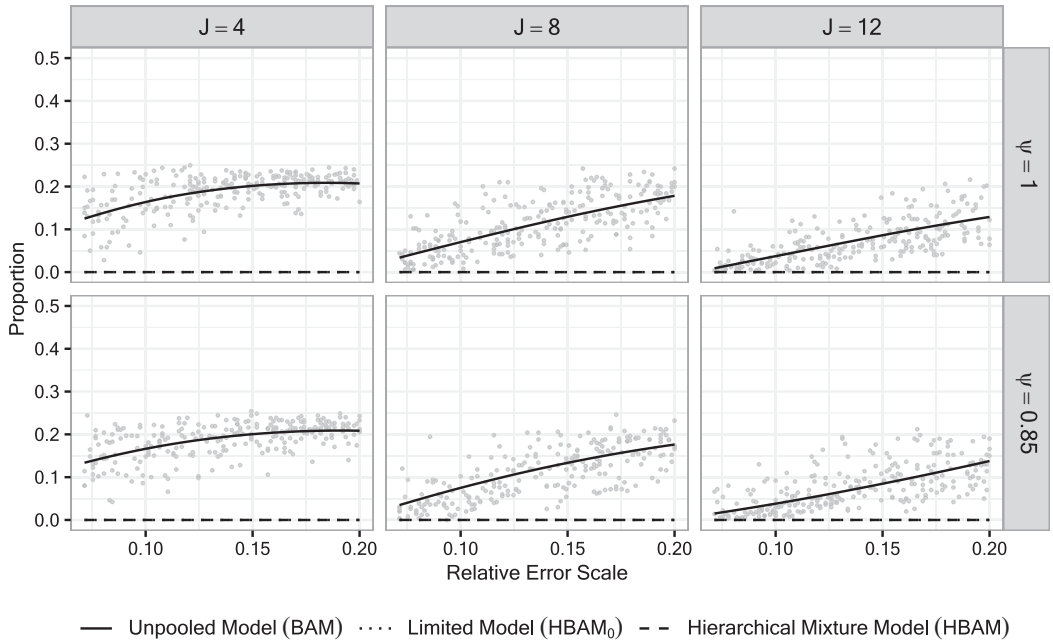

Figure 8. Shares of respondents with extremely wide credible intervals. The lines show loess smoothed curves based on 1,000 trials per panel. The dots represent a random sample of 250 results for the BAM model. J is the number of stimuli, ψ is the probability that respondents do *not* flip the scale, and the relative error scale is τ divided by the scale length.

Figure 8 shows the share of the respondents who receive extremely wide credible intervals. To count as extremely wide, the 95% intervals must be more than 50% wider than the interval containing 95% of the draws for all respondents. For the BAM model, the share of respondents who fall into this category is up to 25% in the worst cases and close to 0% in the best

cases—suggesting the result from the empirical application is fairly typical. Notably, there is no scenario in which the BAM model is guaranteed not to produce problematic posterior distributions for the respondent positions.

Results for the stimulus positions are reported in the Supplementary Material. The correlations between estimated and true stimulus positions are very strong for all models across all scenarios. However, the BAM model still performs marginally worse than the two other models, especially as the scale of the errors increases. The $HBAM_0$ model performs marginally better than the HBAM model when the error scale gets extreme, but both models perform well—and equally well for all realistic scenarios.

5 Conclusion

As our tools for statistical computing become more powerful and accessible, applied researchers will increasingly face difficult choices in specifying models and drawing inferences. In situations where the data are clustered, a key question is how to model parameters that vary by cluster. Researchers who have substantive interest in such parameters are often tempted to fit unpooled models: In a standard regression context, they may fit a separate model for each cluster, while in a Bayesian setting, they may fit a joint model where the cluster-level parameters are given uniform priors. However, these approaches are rarely optimal, as they tend to overfit the data.

The new version of the Bayesian Aldrich–McKelvey model developed in this article illustrates the advantages of Bayesian hierarchical modeling. The AM model makes an ideal case for hierarchical modeling because the individual-level parameters are among its most important outputs, and a typical dataset entails very few observations per individual. However, the new model also entails other improvements over previous versions: It treats respondents' self-placements as data to be modeled probabilistically and included in the likelihood, and their latent positions as parameters to be estimated. In addition, the model expands the likelihood function to allow for scale flipping while using hierarchical priors.

The new model (HBAM) outperforms the unpooled model (BAM) on all assessment criteria considered: (1) It has higher out-of-sample prediction accuracy on real data, (2) it is better at recovering the true positions of both respondents and stimuli using simulated data, and (3) it yields proper marginal posterior distributions for all respondent positions. This article has also reported results for a simpler version of the new model ($HBAM_0$), which could be sufficient in some rare scenarios where scale flipping is non-existent. However, the HBAM model performs as well as the simpler model even in these scenarios, and HBAM is therefore the default model in the R package accompanying this article.

Existing studies using the BAM model have often ignored the estimated respondent positions, which may partly be due to the issues highlighted in this article. By providing more accurate estimates of the respondent positions and their associated uncertainties, the new model invites greater use of these estimates in future research. The other individual-level estimates from the new model are also less noisy, and a recurring feature of the plots in this article is that the new model makes aggregate patterns clearer. In some instances, like Figure 3, the HBAM model brings out a theoretically expected pattern that cannot be seen in the results from the BAM model. This illustrates that these models can yield entirely different conclusions even at the aggregate level.

The *hbamr* package, which accompanies this article, can be used to address a wide range of questions. The package permits missing values in the data and therefore allows users to scale actors from different geographical districts onto the same political space—provided at least a few bridging stimuli are available. In this context, bridging stimuli are ones that respondents in multiple districts are exposed to, which makes them valuable as common anchor points. The BAM

model has been used for cross-district scaling in the past,[23] and the models in the *hbamr* package can be used the same way.[24]

A key finding of Hare *et al.* (2015) was that respondents tend to shift actors on their own side of the scale toward the center while shifting those on the other side further away. This is important because it will lead us to underestimate the degree of polarization if we rely on the raw data. Hare et al.'s finding was based on the ANES 2012, and their finding is substantively replicated in this article (see Figure 2). However, the Supplementary Material includes analyses of the ANES 1980, and neither model finds a clear pattern of the same kind in this older dataset. A plausible hypothesis is that this type of shifting has increased along with the degree of polarization in the electorate, and this is one of the many questions one could investigate using the *hbamr* package.

Acknowledgments

I am grateful to Bjørn Høyland, Philipp Broniecki, and three anonymous reviewers for helpful comments. I also thank those who have provided input on my earlier work in this area, especially Pablo Barberá and Liam Beiser-McGrath.

Data Availability Statement

Replication code for this article is available at https://doi.org/10.7910/DVN/VX0YPW (Bølstad 2023). The current release of the R package *hbamr* is available at https://cran.r-project.org/package=hbamr. The development version is available at https://github.com/jbolstad/hbamr/. For an introduction to the package, users should consult the vignette, which is linked from the CRAN and GitHub pages.

Supplementary Material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2023.18.

References

Aldrich, J. H., and R. D. McKelvey. 1977. "A Method of Scaling with Applications to the 1968 and 1972 Presidential Elections." *American Political Science Review* 71 (1): 111–130.

Alemán, E., J. P. Micozzi, P. M. Pinto, and S. Saiegh. 2018. "Disentangling the Role of Ideology and Partisanship in Legislative Voting: Evidence from Argentina." *Legislative Studies Quarterly* 43 (2): 245–273.

Bafumi, J., A. Gelman, D. K. Park, and N. Kaplan. 2005. "Practical Issues in Implementing and Understanding Bayesian Ideal Point Estimation." *Political Analysis* 13 (2): 171–187.

Bafumi, J., and M. C. Herron. 2010. "Leapfrog Representation and Extremism: A Study of American Voters and Their Members in Congress." *American Political Science Review* 104 (3): 519–542.

Bakker, R., S. Jolly, and J. Polk. 2020. "Analyzing the Cross-National Comparability of Party Positions on the Socio-Cultural and EU Dimensions in Europe." *Political Science Research and Methods* 10: 408–418.

Bakker, R., S. Jolly, J. Polk, and K. Poole. 2014. "The European Common Space: Extending the Use of Anchoring Vignettes." *Journal of Politics* 76 (4): 1089–1101.

Barberá, P. 2015. "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation using Twitter Data." *Political Analysis* 23 (1): 76–91.

Bølstad, J. 2019. "How Efficient Is Stan Compared to JAGS? Conjugacy, Pooling, Centering, and Posterior Correlations." Playing with Numbers: Notes on Bayesian Statistics. http://www.boelstad.net/post/stan_vs_jags_speed/.

Bølstad, J.. 2020. "Capturing Rationalization Bias and Differential Item Functioning: A Unified Bayesian Scaling Approach." *Political Analysis* 28 (3): 340–355.

Bølstad, J.. 2023. "Replication Materials for: Hierarchical Bayesian Aldrich–McKelvey Scaling." Harvard Dataverse. https://doi.org/10.7910/DVN/VX0YPW.

Carpenter, B., et al. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1): 1.

23 See, for example, Bakker *et al.* (2020).

24 If the HBAM model proves too slow for large datasets, the package includes a simplified model called $\text{HBAM}_{\text{MINI}}$, which samples considerably faster. The package also offers an even simpler model called $\text{FBAM}_{\text{MINI}}$, which can be fit using optimization when minimizing execution time is the main concern.

Carroll, R., and H. Kubo. 2018. "Explaining Citizen Perceptions of Party Ideological Positions: The Mediating Role of Political Contexts." *Electoral Studies* 51: 14–23.
Clay, K. C., R. Bakker, A.-M. Brook, J. Daniel, W. Hill, and A. Murdie. 2020. "Using Practitioner Surveys to Measure Human Rights: The Human Rights Measurement Initiative's Civil and Political Rights Metrics." *Journal of Peace Research* 57 (6): 715–727.
Clinton, J., S. Jackman, and D. Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98 (2): 355–370.
Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2014. *Bayesian Data Analysis*. 3rd edition. London: Chapman & Hall/CRC Press.
Hare, C., D. A. Armstrong, R. Bakker, R. Carroll, and K. T. Poole. 2015. "Using Bayesian Aldrich–McKelvey Scaling to Study Citizens' Ideological Preferences and Perceptions." *American Journal of Political Science* 59 (3): 759–774.
Hoffman, M. D., and A. Gelman. 2014. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research* 15 (1): 1593–1623.
Hollibaugh, G. E., L. S. Rothenberg, and K. K. Rulison. 2013. "Does It Really Hurt To Be Out of Step?" *Political Research Quarterly* 66 (4): 856–867.
Imai, K., J. Lo, and J. Olmsted. 2016. "Fast Estimation of Ideal Points with Massive Data." *American Political Science Review* 110 (4): 631–656.
Lo, J., S.-O. Proksch, and T. Gschwend. 2014. "A Common Left-Right Scale for Voters and Parties in Europe." *Political Analysis* 22 (2): 205–223.
Palfrey, T. R., and K. T. Poole. 1987. "The Relationship between Information, Ideology, and Voting Behavior." *American Journal of Political Science* 31 *:* 511–530.
Plummer, M. 2003. "JAGS: A Program for Analysis of Bayesian Graphical Models using Gibbs Sampling." In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Vienna, Austria.
Poole, K., J. Lewis, H. Rosenthal, J. Lo, and R. Carroll. 2016. "Recovering a Basic Space from Issue Scales in R." *Journal of Statistical Software* 69 (7): 1–21.
Poole, K. T. 1998. "Recovering a Basic Space from a Set of Issue Scales." *American Journal of Political Science* 42 (3): 954–993.
Poole, K. T., and H. Rosenthal. 1985. "A Spatial Model for Legislative Roll Call Analysis." *American Journal of Political Science* 29 (2): 357–384.
Poole, K. T., and H. Rosenthal. 1991. "Patterns of Congressional Voting." *American Journal of Political Science* 35 (1): 228–278.
Saiegh, S. M. 2009. "Recovering a Basic Space from Elite Surveys: Evidence from Latin America." *Legislative Studies Quarterly* 34 (1): 117–145.
Saiegh, S. M. 2015. "Using Joint Scaling Methods to Study Ideology and Representation: Evidence from Latin America." *Political Analysis* 23 (3): 363–384.
Vehtari, A., A. Gelman, and J. Gabry. 2017. "Practical Bayesian Model Evaluation using Leave-One-Out Cross-Validation and WAIC." *Statistics and Computing* 27 (5): 1413–1432.
Zakharova, M., and P. V. Warwick. 2014. "The Sources of Valence Judgments: The Role of Policy Distance and the Structure of the Left–Right Spectrum." *Comparative Political Studies* 47 (14): 2000–2025.