Appendix: A Multinomial Framework for Ideal Point Estimation

Max Goplerud*

March 22, 2018

A. APPENDIX A: THE CHOICE OF ORDERING

The main paper noted that an ordering of categories must be imposed to use the stick-breaking representation. The choice of ordering is not discussed in detail in Linderman et al. (2015) (although they note that their results are similarly robust to order permutations), perhaps because the models estimated do not parameterize the stick-breaks in a way where the ordering may have more importance. I claimed that the results of the model for the key quantities of interest (ideal points and the predicted probabilities of responses as ideal points vary) are fairly invariant to the choice of ordering. Thus, this section provides an analytic and simulation-based justification for this claim.

First, I show that one can think of the stick-breaking representation as an approximation of the classic multinomial logistic regression. As the latter does not specify an ordering, the fact that the stick-breaking representation is an approximation of a representation where order does not matter should give us some hope that the order chosen is fairly unimportant.

Second, I show that in both the simulated and actual ANES data, the quantities of interest estimated via the stick-breaking representation are highly insensitive to the ordering chosen. Indeed, I show that any differences in the ideal points recovered by different orders are generally much smaller than the uncertainty associated with those estimates and thus, whilst there are limited effects to choosing an ordering, they will tend to 'wash out' compared to other sources of uncertainty. For the ANES data, I compare the point estimates from the EM algorithm under multiple orderings against the ideal points estimated using gradient descent on a 'classic' multinomial formulation; I show that under all permutations, they are nearly perfectly correlated.

Third, I discuss some diagnostic tools to examine which ordering performs 'best', although they should be used with caution in conjunction with the EM estimation framework.

Finally, I discuss some extensions to the model above that would further relax the sensitivity of the model to the choice of ordering.

A.1. *Linear Approximations to Softmax Regression*

As the classic multinomial version of logistic regression is parameterized via a softmax formulation, I now derive some results to show that two approximations of this framework coincide with the stick-breaking formulation above. For notation, the classic way of defining a multinomial outcome in the ideal point context is as follows:

$$\Pr\left(y_{ij} = k\right) = \frac{\exp(\alpha_j^k + \gamma_j^k x_i)}{\sum_{l=1}^{K_j} \exp(\alpha_j^l + \beta_j^l x_i)}$$

Given some arbitrary ordering O_j, we can write the stick-break for some category k as follows (the sum now ranges from k to K_j rather than from 1 to K_j):

*Code to implement the models in the paper, and the `mIRT` more generally, can be found at http://dx.doi.org/10.7910/DVN/LD0ITE

$$\Pr\left(y_{ij} = k | y_{ij} \geq k\right) = \frac{\exp(\alpha_j^k + \gamma_j^k x_i)}{\sum_{l=k}^{K_j} \exp(\alpha_j^l + \gamma_j^l x_i)}$$

With some re-arrangement, we can write this as:

$$\Pr\left(y_{ij} = k | y_{ij} \geq k\right) = \frac{\exp\left(\alpha_j^k + \gamma_j^k x_i - \ln(\sum_{l=k+1}^{K_j} \exp(\alpha_j^l + \gamma_j^l))x_i)\right)}{1 + \exp\left(\alpha_j^k + \gamma_j^k x_i - \ln(\sum_{l=k+1}^{K_j} \exp(\alpha_j^l + \gamma_j^l x_i))\right)}$$

The stick-breaking representation can thus be seen as approximating the complicated term inside the exponential with an affine function of x_i that hopefully captures much of the salient information. I show that two common functional approximations return this linear form. For clarity, I focus on the unidimensional models. First, consider a way of bounding the log-sum-of-exponentials:

$$\max\{\alpha_j^l + \gamma_j^l x_i\}_{l=k+1}^{K_j} \leq \ln\left(\sum_{l=k+1}^{K_j} \exp\left(\alpha_j^l + \gamma_j^l x_i\right)\right) \leq \max\{\alpha_j^l + \gamma_j^l x_i\}_{l=k+1}^{K_j} + \log(\#terms)$$

This states that the log-sum-of-exponentials is bounded below by the largest term in the summation and above by the largest term plus the log of the number of terms in the summation, i.e. the number of categories that are after k in the ordering. As most questions rely on fairly small numbers of categories (e.g. a seven-point scale), this bound is thus quite tight in most actual applications. Thus, if it is the case that there is a unique maximum, i.e. $\max \alpha_j^l + \gamma_j^l x_i{}_{l=k+1}^{K_j} = \alpha_j^* + \gamma_j^* x_i$ for all x_i or all x_i in the space that contains the estimated ideal points, then the linear approximation is quite good as the multinomial softmax can be tightly approximated in terms of stick-breaks as the following:

$$\Pr\left(y_{ij} = k | y_{ij} \geq k\right) \approx \frac{\exp(\alpha_j^k - \alpha_j^* + (\gamma_j^k - \gamma_j^*)x_i)}{1 + \exp(\alpha_j^k - \alpha_j^* + (\gamma_j^k - \gamma_j^*)x_i)}$$

With this approximation, it is clear that the stick-breaking formulation shown above recovers a model of this form where $\kappa_j^k = \alpha_j^k - \alpha_j^*$ and $\beta_j^k = \gamma_j^k - \gamma_j^*$. As this approximation requires the maximum of the constituent terms of the log-sum being constant or approximately so (and, indeed, knowing the ordering that makes this hold), a different and more flexible approximation relies on a Taylor expansion of the log-sum-of-exponentials. Calculating the expansion around $x_i = 0$ yields:[1]

$$\ln\left(\sum_{l=k+1}^{K_j} \exp\left(\alpha_j^l + \gamma_j^l x_i\right)\right) \approx \ln\left(\sum_{l=k+1}^{K_j} \exp(\alpha_j^l)\right) + x_i \frac{\sum_{l=k+1}^{K_j} \gamma_j^l \exp(\alpha_j^l)}{\sum_{l=k+1}^{K_j} \exp(\alpha_j^l)}$$

By a similar argument to above, we see that the stick-breaking representation can be thought of as encoding a first order approximation to the classic multinomial representation.

A.2. *Simulations on Differing Orderings*

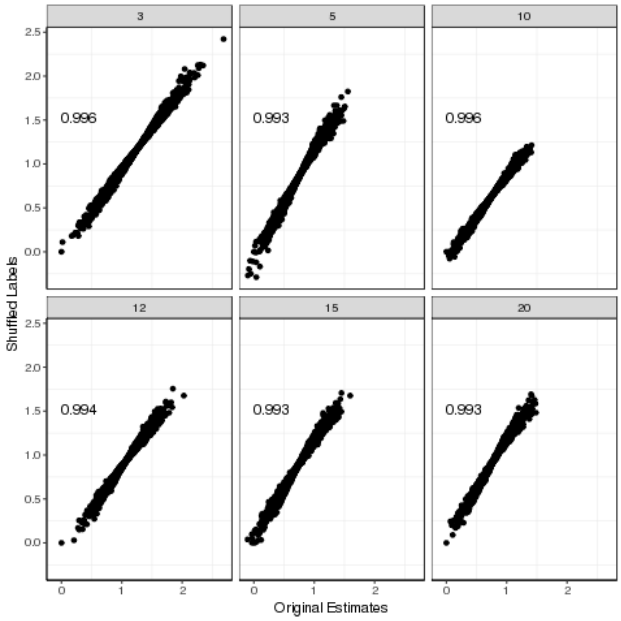
Given an analytic justification for why the ordering of the stick-breaking representation may not be especially important, it is important to test this in practice with simulated data.[2] I return to the simulated data in the main text and re-run the model ten times for each set of simulated parameters with different random ordering of the categories for each question. To provide some visual evidence, Figure 1 reports the correlation

[1] As the priors on x_i ensure that the resulting distribution is anchored around 0, this seems to be a reasonable value around which to calculate the expansion.

[2] In the following empirical analysis on the American National Election Study, I also report the results of varying the ordering to show that the near equivalence of results in response to changing orders is not merely an artifact of my 'nice' simulated data.

between one permutation and the original estimates shown above. We see they are highly correlated across all M (maximum number of response categories). We also see that the differences that do exist are quite small and would be washed out by taking into account the uncertainty inherent in each estimated ideal point. Across all permutations, the lowest correlation between any permuted set of labels and the truth is 0.925; the lowest correlation between any set of estimates is 0.99!

Figure 1: Re-Ordering Multinomial Data



NB: Each panel indicates the M, i.e. that each question j is sampled from $K_j \in \{2, \cdots M\}$.

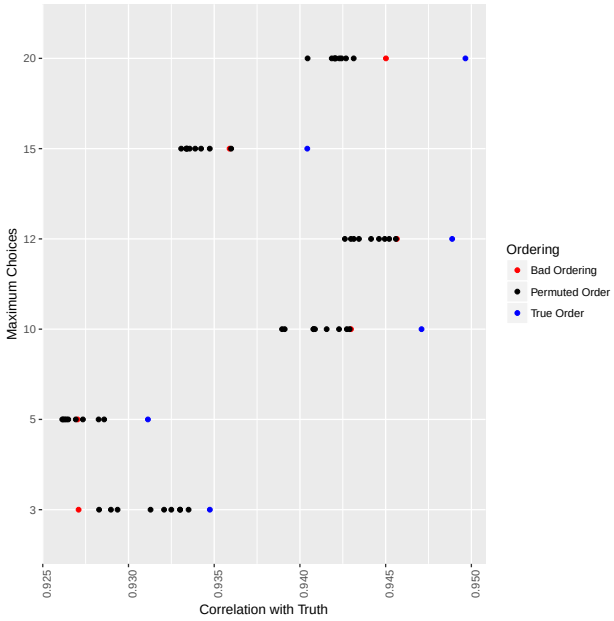
At the suggestion of a helpful reviewer, I also ran a test using a deliberately *bad* ordering; specifically, assume that we can order the choices based on where the modes of their predicted probabilities are. Call this the 'true' ordering T_j. I create a bad ordering B_j by placing the median item in T_j as the first item in B_j; the item in position two in B_j is the first item in T_j; the item in position three in B_j is the last item in T_j. I continue this 'zipper' type relationship to build an ordering that is very far from the true one implied by the model. The results from this bad ordering are shown below in Figure 2. The y-axis shows the maximum number of choices permissible. The x-axis shows the correlation with the truth. The 'true' ordering is shown in blue; the random permutations are in black; the 'bad' permutation is in red. We see that using the correct ordering typically has a higher correlation with the truth, but that the bad order is not systematically below the (still very high) correlation with random permutations. It's important to note these differences are slight; the correlations between *any* ordering and the truth never falls below 0.925.

A.3. *Analysis of the ANES*

Turning to actual data, I show that the same properties of invariance of the quantities of interest hold when using 'messier' real data. Specifically, I re-run the model where I 'insert' the non-response response in different places in the question ordering: I do this where I include it in the first, second, third, fourth, fifth, and last positions.[3] The ideal points are nearly identical across all models; even including when they are placed first and thus required to have a monotonic effect, the smallest pairwise correlation between the

[3]If there are fewer than five responses, I put it as the last position.

Figure 2: Re-Ordering Multinomial Data with a 'Bad' Ordering



estimates is above 0.999. Despite some differences between some of the ideal points at the extremes, this should give us confidence that the procedure is robust to the ordering imposed on the multinomial outcomes. It is also clear that when looking at the uncertainty attached to any of these ideal points, that will dwarf the variability that comes from scaling the questions using a different implied ordering.

I then replicated Figures 5 ('Predicted Probabilities for Moral Questions') and 6 ('Probability of Non-Response') from the main paper using the permuted orderings. We see that, whilst there are some differences in the posterior means as indicated by the solid lines, the intervals mostly overlap especially in the region where most ideal points are located. To the extent we see differences that are distinguishable from variability in the quantities of interest, they mostly occur in the tails of the distribution. It is also interesting to note that for categories with relatively large numbers of outcomes (i.e. the categories in the 'often prayer' and 'same sex marriage' question), the differences between the orderings are very small indeed.

Finally, I also used gradient descent in Python to estimate a model based on the classic multinomial formulation (softmax) noted above. In this framework, it is not easy to impute missing data and thus I run the model based only on the observed responses. Figure 6 shows the scatterplot between the preferred `mIRT` method (i.e. 'don't know' being position last), the classic multinomial point estimates, and factor analysis.

Visually, we see that both methods for scaling multinomial data return highly correlated responses with the correlation between the classic multinomial and the mIRT being 0.989! We also see that this does not depend on the ordering as the smallest entry in Table 1 is 0.988. My sense is that the differences that arise are also partially driven by the lack of imputation in the classic multinomial framework estimated via gradient descent.

This is hopeful decisive proof that, at least for the ANES data, using the stick-breaking approximation returns (a) nearly identical ideal points to that from using a classic multinomial formulation and (b) returns the correct results regardless of where the 'don't know' is inserted into the ordering.

Figure 3: Re-Ordering ANES: Same Sex Marriage

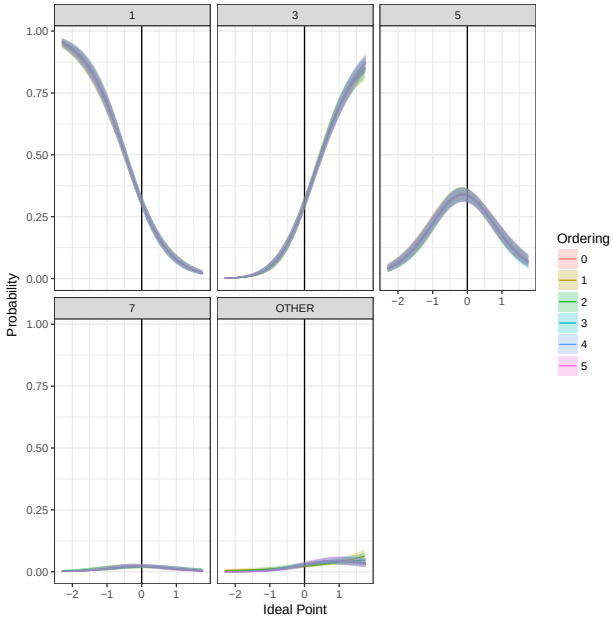


Table 1: Correlation of Ideal Points with Re-Ordering and Classic Multinomial

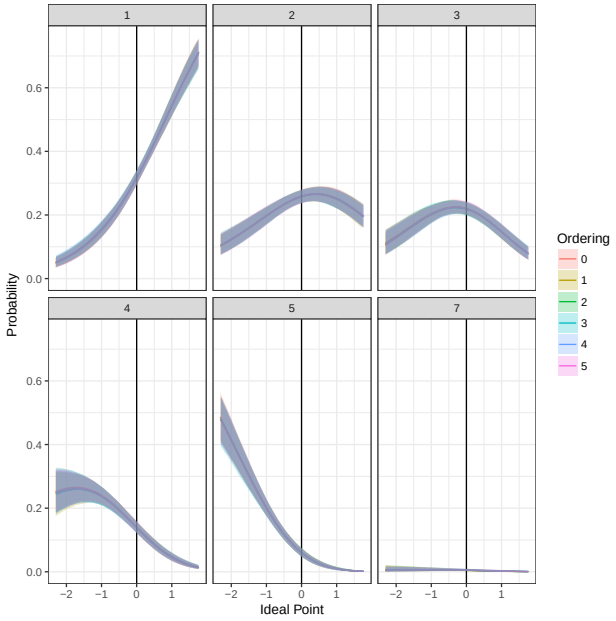
	0	1	2	3	4	5	Classic Multinomial
0	1.000	0.993	0.998	0.998	0.998	0.999	0.989
1	0.993	1.000	0.998	0.997	0.996	0.996	0.988
2	0.998	0.998	1.000	1.000	0.999	0.999	0.990
3	0.998	0.997	1.000	1.000	1.000	1.000	0.990
4	0.998	0.996	0.999	1.000	1.000	1.000	0.991
5	0.999	0.996	0.999	1.000	1.000	1.000	0.991
Classic Multinomial	0.989	0.988	0.990	0.990	0.991	0.991	1.000

A.4. *Model Selection Based on Orderings*

One possible strategy for deal with choosing an ordering is to run the model with a number of different orderings and then select the one that has the best model fit. There are many different ways to examining model fit in a Bayesian context, see Gelman et al. (2013) for a detailed discussion. One simple approach that could be done using the quantities returned by the EM estimation framework is to use those estimates of the posterior model and calculate the associated log-likelihood of the data. As the number of free parameters is constant in the model, this can be seen as roughly analogous to using an AIC or BIC selection rule insofar as the penalty term is constant and thus the variation in fit is driven solely by the log-likelihood. As there are good objections to the AIC and BIC in a Bayesian context (Gelman et al., 2013), one could also estimate other criteria, e.g. the WAIC (Watanabe, 2010) or Pareto Smoothing Importance Sampling (Vehtari et al., 2017), to engage in model selection. Note, however, that these methods require a sample from the posterior and thus one would need to use the Gibbs Sampler implementation to evaluate the orderings using this metric. A combination approach, that I use later to generate estimate of uncertainty, would take the EM point estimates of the posterior mode as starting values for a Gibbs Sampler to hopefully achieve fairly rapid convergence even with large datasets.

Another potential downside of relying on the EM point estimates is that even when the model is run

Figure 4: Re-Ordering ANES: Prayer



for a long period of time (e.g. until all parameters correlate with their previous iteration at $1 - 10^{-6}$), the values of the AIC and BIC appear reasonably sensitive to different initializations and stopping rules. This is perhaps because given a large amount of data, slight differences in the point estimates (and the lack of an inherent underlying scale of the ideal point model) might translate into reasonably different log-likelihoods. However, to show that this approach is at least roughly on the correct track, Figure 7 plots the log-likelihood evaluated at the posterior mode against the correlation between the estimated ideal points and the truth for the ten permutations run above.

We see that whilst there are differences in the log-likelihood values, these are both comparably small relative to the overall magnitude of the log-likelihood as well as corresponding to very small changes in the correlation between the estimated ideal points and the truth. It is important to note, however, that the 'correct' ordering does have both the highest correlation with the truth and the highest log-likelihood.

Given the sensitivity of the value of the log-likelihood when using the plug-in method from the EM algorithm, my tentative advice to researchers would be to try to impose a sensible ordering based on prior knowledge as well as ensuring the option in the first category can be assumed to be monotonic based on prior knowledge about the latent scale. If concerns about the ordering are especially salient, using the model with different orderings and then using a Gibbs Sampler to allow a proper estimation of model fit (rather than relying on the point estimates used in the discussion above) is probably a safer approach. However, as all of the results in this section suggest that, in practice, the ordering does not matter too much and thus picking a plausible ordering and checking that the results are stable to some number of random permutations is likely also a sensible way to proceed.

A.5. *Further Ways to Relax The Functional Form Assumption*

This section briefly outlines two further ways to relax the functional form assumption. First, one could rely on a quadratic link function in terms of the ideal points: In the notation above, $\psi_{ij}^n = \kappa_j^n + \beta_j^n x_i + \nu_j^n x_i^2$. This would correspond to a second-order approximation to the log-sum-exponential and thus may have better correspondence with that model. As the E-step remains materially unchanged, the only further

Figure 5: Re-Ordering ANES: 'Non-Response'

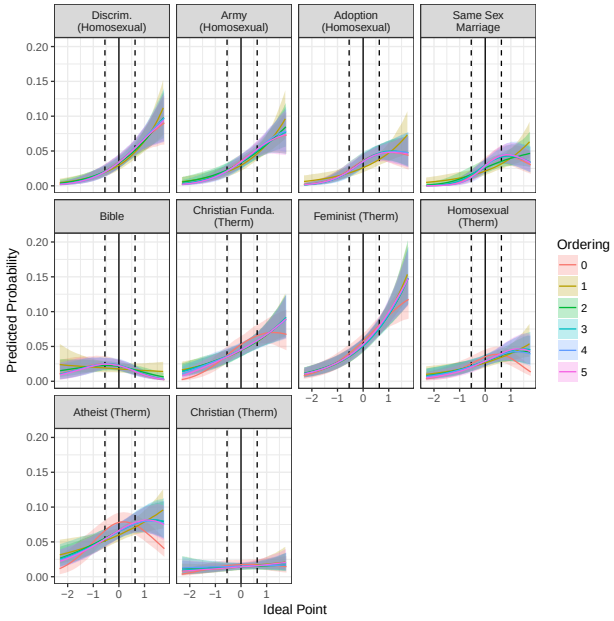
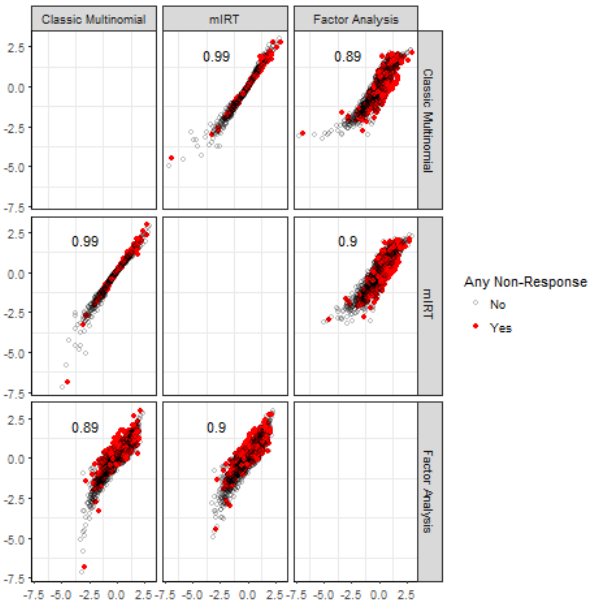
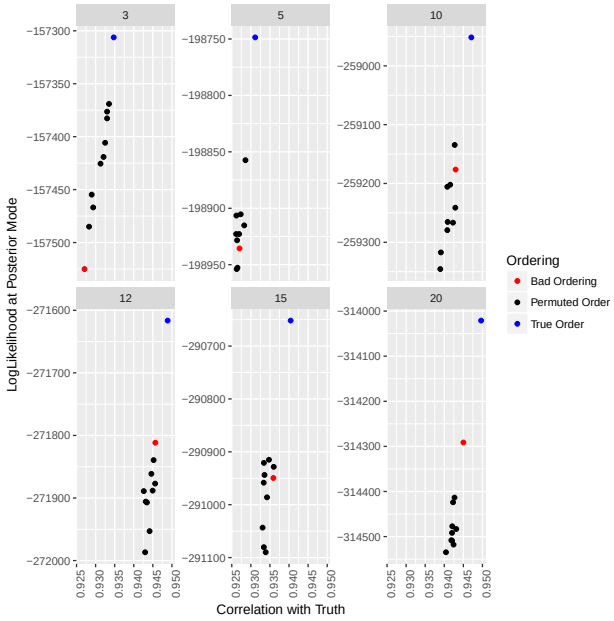


Figure 6: ANES: mIRT vs Classic Multinomial



challenge of including a quadratic link function is that the M-step may become more complicated. As this simply involves finding the roots of a polynomial equation, however, this can be done quickly using standard statistical software and still allows for an exact EM implementation. It may however lead to complications in the x_i updates for a Gibbs Sampler, but one could then rely on a 'Metropolis-in-Gibbs' or some alternative framework for those updates.

Figure 7: Comparison of Orderings



NB: Each panel indicates the M, i.e. that each question j is sampled from $K_j \in \{2, \cdots M\}$.

A different tact to relaxing the functional form assumption, as well as the particular form of IIA imposed above, would be to implement the IRT equivalent of a 'mixed logistic regression'. In the classic formulation, this is an attempt to relax IIA by introduce a random error term to each choice level, i.e. ϵ_{ij}^n, where the ϵ are allowed to be correlated as they are assumed to come from a multivariate normal distribution with an unconstrained correlation matrix. This could be thought of as a particular type of random effect for each i-j combination that induces dependence between the outcome categories and thus increases the flexibility of the model. It can be shown in a regression context that a particular formulation of ϵ leads to the multinomial logistic regression approximating the multinomial probit (Train, 1998; McFadden and Train, 2000) and thus a similar insight may apply here. This would likely require a more complicated E and M step and thus is outside of the paper to derive directly, but it is an interesting question for future research.

B. APPENDIX B: IDENTIFICATION

As Rivers (2003) notes, identification in ideal point models turns on the concept of 'observational equivalence'. To define the concept briefly, assume the true parameter values are $\theta = (\boldsymbol{\kappa}_j^n, \boldsymbol{\beta}_j^n, \boldsymbol{x}_i)$ and that these are recovered from the Bayesian or EM approach. Identification means that there must not be some other θ' that is observationally equivalent, i.e. that the likelihood function or posterior distribution is identical for all observations for events that occur with non-zero probability. A model is said to be 'locally identified' if and only if there is some neighbourhood N around the proposed solution θ such that no other θ' in said neighbourhood is observationally equivalent to θ. Rivers (2003) provides a more extensive discussion.

For my purposes, it is sufficient to note that Rivers's proof (or others) of the restrictions necessary to identify binary ideal point models can be applied without modification to the `mIRT`'s core multinomial model. Recall that the stick-breaking representation says that for some bill j and MP i, their decision can be thought of as making $K_j - 1$ binary choices, e.g. $\Pr(y_{ij} = 1)$, $\Pr(y_{ij} = 2 | y_{ij} \neq 1)$, etc., that are independent. i's revealed outcome $y_{ij} = n$ is thus a deterministic function of those stick-breaking binary choices, e.g. reveal 2 if the first binary choice is 'no' and the second binary choice is 'yes'. Given this view of the model, it fits exactly into the data generating process discussed by Rivers (2003), i.e. a series of binary choices that are independent with the linear two parameter (κ, β) specification. Thus, his proof of multidimensional identification (or others) is sufficient for identification in the `mIRT`.

C. APPENDIX C: DERIVATION OF DYNAMIC IDEAL POINT MODELS

Recall from before that g denotes individuals and i denotes individuals-in-particular Congresses, e.g. $x_i^{(g)}$ is John Kerry in a particular Congress. This model changes the independent normal priors on x_i to a normal prior that has mean $x_{i-1}^{(g)}$ and variance Δ. I derive results here for the binary case for expositional simplicity; the multinomial case is derived analogously. The Q function now sums over g, i, and j and becomes, suppressing the notation of constants and unrelated priors. Further, to avoid unbearable notational clutter, I suppress the indexing of the parameters by iteration t.

$$Q(\theta, \theta^{(t-1)}) \propto \sum_g \sum_i \sum_j (y_{ij} - 1/2)\psi_{ij} - \omega_{ij}^* \psi_{ij}^2/2 - \frac{1}{2\Delta}(x_i^{(g)} - x_{i-1}^{(g)})^2; \quad \omega_{ij}^* = \mathbb{E}\left[\omega_{ij} | y_{ij}, \theta^{(t-1)}\right]$$

The E-step is unchanged as the prior does not affect this calculation. The M-step for the $x_i^{(g)}$ is the only different step; I focus on the one-dimensional case. To begin, we note that the posterior of $p(\boldsymbol{x}^{(g)}|-)$ is multivariate normal given the Pólya-Gamma augmentation variables. If this was an MCMC approach, we could thus rely on the standard Kalman filtering and smoothing to sample from the full conditional. For our purposes, however, we can simply find the means of the multivariate normal $\boldsymbol{x}^{(g)}|-$ and this is the EM update. We note that this will occur where the partial derivative of Q equal zero for all $x_i^{(g)}$. Define f as the first period and l as the last period where unit g appears. Note that the prior on $x_f^{(g)} \sim N(\mu_0^{(g)}, \Delta_0^{(g)})$

Let us focus on some unit g. Define $B_i^{(g)} = \sum_j \beta_j (y_{ij} - 1/2 - \omega_{ij}^* \kappa)$ and $O_i^{(g)} = \sum_j \omega_{ij}^* \beta_j^2$. For all $i > f$, i.e. all but the initial period, I also add Σ_x^{-1} to $O_i^{(g)}$ to add an additional degree of regularization. This also means that as $\Delta \to \infty$, all periods still have the stabilizing effect of the prior in the basic `mIRT`. I have found from experience on the SCOTUS data that this random walk bridging is unstable and thus needs extra regularization to stabilize the model. This is especially true in cases where the i may have relatively few votes. If their ideal point is unstably estimated and the pooling effect from adjacent periods is weird, this seems to disrupt their entire ideal point sequence.

$$\frac{\partial Q}{\partial x_f^{(g)}} = B_f^{(g)} - O_f^{(g)} x_f^{(g)} + \frac{1}{\Delta}(x_{f+1}^{(g)} - x_f^{(g)}) - \frac{1}{\Delta_0}(x_f^{(g)} - \mu_0^{(g)})$$

$$\frac{\partial Q}{\partial x_i^{(g)}} = B_i^{(g)} - O_i^{(g)} x_i^{(g)} + \frac{1}{\Delta}(x_{i+1}^{(g)} - x_i^{(g)}) - \frac{1}{\Delta}(x_i^{(g)} - x_{i-1}^{(g)})$$

$$\frac{\partial Q}{\partial x_l^{(g)}} = B_l^{(g)} - O_l^{(g)} x_l^{(g)} - \frac{1}{\Delta}(x_l^{(g)} - x_{l-1}^{(g)})$$

The set of $\boldsymbol{x}^{(g)}$ that set all of these equations equal to zero thus is the M-update. There are many ways to solve this. Given the relatively short duration of the time series, we can simply set this up as a (sparse) system of linear equations and solve. To note:

$$\begin{aligned}
B_f^{(g)} + \frac{1}{\Delta_0}\mu_0^{(g)} &= 0x_{f-1}^{(g)} + & \left(O_f^{(g)} + \frac{1}{\Delta} + \frac{1}{\Delta_0}\right) x_f^{(g)} - \frac{1}{\Delta} x_{f+1}^{(g)} \\
B_i^{(g)} &= -\frac{1}{\Delta} x_{i-1}^{(g)} + & \left(O_i^{(g)} + \frac{2}{\Delta}\right) x_i^{(g)} - \frac{1}{\Delta} x_{i+1}^{(g)} \\
B_l^{(g)} &= -\frac{1}{\Delta} x_{l-1}^{(g)} + & \left(O_l^{(g)} + \frac{1}{\Delta}\right) x_l^{(g)} - 0x_{l+1}^{(g)}
\end{aligned}$$

Given the sensitivity of these models to Δ and the starting variances and means of the ideal points, I would suggest that this model be run with some caution in the absence of other bridging mechanisms.[4] For example, Bailey et al. (2017) use repeated resolutions to bridge certain β_j^n across sessions. If one does not have bridging legislation and uses a moderate Δ, one might try running the model to completion and if an individual i with a known polarity changes sign or moves dramatically without good theoretical cause, then try re-running the model with stronger, i.e. more informative, starting priors on that individual. Alternatively, one could run the session-by-session model and ensure that, say, Scalia was always given a positive ideal point. Then, using the estimated θ as starting values for the dynamic model may provide sufficient anchoring to ensure that plausible results are recovered. This could be seen as a solution in the spirit of Nokken and Poole (2004)'s work on estimating Congress-by-Congress ideal points. These questions remain unresolved and are a plausible avenue for future research.

D. APPENDIX D: IMPUTATION IN THE `MIRT` FRAMEWORK

The model derived in the main text assumes that y_{ij} is observed for all i, j. If it is not, and one wishes to model non-response as missing-at-random (rather than a separate category), imputation of the y_{ij} should occur in the E step.[5] The Q function must now consider that we are augmenting over not only ω_{ij} but also the missing y_{ij}. Yet, we know that given θ, y_{ij} is distributed binomial with probabilities given to us by the logistic link. To begin, focus on a single vote y_{ij} that is missing. The relevant terms of the Q function are shown below. The summand must run from $n = 1$ to $K_j - 1$ (the largest value) since we cannot exclude any augmented ω_{ij} *a priori* because we do not observe y_{ij}.

$$\sum_{n=1}^{K_j-1} s_{ij}^n \psi_{ij}^n - \omega_{ij}^{n*}/2(\psi_{ij}^n)^2$$

We note that y_{ij} only enters into the Q function via the $s_{ij}^n = I(y_{ij} = n) - 1/2$ term. Thus, for the E-step, we simply need to calculate the following additional terms (in addition to the $\mathbb{E}\left[\omega_{ij}^n | \theta^{(t-1)}\right]$):

$$p_{ij}^n = \mathbb{E}\left[I(y_{ij} = n)|\theta^{(t-1)}\right] = \Pr\left(y_{ij} = n|\theta^{(t-1)}\right)$$

[4] Indeed, even when increasing the stopping tolerance for the EM algorithm, different random restarts lead to sometimes different ideal points—especially for those justices in the first session (i.e. with relatively little prior information to bridge on).

[5] Whilst we could drop the non-response, as would be needed in bridging applications, we should generally impute it to avoid some respondents being estimated on a very small number of votes and thus having highly unstable or inappropriately extreme ideal points. This is the approach taken in Imai et al. (2016) that requires imputation of missing data—which unfortunately renders their package unable to engage in bridging applications at the time of writing.

This can be done directly from the data generating process: Find the stick-breaking probabilities using the logistic link and transform them to get the 'true' probabilities. For helpful notation define, $c_{ij}^n = 1 - \sum_{k>n}^{K_j} p_{ij}^n$, i.e. c_{ij}^n is the probability that $y_{ij} > n$. Plugging these into the above equation and re-arranging terms shows that the relevant portion of the Q function can be expressed as

$$\sum_{n=1}^{K_j-1} \frac{1}{2}(p_{ij}^n - c_{ij}^n)\psi_{ij}^n - \omega_{ij}^{n*}/2(\psi_{ij}^n)^2(p_{ij}^n + c_{ij}^n)$$

Thus, for observations (i, j) where y_{ij} is missing, one can use the same M-step results derived above if one re-defines (with a notational slight of hand) s_{ij}^n as $\frac{1}{2}(p_{ij}^n - c_{ij}^n) + 1/2$ and $(\omega_{ij}^n)^*$ as $(\omega_{ij}^n)^*(p_{ij}^n + c_{ij}^n)$.

To demonstrate the importance of imputation, consider the Ashai Todai dataset analysed earlier; it has 20,000 respondents and 100 questions, although most respondents answer only a small fraction of the data. In the figure in the main text, I deal with missing data by 'imputing' outcomes as the EM algorithm runs, following Imai et al. (2016).[6] This 'rolling imputation' is, however, only one strategy to deal with large quantities of missing data and implies that even if a respondent answers very few questions or votes on few votes, they can still be included in the analysis. To show how this affects the results, I present two additional figures that run the `mIRT` model (a) without imputation, i.e. dropping all missing responses, and (b) the model without imputation but excluding the approximately 6,300 respondents who answered fewer than 5 questions and, arguably, should not be included in the analysis. Both strategies return very similar results as the the correlation between those two methods is rounded up to '1' at three decimal points.

The figures show that, unsurprisingly, without imputation there are a number of respondents who are assigned relatively different ideal points than the `emIRT` estimates, although the correlation between all methods is still quite high. This seems to vary based on starting values (even with the same, high, stopping tolerance) and can be seen suggestively in the figure by the set of individuals who are given ideal points around 0 in the `mIRT` but have a wider variation of ideal points in the `emIRT` model. It is hard to know exactly what is driving this, e.g. the variational approximations of the `emIRT` or the lack of data for these individuals, but this divergence may be driven by some β_j^n being highly mis-estimated in the presence of large amounts of missing data without imputation.

E. APPENDIX E: COMPARISON OF ESTIMATION METHODS FOR IDEAL POINTS

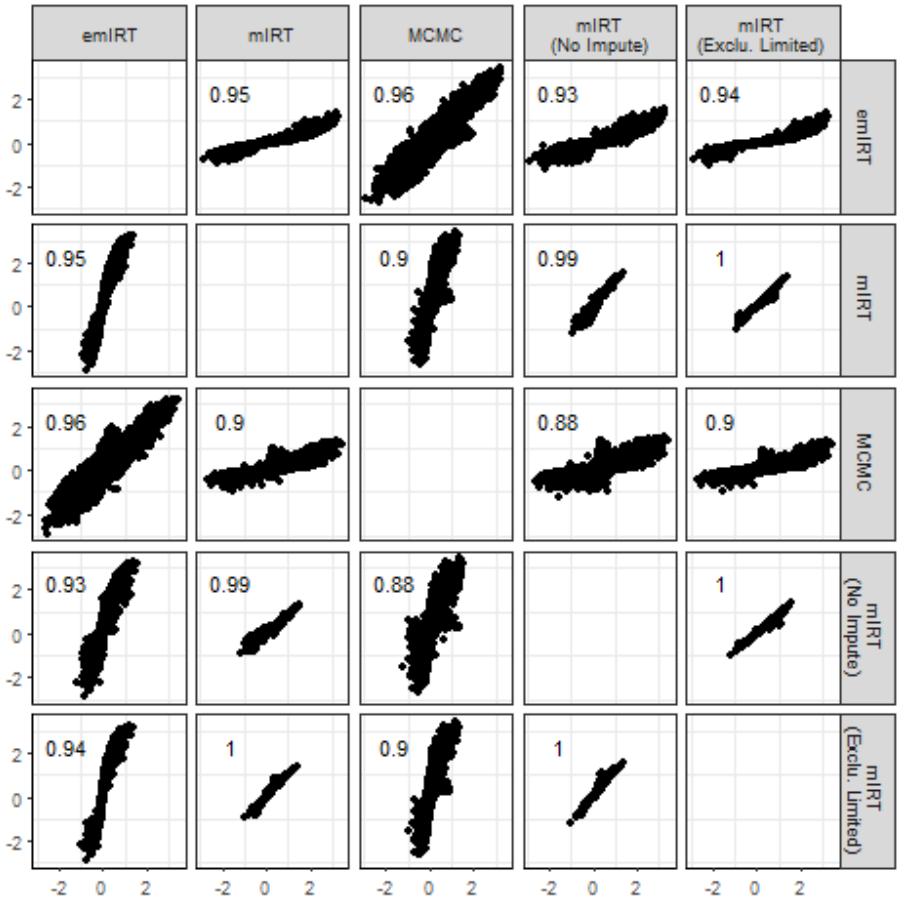
When estimating ideal point models, a key question is whether one cares about estimates of uncertainty. If the answer is 'no', then a classic approach (e.g. NOMINATE) is to use some optimization algorithm (e.g. gradient descent) to find the maximum likelihood or posterior mode of the model. A different approach, and the one used in this paper, is to use EM to find the posterior mode; whilst sometimes slower than gradient descent, it is more stable in that it is *guaranteed* to find a (local) mode without any tuning parameters being required and will, at each iteration, be guaranteed to increase the objective function.

The other common approach is 'variational inference' [VI] (see Grimmer (2010) for an introduction for political scientists). Imai et al. (2016) deploy this strategy to great effect on ideal point models by showing that we can *approximate* the posterior (or likelihood) and then use EM to deterministically find the maximum of this *approximate* distribution. VI also provides some rough measure of uncertainty of the parameter estimates, although these are common agreed severely biased downwards and thus unreliable.

In terms of speed, VI and EM are roughly equivalent in theoretical terms (though will depend on the specific implementation) as they are deterministic. For the purposes of ideal point estimation, therefore, EM is more theoretically justified as it is finding the maximum of the *actual* model whereas VI is finding the maximum of a (perhaps poor) approximating distribution. Typically scholars resort to VI when the underlying model is intractable as it is a 'last resort' when exact inference is intractable or computationally demanding, but as the underling `mIRT` model makes ideal point models for binary, multinomial and common

[6] The MCMC framework deals with this by 'imputing' outcomes as the sampler runs and Imai et al. (2016) do something analogous in the E-step by noting that when y_{ij} is missing, there is no information on how the latent utility is truncated.

Figure 8: Alternative Strategies for Missing Data



extensions (e.g. dynamic ideal points) tractable, it seems to be preferable on theoretical grounds. Of course, not all models admit a simple EM representation and in those cases, VI is a reasonable strategy.

If one cares about uncertainty in the estimates, it is necessary to use Monte Carlo methods (classically MCMC but see also Hamiltonian Monte Carlo) to generate samples that will approximate the true posterior distribution. These models also have the benefit to not get stuck in local modes in ways that are possible for both EM and VI. Existing ideal point frameworks for this exist, e.g. in `MCMCpack`, that rely on more complicated samplers for non-binary models as the implied models lack a tractable distribution. These procedures therefore have internal tuning parameters that must be set or calibrated for the particular case. By contrast, the `mIRT` allows for these models to be estimated via a straightforward data augmentation procedure where the full conditionals of the parameters are Gaussian, conditional on the augmentation variables. This procedure is transparent, stable, and lacks tuning parameters. It may be somewhat slower, but there is a large literature on how to speed up or parallelize Gibbs Samplers. In some sense, the MCMC models are the most flexible (as one can model whatever one desires), but the reason to focus on the EM or VI representations is for both speed and scalability for large datasets.

Overall, the `mIRT` represents a synthesis of existing frameworks in that it allows most of the common ideal point models estimated in Imai et al. (2016) to be estimated *exactly* (without need for variational approximations) using an EM framework whereas *also* allowing the same models to be estimated and extended via the implied MCMC framework. For some models (e.g. Poisson scaling), the VI approach of Imai et al. (2016) is superior; for other complex models, it is necessary to resort to MCMC. Yet, the benefit of this data

augmentation framework is that it unifies many existing models into a more tractable framework.

References

Bailey, M. A., Strezhnev, A., & Voeten, E. (2017). Estimating dynamic state preferences from united nations voting data. *Journal of Conflict Resolution*, *61*, 430–456.

Gelman, A., Carlin, J. B., Sterm, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Chapman and Hall.

Grimmer, J. (2010). An introduction to bayesian inference via variational approximations. *Political Analysis*, *19*, 32–47.

Imai, K., Lo, J., & Olmsted, J. (2016). Fast Estimation of Ideal Points with Massive Data. *American Political Science Review*, *110*, 631–656.

Linderman, S. W., Johnson, M. J., & Adams, R. P. (2015). Dependenfimat Multinomial Models Made Easy. In *Neural Information Processing Systems 2015*. Retrieved from https://hips.seas.harvard.edu/files/linderman-dependent-nips-2015.pdf

McFadden, D. & Train, K. E. (2000). Mixed mnl models for discrete response. *Journal of Applied Econometrics*, *15*, 447–470.

Nokken, T. P. & Poole, K. T. (2004). Congressional party defection in american history. *Legislative Studies Quarterly*, *29*, 545–568.

Rivers, D. (2003). Identification of Multidimensional Spatial Voting Models.

Train, K. E. (1998). Recreation demand models with taste differences over people. *Land Economics*, *74*, 230–239.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, *27*, 1413–1432.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594.