Supplementary Materials for "Hierarchical Item Response Models for Analyzing Public Opinion"

Xiang Zhou

October 23, 2018

A: Estimation and Inference

For notational simplicity, let us consider a simple random sample with no missing data. In practice, survey weights and nonresponses can be easily incorporated. In the current implementation, all item nonresponses are omitted from the level-I likelihood, meaning that they are treated as missing as random and can be predicted a posteriori.

First, let us define the following shorthands

$$\begin{aligned}
&\boldsymbol{\alpha} = \{\alpha_{jh};\, 1 \le j \le J,\, 0 \le h \le H_j - 1\}, \qquad \boldsymbol{\alpha_j} = \{\alpha_{jh};\, 0 \le h \le H_j - 1\},\\
&\boldsymbol{\beta} = \{\beta_{jh};\, 1 \le j \le J,\, 0 \le h \le H_j - 1\}, \qquad \boldsymbol{\beta_j} = \{\beta_{jh};\, 0 \le h \le H_j - 1\},\\
&\boldsymbol{\theta} = \{\theta_i;\, 1 \le i \le N\}, \quad \boldsymbol{x} = \{\boldsymbol{x_i};\, 1 \le i \le N\},\\
&\boldsymbol{y} = \{y_{ij};\, 1 \le i \le N, 1 \le j \le J\}, \qquad \boldsymbol{y_i} = \{y_{ij};\, 1 \le j \le J\}.
\end{aligned}$$

Since the covariates $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{z}}_i$ are treated as fixed quantities, I suppress them in most of the following derivation. Given equation (1) and the prior distribution (6), we can write the complete data likelihood as

$$\begin{aligned}
p(\boldsymbol{y}, \boldsymbol{\theta} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) &= p(\boldsymbol{y} | \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})\, p(\boldsymbol{\theta} | \boldsymbol{\gamma}, \boldsymbol{\lambda})\\
&= \prod_{i=1}^{N} \Big\{ \prod_{j=1}^{J} p(y_{ij} | \theta_i, \boldsymbol{\alpha}_j, \boldsymbol{\beta}_j) \Big\}\, p(\theta_i | \boldsymbol{\gamma}, \boldsymbol{\lambda}).
\end{aligned}$$

Suppose we now have a set of existing parameter estimates $\boldsymbol{\alpha}^*$, $\boldsymbol{\beta}^*$, $\boldsymbol{\gamma}^*$, $\boldsymbol{\lambda}^*$. Treating $\boldsymbol{\theta}$ as missing data, the Q-function of the EM algorithm, i.e., the conditional expectation of the log complete data

likelihood, is

$$
\begin{aligned}
Q(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \boldsymbol{\lambda}) &= \mathbb{E}\left[\log p(\boldsymbol{y}, \boldsymbol{\theta} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \boldsymbol{\lambda}) | \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma^*, \boldsymbol{\lambda}^*, \boldsymbol{y}\right] \\
&= \int_{\boldsymbol{\theta}} \Big\{ \sum_{i=1}^{N} \big[\sum_{j=1}^{J} \log p(y_{ij} | \theta_i, \boldsymbol{\alpha}_j, \boldsymbol{\beta}_j) + \log p(\theta_i | \gamma, \boldsymbol{\lambda}) \big] \Big\} \, p(\boldsymbol{\theta} | \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma^*, \boldsymbol{\lambda}^*, \boldsymbol{y}) d\boldsymbol{\theta} \\
&= \sum_{i=1}^{N} \int_{\theta_i} \big[\sum_{j=1}^{J} \log p(y_{ij} | \theta_i, \boldsymbol{\alpha}_j, \boldsymbol{\beta}_j) + \log p(\theta_i | \gamma, \boldsymbol{\lambda}) \big] \, p(\theta_i | \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma^*, \boldsymbol{\lambda}^*, \boldsymbol{y}_i) d\theta_i. \qquad (1)
\end{aligned}
$$

The latter equation holds because the posterior distribution of the ability parameters are independent across individuals:

$$
\begin{aligned}
p(\boldsymbol{\theta} | \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma^*, \boldsymbol{\lambda}^*, \boldsymbol{y}) &\propto p(y | \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\theta}) \, p(\boldsymbol{\theta} | \gamma, \boldsymbol{\lambda}) \\
&= \prod_{i=1}^{N} \prod_{j=1}^{J} p(y_{ij} | \boldsymbol{\alpha}_j^*, \boldsymbol{\beta}_j^*, \theta_i) \prod_{i=1}^{N} p(\theta_i | \gamma^*, \boldsymbol{\lambda}^*) \\
&= \prod_{i=1}^{N} \Big\{ \big[\prod_{j=1}^{J} p(y_{ij} | \boldsymbol{\alpha}_j^*, \boldsymbol{\beta}_j^*, \theta_i) \big] \, p(\theta_i | \gamma^*, \boldsymbol{\lambda}^*) \Big\} \\
&\propto \prod_{i=1}^{N} p(\theta_i | \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma^*, \boldsymbol{\lambda}^*, y_i).
\end{aligned}
$$

The unidimensional integrals in equation (1) can then be evaluated using quadrature methods. The basic idea is to select a number of nodes, say θ^k $(1 \leq k \leq K)$ that range from $-C$ to C, where C is a sufficiently large number such that $[-C, C]$ captures almost all of the mass of the posterior distribution $p(\theta_i | \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma^*, \boldsymbol{\lambda}^*, \boldsymbol{y}_i)$ for all individuals. In practice, if we impose the scale constraint $\sum_i \boldsymbol{\lambda}^T \tilde{\boldsymbol{z}}_i = 0$ such that the geometric average of estimated error variances $\hat{\sigma}_i^2$ equals one, setting $K = 25$ and $C = 5$ would be sufficient. Given a set of quadrature points θ^k and quadrature weights w_k, the final weights that enter the numerical evaluation of integral (1) will be

$$
w_{ik} = \frac{w_k \big[\prod_{j=1}^{J} p(y_{ij} | \boldsymbol{\alpha}_j^*, \boldsymbol{\beta}_j^*, \theta^k) \big] \, p(\theta^k | \gamma^*, \boldsymbol{x}_i, \boldsymbol{z}_i)}{\sum_{k=1}^{K} w_k \big[\prod_{j=1}^{J} p(y_{ij} | \boldsymbol{\alpha}_j^*, \boldsymbol{\beta}_j^*, \theta^k) \big] \, p(\theta^k | \gamma^*, \boldsymbol{x}_i, \boldsymbol{z}_i)}. \qquad (2)
$$

Thus equation (1) can be approximated as

$$\begin{aligned} Q(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) &\approx \sum_{i=1}^{N} \sum_{k=1}^{K} w_{ik} \big[\sum_{j=1}^{J} \log p(y_{ij} | \theta^k, \boldsymbol{\alpha}_j, \boldsymbol{\beta}_j) + \log p(\theta^k | \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{x}_i, \boldsymbol{z}_i) \big] \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{h=0}^{H_j - 1} w_{ik} \mathbf{1}(y_{ij} = h) \log P_{jh}(\theta^k) + \sum_{i=1}^{N} \sum_{k=1}^{K} w_{ik} \log p(\theta^k | \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{x}_i, \boldsymbol{z}_i) \\ &= \sum_{j=1}^{J} \big[\sum_{k=1}^{K} \sum_{h=0}^{H_j - 1} f_k^{jh} \log P_{jh}(\theta^k) \big] + \sum_{i=1}^{N} \sum_{k=1}^{K} w_{ik} \log p(\theta^k | \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{x}_i, \boldsymbol{z}_i), \end{aligned}$$

where $f_k^{jh} = \sum_{i=1}^{N} w_{ik} \mathbf{1}(y_{ij} = h)$ can be interpreted as the number of individuals around the preference level θ^k who choose category h for item j (given $\boldsymbol{\alpha}_j^*$ and $\boldsymbol{\beta}_j^*$). As a result, the M-step of the EM algorithm boils down to

$$\text{argmax}_{\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j} \sum_{k=1}^{K} \sum_{h=0}^{H_j - 1} f_k^{jh} \log P_{jh}(\theta^k) \text{ for all } j, \quad \text{and} \quad \text{argmax}_{\boldsymbol{\gamma}, \boldsymbol{\lambda}} \sum_{i=1}^{N} \sum_{k=1}^{K} w_{ik} \log p(\theta^k | \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{x}_i, \boldsymbol{z}_i).$$

It is not hard to show that the first optimization problem is equivalent to fitting J separate generalized linear models—one for each item—to the "pseudo data" f_k^{jh}. Specifically, binary logit (or probit) models are fitted for items with dichotomous responses, proportional odds models (or adjacent category logit models) for items with ordinal responses, and multinomial logit models for items with nominal responses. The second optimization problem is akin to the heteroscedastic regression model developed in Cook and Weisberg (1983), Aitkin (1987), and Verbyla (1993), except for the weights w_{ik} attached to the log likelihood $\log p(\theta^k | \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{x}_i, \boldsymbol{z}_i)$. To solve for $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$, we can employ the conditional maximization procedures outlined in Aitkin (1987) with a slight modification. The algorithm is detailed in Appendix B. Although both components of the M-step involve iterative procedures, they prove to be very fast in practice. For the first optimization, the generalized linear models are fitted to grouped data, where the number of observations equals the number of quadrature points (K) times the number of response categories (H_j) for the corresponding item. For the second optimization, the procedures described in Appendix B typically take few steps to converge. As a result, the runtime of the entire EM algorithm on a personal computer rarely exceeds a minute even for fairly large data sets (N=20,000-40,000; J=10-40).

Upon convergence of the EM algorithm, we obtain our final estimates $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\lambda}}$. We can then treat them as true parameters and conduct empirical Bayes inference of the latent preferences θ_i. For example, we can directly use the final posterior means, giving the expected a posterior

(EAP) estimates

$$\hat{\theta}_i = \mathbb{E}(\theta_i|\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\gamma}, \hat{\boldsymbol{\lambda}}, \boldsymbol{y}) = \sum_{k=1}^{K} w_{ik}\theta^k. \tag{3}$$

Finally, to conduct inference for the key parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, γ and $\boldsymbol{\lambda}$, we can calculate the asymptotic variance-covariance matrix $\hat{I}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \boldsymbol{\lambda})$ using either the Hessian matrix or the outer product of gradients of the log marginal likelihood. The latter approach is illustrated in Appendix C.

B: The M-step for Updating γ and $\boldsymbol{\lambda}$

To update γ and $\boldsymbol{\lambda}$, we first note that the objective function can be written as

$$\begin{aligned} f(\gamma, \boldsymbol{\lambda}) &= \sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik} \log p(\theta^k|\gamma, \boldsymbol{\lambda}, \boldsymbol{x}_i, \boldsymbol{z}_i) \\ &= -\frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{K}[w_{ik}\log\, 2\pi + w_{ik}\boldsymbol{\lambda}^T\tilde{\boldsymbol{z}}_i + \frac{w_{ik}(\theta_k - \gamma^T\tilde{\boldsymbol{x}}_i)^2}{\exp(\boldsymbol{\lambda}^T\tilde{\boldsymbol{z}}_i)}] \\ &= -\frac{1}{2}\sum_{i=1}^{N}[\log\, 2\pi + \boldsymbol{\lambda}^T\tilde{\boldsymbol{z}}_i + \frac{(\tilde{\theta}_i - \gamma^T\boldsymbol{x}_i)^2 + \tilde{\sigma^2_{\theta_i}}}{\exp(\boldsymbol{\lambda}^T\tilde{\boldsymbol{z}}_i)}], \end{aligned}$$

where $\tilde{\theta}_i = \sum_{k=1}^{K} w_{ik}\theta^k$ is the working posterior mean of θ_i and $\tilde{\sigma^2_{\theta_i}} = \sum_{k=1}^{K} w_{ik}(\theta^k)^2 - \tilde{\theta}_i^2$ is the working posterior variance of θ_i (given $\boldsymbol{\alpha}^*$, $\boldsymbol{\beta}^*$, γ^*, $\boldsymbol{\lambda}^*$). Thus we can maximize $f(\gamma, \boldsymbol{\lambda})$ iteratively:

1. Fit a simple least squares of $\tilde{\theta}_i$ on $\boldsymbol{x}_i$, saving the residuals r_i,
2. Fit a gamma regression with a log link of $r_i^2 + \tilde{\sigma^2_{\theta_i}}$ on z_i, saving the fitted values $s_i^2 = \exp(\hat{\boldsymbol{\lambda}}^T\tilde{\boldsymbol{z}}_i)$,
3. Fit a weighted least squares of $\tilde{\theta}_i$ on $\boldsymbol{x}_i$ with weights $1/s_i^2$, updating the the residuals r_i,
4. Iterate steps 2 and 3 until convergence, updating γ^* and $\boldsymbol{\lambda}^*$.

C: Asymptotic Inference for Hierarchical IRT Models

To construct the observed information matrix, we use the outer product of gradients of the log marginal likelihood. For individual i, the log marginal likelihood can be numerically evaluated as

$$\log L_i \approx \log \sum_{k=1}^{K} L_{ik} p_{ik} w_k,$$

where

$$L_{ik} = \prod_{j=1}^{J} p(y_{ij}|\theta^k, \boldsymbol{\alpha_j}, \beta_j)$$
$$p_{ik} = [2\pi \exp(\boldsymbol{\lambda}^T \boldsymbol{z_i})]^{-\frac{1}{2}} \exp[-\frac{(\theta^k - \boldsymbol{\gamma}^T x_i)^2}{2\exp(\boldsymbol{\lambda}^T \boldsymbol{z_i})}],$$

and w_k are quadrature weights associated with θ^k. Given the above expression, we can derive the score function for each of the level I models presented in the paper. For instance, for the graded response model (3), we can show that

$$\frac{\partial \log L_i}{\partial \alpha_{jh}} = \frac{\sum_{k=1}^{K} w_k p_{ik} L_{ik}^{-j}}{L_i} \begin{cases} \frac{\exp(\alpha_{jh}+\beta_j\theta^k)}{[1+\exp(\alpha_{jh}+\beta_j\theta^k)]^2}, & \text{if } h = y_{ij} \geq 1 \\ -\frac{\exp(\alpha_{jh}+\beta_j\theta^k)}{[1+\exp(\alpha_{jh}+\beta_j\theta^k)]^2}, & \text{if } h = y_{ij}+1 \leq H_j - 1 \\ 0, \quad \text{otherwise} \end{cases}$$

where $L_{ik}^{-j} = \prod_{l \neq j}^{J} p(y_{ij}|\theta^k, \boldsymbol{\alpha_j}, \beta_j)$. Similarly, by taking the partial derivatives with respect to β_j, γ_p, λ_q, we obtain

$$\frac{\partial \log L_i}{\partial \beta_j} = \frac{1}{L_i} \sum_{k=1}^{K} w_k p_{ik} L_{ik}^{-j} \theta^k \{ \frac{\exp(\alpha_{jy_{ij}} + \beta_j\theta^k)}{[1+\exp(\alpha_{jy_{ij}} + \beta_j\theta^k)]^2} - \frac{\exp(\alpha_{j\,y_{ij}+1} + \beta_j\theta^k)}{[1+\exp(\alpha_{j\,y_{ij}+1} + \beta_j\theta^k)]^2} \}$$
$$\frac{\partial \log L_i}{\partial \gamma_p} = \frac{1}{L_i} \sum_{k=1}^{K} w_k p_{ik} L_{ik} \exp(-\boldsymbol{\lambda}^T \boldsymbol{z_i})(\theta_k - \boldsymbol{\gamma}^T \boldsymbol{x_i}) x_{ip}$$
$$\frac{\partial \log L_i}{\partial \lambda_q} = \frac{1}{2L_i} \sum_{k=1}^{K} w_k p_{ik} L_{ik} [\exp(-\boldsymbol{\lambda}^T \boldsymbol{z_i})(\theta_k - \boldsymbol{\gamma}^T \boldsymbol{x_i})^2 - 1] z_{iq}.$$

We then concatenate all these terms to form the score vector $\nabla \log L_i$ and construct the asymptotic variance-covariance matrix of parameter estimates as

$$\hat{V}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\lambda}}) = \hat{I}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\lambda}})^{-1}$$
$$= [\sum_{i=1}^{N} \nabla \log L_i (\nabla \log L_i)^T]^{-1}.$$

Note that in constructing the score vector, we must discard one component of $\boldsymbol{\gamma}$ and one component of $\boldsymbol{\lambda}$ to avoid a singular information matrix (due to the identification constraints). In practice, we can discard $\frac{\partial \log L_i}{\partial \gamma_0}$ and $\frac{\partial \log L_i}{\partial \lambda_0}$ as the intercepts are usually the least substantively interesting parameters.

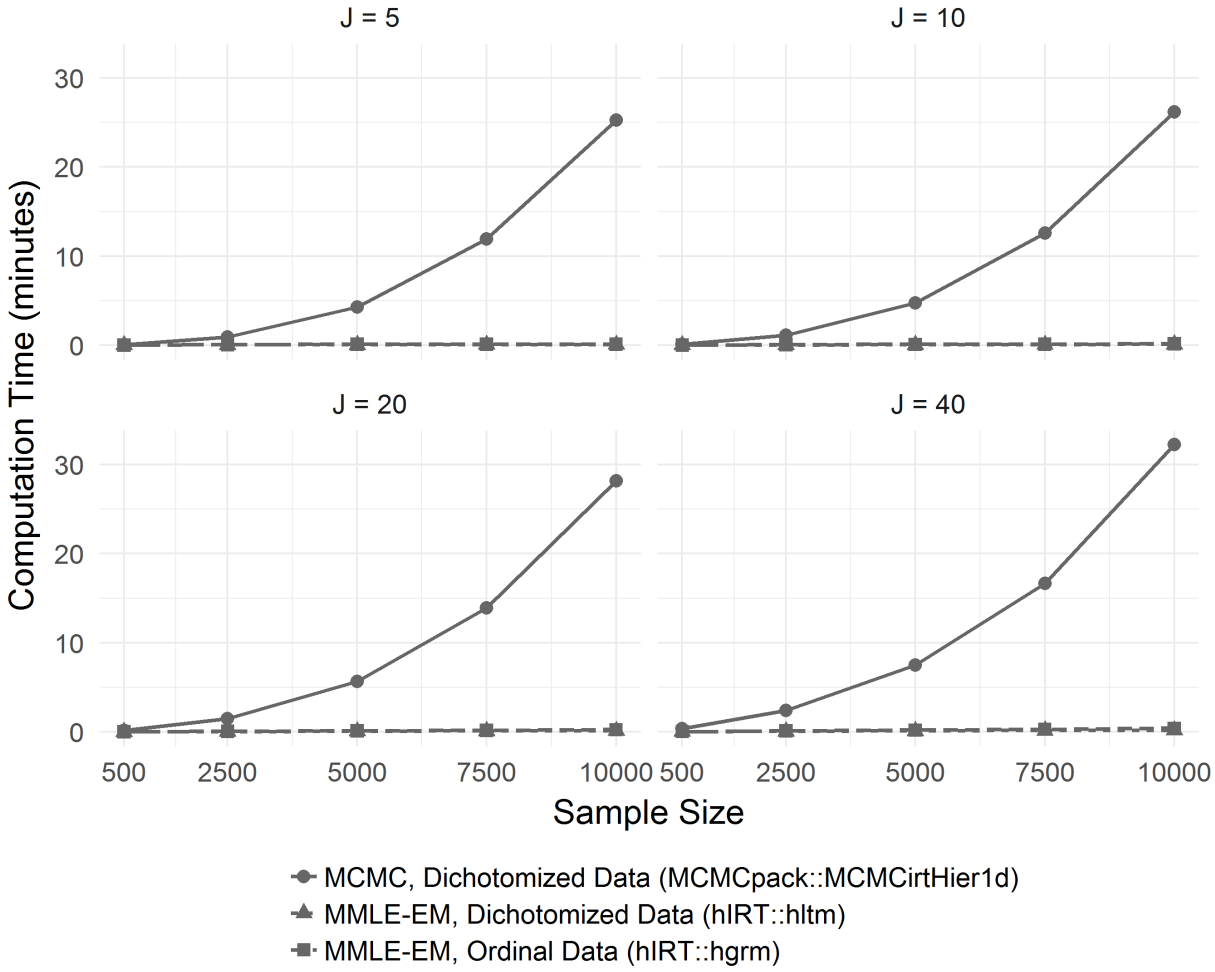
D: EM Algorithm versus MCMC Simulation in Computation Time

This appendix provides a brief yet systematic comparison in computation time between the EM algorithm and a full Bayesian approach for fitting hierarchical IRT models. The latter has been implemented in the R function `MCMCpack::MCMCirtHier1d` for the simplest case—binary response data with homoscedastic preferences (Martin, Quinn and Park 2011). I use the same data generating process as in my Monte Carlo study. When applying `MCMCpack::MCMCirtHier1d`, I dichotomize the response data using their sample means as cutoff points. To facilitate comparison, I run the EM algorithm both for the dichotomized data and for the ordinal data, using `hIRT::hltm` and `hIRT::hgrm` respectively. To illustrate the scalability of different methods, I vary the number of respondents N from 500 to 10,000 and the number of items from 5 to 40. The results are shown in Figure A1, where the horizontal axis denotes sample size N, the vertical axis denotes computation time (in minutes), and different algorithms are represented by different point shapes and line types. It is easy to see that the EM algorithm is extremely fast for all combinations of N and J, whereas the full Bayesian implementation is not only much slower but also much less scalable as the number of respondents grows.

E: Estimates of Item Discrimination Parameters

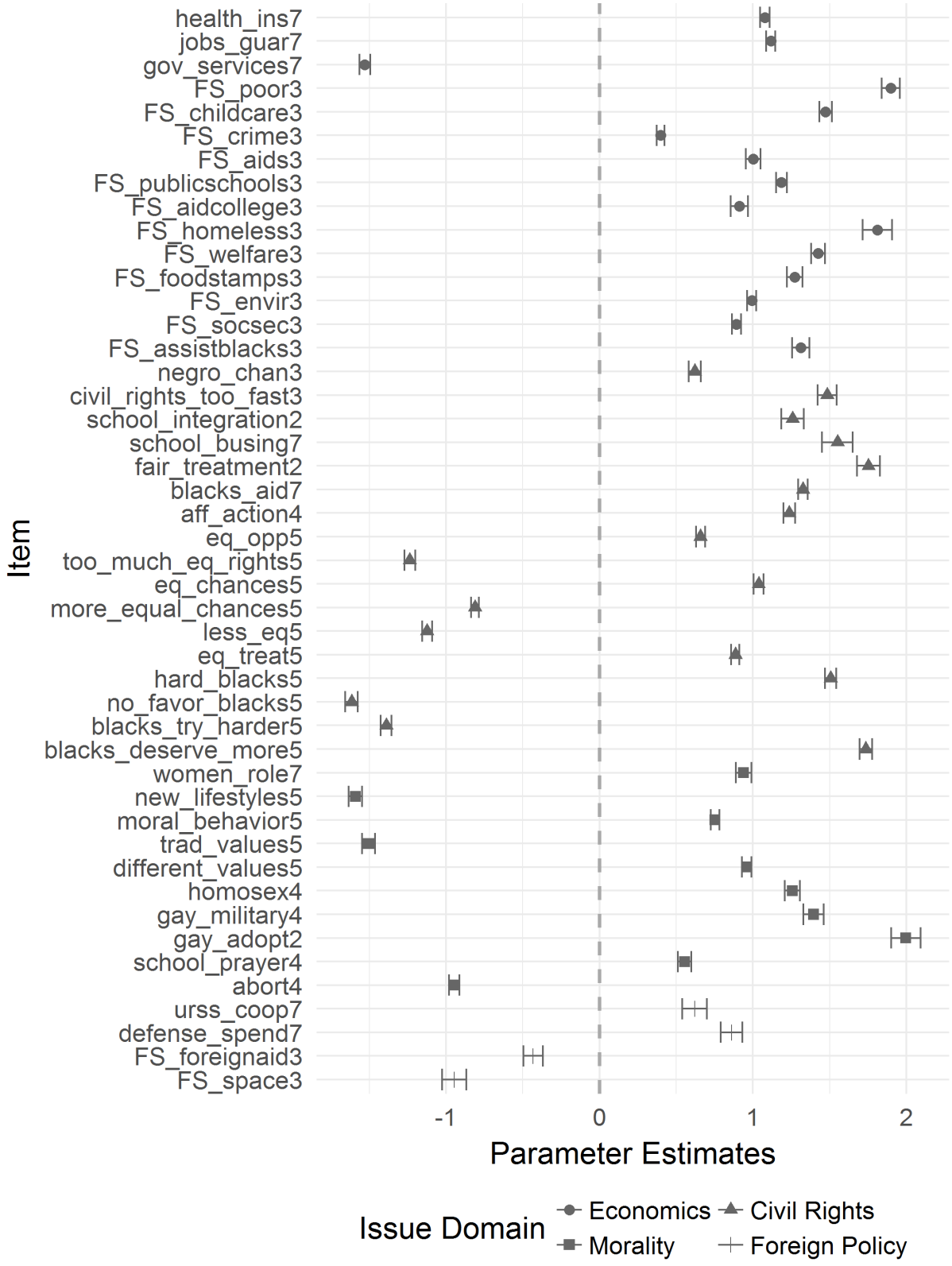
Figure A2 shows the estimates of the item discrimination parameters, along with their 95% asymptotic confidence intervals, for the party polarization example. We can see that the discrimination parameter estimates vary greatly across items, although all of them are statistically significantly different from zero.

Figure S1: EM Algorithm versus MCMC Simulation in Computation Time.



Note: MCMC = Markov Chain Monte Carlo; MMLE = Marginal Maximum Likelihood Estimation; EM = Expectation-Maximization.

Figure S2: Estimates of Item Discrimination Parameters for the Party Polarization Example.



Note: Error bars represent 95% asymptotic confidence intervals.

References

Aitkin, Murray. 1987. "Modelling Variance Heterogeneity in Normal Regression Using GLIM." *Applied Statistics* 36(3):332–339.

Cook, R Dennis and Sanford Weisberg. 1983. "Diagnostics for Heteroscedasticity in Regression." *Biometrika* 70(1):1–10.

Martin, Andrew D, Kevin M Quinn and Jong Hee Park. 2011. "MCMCpack: Markov Chain Monte Carlo in R." *Journal of Statistical Software* 42(9):1–21.

Verbyla, Arunas Petras. 1993. "Modelling Variance Heterogeneity: Residual Maximum Likelihood and Diagnostics." *Journal of the Royal Statistical Society. Series B (Methodological)* 55(2):493–508.