Fast Estimation of Ideal Points with Massive Data*

Kosuke Imai† James Lo‡ Jonathan Olmsted§

First Draft: December 22, 2014
This Draft: March 9, 2015

Abstract

Recently, many scholars have begun to estimate ideological preferences across time and institutions, analyzing data sets that are orders of magnitude larger than a canonical single-chamber roll call matrix for a single time period. To overcome the resulting computational challenges, we propose fast estimation methods for ideal points with massive data. Specifically, we derive the Expectation-Maximization (EM) algorithms to estimate the standard ideal point model with binary, ordinal, and continuous outcome variables. We then extend this methodology to dynamic and hierarchical ideal point models by developing variational EM algorithms for approximate inference. We demonstrate the computational efficiency and scalability of our methodology through a variety of real and simulated data. In cases where a standard Markov chain Monte Carlo algorithm would require several days to compute ideal points, the proposed algorithm can produce essentially identical estimates within minutes. Open-source software is available for implementing the proposed methods.

Key words: Expectation-Maximization algorithm, factor analysis, item response theory, scaling, variational inference

*We thank Simon Jackman, Will Lowe, Michael Peress, and Marc Ratkovic for their helpful discussions, Kevin Quinn for providing a replication data set and code, and Yuki Shiraito for his assistance with the Japanese survey data. The proposed methods are implemented through open-source software `emIRT`, which will be made freely available as an R package at the Comprehensive R Archive Network (CRAN; http://cran.r-project.org/package=emIRT).

†Professor, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 609–258–6601, Email: kimai@princeton.edu, URL: http://imai.princeton.edu

‡Postdoctoral Research Associate, Department of Politics, Princeton University, Princeton NJ 08544. Email: jameslo@princeton.edu

§Specialist, NPD Group, Port Washington NY 11050.

1 Introduction

Ever since the seminal work of Poole and Rosenthal (1991, 1997), a number of scholars have used spatial voting models to estimate ideological preferences of legislators, justices, and other political actors from roll call votes and other data in a variety of settings (e.g., Voeten, 2000; Morgenstern, 2004; Bailey, Kamoie and Maltzman, 2005; McCarty, Poole and Rosenthal, 2006; Londregan, 2007; Spirling and McLean, 2007; Clinton and Lewis, 2008; Ho and Quinn, 2010; Bonica, 2014). These and other substantive applications are made possible by numerous methodological advancements including Bayesian estimation (Jackman, 2001; Clinton, Jackman and Rivers, 2004), optimal classification (Poole, 2000), dynamic modeling (Martin and Quinn, 2002), and models with agenda setting or strategic voting (Londregan, 1999; Clinton and Meirowitz, 2003).

Recently, with the increasing availability of data and methodological sophistication, researchers have turned their attention to the estimation of ideological preferences that are comparable across time and institutions. For example, Bailey (2007) measure ideal points of US presidents, senators, representatives, and Supreme Court justices on the same scale over time (see also Bailey and Chang, 2001; Bailey, 2013). Similarly, Shor and McCarty (2011) compute the ideal points of the state legislators from all US states and compare them with the members of Congress (see also Shor, Berry and McCarty, 2011; Battista, Peress and Richman, 2013). Finally, Bafumi and Herron (2010) estimate the ideological positions of voters and their members of Congress in order to study representation while Clinton et al. (2012) compare the ideal points of agencies with those of presidents and Congressional members.

These efforts to estimate ideological preferences across time and institutions often face a computational challenge of dealing with data sets that are orders of magnitude larger than the canonical single-chamber roll call matrix for a single time period. Indeed, as Table 1 shows, the past decade has witnessed a significant rise in the use of large data sets for ideal point estimation. While most of the aforementioned works are based on the Bayesian models of ideal points, standard Markov chain Monte Carlo (MCMC) algorithms can be prohibitively slow when applied to large data sets. As a result, researchers are often unable to estimate their models using the entire data and are forced to make various shortcuts and compromises. For example, Shor and McCarty (2011) fit their model in multiple steps using subsets of the data whereas Bailey (2007) resorts to a simpler parametric dynamic model in order to reduce computational costs (p. 441) (see also Bailey, 2013). Since a massive data set implies a large number of parameters under these models, the convergence of MCMC algorithms also becomes difficult to assess. Bafumi and Herron (2010), for example, express

	Number of subjects	Number of items	Data types
DW-NOMINATE scores (1789 – 2012)	37,511	46,379	roll calls
Common Space scores (1789 - 2012)	11,833	90,609	roll calls
Martin and Quinn scores (1937 – 2013)	697	5,164	votes
Gerber and Lewis (2004)	2.8 million	12	votes
Bailey (2007)	27,795	2,750	roll calls & votes
Bafumi and Herron (2010)	8,848	4,391	survey & roll calls
Shor and McCarty (2011)	6,201	5,747	survey & roll calls
Tausanovitch and Warshaw (2013)	275,000	311	survey
Peress (2013)	700	16,000	co-sponsorship & roll calls
Bonica (2014)	4.2 million	78,363	contribution

Table 1: Recent Applications of Ideal Point Models to Large Data Sets. The past decade has witnessed a significant rise in the use of large data sets for ideal point estimation. Note that "# of subjects" should be interpreted as the number of ideal points to be estimated. For example, if a legislator serves for two terms and are allowed to have different ideal points in those terms, then this legislator is counted as two subjects.

a concern about the convergence of ideal points for voters (footnote 24).

In addition, estimating ideal points over a long period of time often imposes a significant computational burden. Indeed, the use of computational resources at supercomputer centers has been critical to the development of various NOMINATE scores.[1] Similarly, estimation of the Martin and Quinn (2002) ideal point estimates for US supreme court justices over 47 years took over 5 days to estimate. This suggests that while these ideal point models are attractive, they are often practically unusable for many researchers who wish to analyze a large-scale data set.

In this paper, we propose a fast estimation method for ideal points with massive data. Specifically, we develop the Expectation-Maximization (EM) algorithms (Dempster, Laird and Rubin, 1977) that either exactly or approximately maximizes the posterior distribution under various ideal point models. The main advantage of EM algorithms is that they can dramatically reduce computational time. Through a number of empirical and simulation examples, we demonstrate that in cases where a standard MCMC algorithm would require several days to compute ideal points, the proposed algorithm can produce essentially identical estimates within minutes. The EM algorithms also scale much better than other existing ideal point estimation algorithms. They can estimate an extremely large number of ideal points on a laptop within a few hours whereas current methodologies would require the level of computational resources only available at a supercomputer center to do the same

[1] The voteview website notes that the DW-NOMINATE and Common Space DW-NOMINATE scores are computed using the Rice terascale cluster. See `http://voteview.com/dwnominate.asp` and `http://voteview.com/dwnomjoint.asp` (accessed on November 10, 2014).

computation.

We begin by deriving the EM algorithm for the standard Bayesian ideal point model of Clinton, Jackman and Rivers (2004) in Section 2. We show that the proposed algorithm produces ideal point estimates which are essentially identical to those from other existing methods. We then extend our approach to other popular ideal point models that have been developed in the literature. Specifically, we develop an EM algorithm for the model with mixed ordinal and continuous outcomes in Section 3 (Quinn, 2004) by applying a certain transformation to the original parametrization. We also develop an EM algorithm for the dynamic model in Section 4 (Martin and Quinn, 2002) and the hierarchical model in Section 5 (Bafumi et al., 2005).

For dynamic and hierarchical models, an EM algorithm that directly maximizes the the posterior distribution is not available in a closed-form. Therefore, we rely on variational Bayesian inference, which is a popular machine learning methodology for fast and approximate Bayesian estimation (see Wainwright and Jordan (2008) for a review and Grimmer (2011) for an introductory article in political science). For each case, we demonstrate the computational efficiency and scalability of the proposed methodology by applying it to a wide range of real and simulated data sets. Our proposed algorithms complement a recent application of variational inference to combine ideal point estimation with topic models (Gerrish and Blei, 2012). We implement the proposed algorithms via an open-source R package, `emIRT` (Imai, Lo and Olmsted, 2015), so that others can apply them to their own research.

In the item response theory literature, the EM algorithm is used to maximize the marginal likelihood function where ability parameters, i.e., ideal point parameters in the current context, are integrated out (Bock and Aitkin, 1981). In the ideal point literature, Bailey (2007) and Bailey and Chang (2001) use variants of the EM algorithm in their model estimation. The M-steps of these existing algorithms, however, do not have a closed-form solution. In this paper, we derive closed-form EM algorithms for popular Bayesian ideal point models. This leads to faster and more reliable estimation algorithms.

Finally, the important and well-known drawback of these EM algorithms is that they do not produce uncertainty estimates such as standard errors. In contrast, the MCMC algorithms are designed to fully characterize the posterior, enabling the computation of uncertainty measures for virtually any quantities of interest. Moreover, the standard errors based on variational posterior are often too small, underestimating the degree of uncertainty. While many applied researchers tend to ignore estimation uncertainty associated with ideal points, such a practice can yield misleading

inference. To address this problem, we apply the parametric bootstrap approach of Lewis and Poole (2004) (see also Carroll, Lewis, Lo and Poole, 2009). Although this obviously increases the computational cost of the proposed approach, the proposed EM algorithms still scale much better than the existing alternatives. Furthermore, researchers can reduce this computational cost by a parallel implementation of bootstrap on a distributed system. We note that since our models are Bayesian, it is rather unconventional to utilize bootstrap, which is a frequentist procedure. However, one can interpret the resulting confidence intervals as a measure of uncertainty of our Bayesian estimates over repeated sampling under the assumed model.

2 Standard Ideal Point Model

We begin by deriving the EM algorithm for the standard ideal point model of Clinton, Jackman and Rivers (2004). In this case, the proposed EM algorithm maximizes the posterior distribution without approximation. We illustrate the computational efficiency and scalability of our proposed algorithm by applying it to roll call votes in recent US Congress as well as simulated data.

2.1 The Model

Suppose that we have N legislators and J roll calls. Let y_{ij} denote the vote of legislator i on roll call j where $y_{ij} = 1$ ($y_{ij} = 0$) implies that the vote is in the affirmative (negative) with $i = 1, \ldots, N$ and $j = 1, \ldots, J$. Abstentions, if present, are assumed to be ignorable such that these votes are missing at random and can be predicted from the model using observed data (see Rosas and Shomer, 2008). Furthermore, let $\mathbf{x}_i$ represent the K-dimensional column vector of ideal point for legislator i. Then, if we use y^*_{ij} to represent a latent propensity to cast a "yea" vote where $y_{ij} = \mathbf{1}\{y^* > 0\}$, the standard K-dimensional ideal point model is given by,

$$y^*_{ij} \;=\; \alpha_j + \mathbf{x}_i^\top \boldsymbol{\beta}_j + \epsilon_{ij} \tag{1}$$

where $\boldsymbol{\beta}_j$ is the K-dimensional column vector of item discrimination parameters and α_j is the scalar item difficulty parameter. Finally, ϵ_{ij} is an independently, identically distributed random utility and is assumed to follow the standard Normal distribution.

For notational simplicity, we use $\tilde{\boldsymbol{\beta}}_j^\top = (\alpha_j, \boldsymbol{\beta}_j^\top)$ and $\tilde{\mathbf{x}}_i^\top = (1, \mathbf{x}_i^\top)$ so that equation (1) can be more compactly written as

$$y^*_{ij} \;=\; \tilde{\mathbf{x}}_i^\top \tilde{\boldsymbol{\beta}}_j + \epsilon_{ij}. \tag{2}$$

Following the original article, we place independent and conjugate prior distributions on $\mathbf{x}_i$ and $\tilde{\boldsymbol{\beta}}_j$,

separately. Specifically, we use

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N) \;=\; \prod_{i=1}^{N} \phi_K\left(\mathbf{x}_i; \boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}}\right) \quad \text{and} \quad p(\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_J) \;=\; \prod_{j=1}^{J} \phi_{K+1}\left(\tilde{\boldsymbol{\beta}}_j; \boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}\right) \tag{3}$$

where $\phi_k(\cdot;\cdot)$ is the density of a k-variate Normal random variable, $\boldsymbol{\mu}_{\mathbf{x}}$ and $\boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}}$ represent the prior mean vectors, and $\boldsymbol{\Sigma}_{\mathbf{x}}$ and $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}$ are the prior covariance matrices.

Given this model, the joint posterior distribution of $(\mathbf{Y}^*, \{\mathbf{x}_i\}_{i=1}^N, \{\tilde{\boldsymbol{\beta}}_j\}_{j=1}^J)$ conditional on the roll call matrix $\mathbf{Y}$ is given by,

$$\begin{aligned}
& p(\mathbf{Y}^*, \{\mathbf{x}_i\}_{i=1}^N, \{\tilde{\boldsymbol{\beta}}_j\}_{j=1}^J \mid \mathbf{Y}) \\
\propto & \prod_{i=1}^{N}\prod_{j=1}^{J} \left(\mathbf{1}\{y_{ij}^* > 0\}\mathbf{1}\{y_{ij} = 1\} + \mathbf{1}\{y_{ij}^* \leq 0\}\mathbf{1}\{y_{ij} = 0\}\right)\; \phi_1\left(y_{ij}^*; \tilde{\mathbf{x}}_i^\top \tilde{\boldsymbol{\beta}}_j, 1\right) \\
& \times \prod_{i=1}^{N} \phi_K\left(\mathbf{x}_i; \boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}}\right) \prod_{j=1}^{J} \phi_{K+1}\left(\tilde{\boldsymbol{\beta}}_j; \boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}\right)
\end{aligned} \tag{4}$$

where $\mathbf{Y}$ and $\mathbf{Y}^*$ are matrices whose element in the ith row and jth column is y_{ij} and y_{ij}^*, respectively. Clinton, Jackman and Rivers (2004) describe the MCMC algorithm to sample from this joint posterior distribution and implement it as the `ideal()` function in the open-source R package `pscl` (Jackman, 2012).

2.2 The Proposed Algorithm

We derive the EM algorithm that maximizes the posterior distribution given in equation (4) without approximation. The proposed algorithm views $\{\mathbf{x}_i\}_{i=1}^N$ and $\{\tilde{\boldsymbol{\beta}}_j\}_{j=1}^J$ as parameters and treats $\mathbf{Y}^*$ as missing data. Specifically, at the tth iteration, denote the current parameter values as $\{\mathbf{x}_i^{(t-1)}\}_{i=1}^N$ and $\{\tilde{\boldsymbol{\beta}}_j^{(t-1)}\}_{j=1}^J$. Then, the E-step is given by the following so-called "Q-function," which represents the expectation of the log joint posterior distribution,

$$\begin{aligned}
& Q(\{\mathbf{x}_i\}_{i=1}^N, \{\tilde{\boldsymbol{\beta}}_j\}_{j=1}^J) \\
= \;& \mathbb{E}\left[\log p(\mathbf{Y}^*, \{\mathbf{x}_i\}_{i=1}^N, \{\tilde{\boldsymbol{\beta}}_j\}_{j=1}^J \mid \mathbf{Y}) \mid \mathbf{Y}, \{\mathbf{x}_i^{(t-1)}\}_{i=1}^N, \{\tilde{\boldsymbol{\beta}}_j^{(t-1)}\}_{j=1}^J\right] \\
= \;& \sum_{i=1}^{N}\sum_{j=1}^{J} \mathbb{E}\left\{\log \phi_1\left(y_{ij}^*; \tilde{\mathbf{x}}_i^\top \tilde{\boldsymbol{\beta}}_j, 1\right)\right\} + \sum_{i=1}^{N} \log \phi_K\left(\mathbf{x}_i; \boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}}\right) + \sum_{j=1}^{J} \log \phi_{K+1}\left(\tilde{\boldsymbol{\beta}}_j; \boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}\right) + \text{const.} \\
= \;& -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{J}\left(\tilde{\boldsymbol{\beta}}_j^\top \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top \tilde{\boldsymbol{\beta}}_j - 2\tilde{\boldsymbol{\beta}}_j^\top \tilde{\mathbf{x}}_i y_{ij}^{*(t)}\right) - \frac{1}{2}\sum_{i=1}^{N}\left(\mathbf{x}_i^\top \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{x}_i - 2\mathbf{x}_i^\top \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\mu}_{\mathbf{x}}\right) \\
& -\frac{1}{2}\sum_{j=1}^{J}\left(\tilde{\boldsymbol{\beta}}_j^\top \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}^{-1}\tilde{\boldsymbol{\beta}}_j - 2\tilde{\boldsymbol{\beta}}_j^\top \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}^{-1}\boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}}\right) + \text{const.}
\end{aligned} \tag{5}$$

where

$$y_{ij}^{*}{}^{(t)} \;=\; \mathbb{E}(y_{ij}^{*} \mid \mathbf{x}_i^{(t-1)}, \tilde{\boldsymbol{\beta}}_j^{(t-1)}, y_{ij}) \;=\; \begin{cases} m_{ij}^{(t-1)} + \frac{\phi\left(m_{ij}^{(t-1)}\right)}{\Phi\left(m_{ij}^{(t-1)}\right)} & \text{if } y_{ij} = 1 \\ m_{ij}^{(t-1)} - \frac{\phi\left(m_{ij}^{(t-1)}\right)}{1-\Phi\left(m_{ij}^{(t-1)}\right)} & \text{if } y_{ij} = 0 \\ m_{ij}^{(t-1)} & \text{if } y_{ij} \text{ is missing} \end{cases} \tag{6}$$

with $m_{ij}^{(t-1)} = (\tilde{\mathbf{x}}_i^{(t-1)})^{\top} \tilde{\boldsymbol{\beta}}_j^{(t-1)}$.

Straightforward calculation shows that the maximization of this Q-function, i.e., the M-step, can be achieved via the following two conditional maximization steps,

$$\mathbf{x}_i^{(t)} \;=\; \left(\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} + \sum_{j=1}^{J} \boldsymbol{\beta}_j^{(t-1)} \boldsymbol{\beta}_j^{(t-1)\top} \right)^{-1} \left(\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \boldsymbol{\mu}_{\mathbf{x}} + \sum_{j=1}^{J} \boldsymbol{\beta}_j^{(t-1)} (y_{ij}^{*}{}^{(t)} - \alpha_j^{(t-1)}) \right) \tag{7}$$

$$\tilde{\boldsymbol{\beta}}_j^{(t)} \;=\; \left(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}^{-1} + \sum_{i=1}^{N} \tilde{\mathbf{x}}_i^{(t)} \left(\tilde{\mathbf{x}}_i^{(t)} \right)^{\top} \right)^{-1} \left(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}^{-1} \boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}} + \sum_{i=1}^{N} \tilde{\mathbf{x}}_i^{(t)} y_{ij}^{*}{}^{(t)} \right) \tag{8}$$

The algorithm repeats these E and M steps until convergence. Given that the model is identified up to an affine transformation, we use the correlation-based convergence criteria where the algorithm terminates when the correlation between the previous and current values of all parameters reaches a pre-specified threshold.

Finally, to compute uncertainty estimates, we apply the parametric bootstrap (Lewis and Poole, 2004). Specifically, we first estimate ideal points and bill parameters via the proposed EM algorithm. Using these estimates, we calculate the choice probabilities associated with each outcome. Then, we randomly generate roll-call matrices given these estimated outcome probabilities. Where there are missing votes, we simply induce the same missingness patterns. This is repeated a sufficiently large number of times and the resulting bootstrap replicates for each parameter are used to characterize estimation uncertainty.

2.3 An Empirical Application

To assess its empirical performance, we apply the proposed EM algorithm to roll-call voting data for both the Senate and the House of Representatives for sessions of Congress 102 through 112. Specifically, we compare the ideal point estimates and their computation time from the proposed algorithm to those from three other methods; the MCMC algorithm implemented as `ideal()` in the R package **pscl** (Jackman, 2012), the alternating Maximum Likelihood estimator implemented as `wnominate()` in the R package **wnominate** (Poole et al., 2011), and the non-parametric optimal classification estimator implemented as `oc()` in the R package **oc** (Poole et al., 2012). For all

roll-call matrices, we restrict attention to just those legislators with at least 25 observed votes on non-unanimous bills. In all cases, we assume a single spatial dimension.

We caution that the comparison of computational efficiency presented here is necessarily illustrative. The performance of any numerical algorithm may depend on starting values, and no absolute convergence criteria exists for any of the algorithms we examine. For the MCMC algorithm, we run the chain for 100,000 iterations beyond a burn-in period of 20,000 iterations. Inference is based on a thinned chain where we keep every 100 draws. While the length of chain and its thinning interval are within the recommended range by Clinton, Jackman and Rivers (2004), we emphasize that any convergence criteria used for deciding when to terminate the MCMC algorithm is somewhat arbitrary. The default diffuse priors, specified in `ideal()` from the R package **pscl**, are used and propensities for missing votes are not imputed during the data-augmentation step. In particular, we assume a single dimensional Normal prior for all ideal point parameters with a mean of 0 and a variance of 1. For the bill parameters, we assume a two dimensional Normal prior with a mean vector of 0's and a covariance matrix with each variance term equal to 25 and no covariance. The standard normalization of ideal points, i.e., a mean of 0 and a standard deviation of 1, is used for local identification.

For the proposed EM algorithm, we use random starting values for the ideal point and bill parameters. The same prior distributions as the MCMC algorithm are used for all parameters. We terminate the EM algorithm when each block of parameters has a correlation with the values from the previous iteration larger than $1 - p$. With one spatial dimension, we have three parameter blocks: the bill difficulty parameters, the bill discrimination parameters and the ideal point parameters. Following Poole and Rosenthal (1997) (see p. 237), we use $p = 10^{-2}$. We also consider a far more stringent criterion where $p = 10^{-6}$, requiring parameters to correlate greater than 0.999999. In the following results, we focus on the latter "high precision" variant, except in the case of computational performance where results for both criteria are presented (labeled "EM" and "EM (high precision)", respectively). The results from the EM algorithm do not include measures of uncertainty. For this, we include the "EM with Bootstrap" variant, which uses 100 parametric bootstrap replicates.

Finally, for W-NOMINATE, we do not include any additional bootstrap trials for characterizing uncertainty about the parameters, so the results will only include point estimates. For optimal classification, we use the default setting of `oc()` in the R package **oc**. Because Bayesian the MCMC algorithm is stochastic and the EM algorithm has random starting values, we run each estimator 50 times for any given roll-call matrix and report the median of each performance measurement.

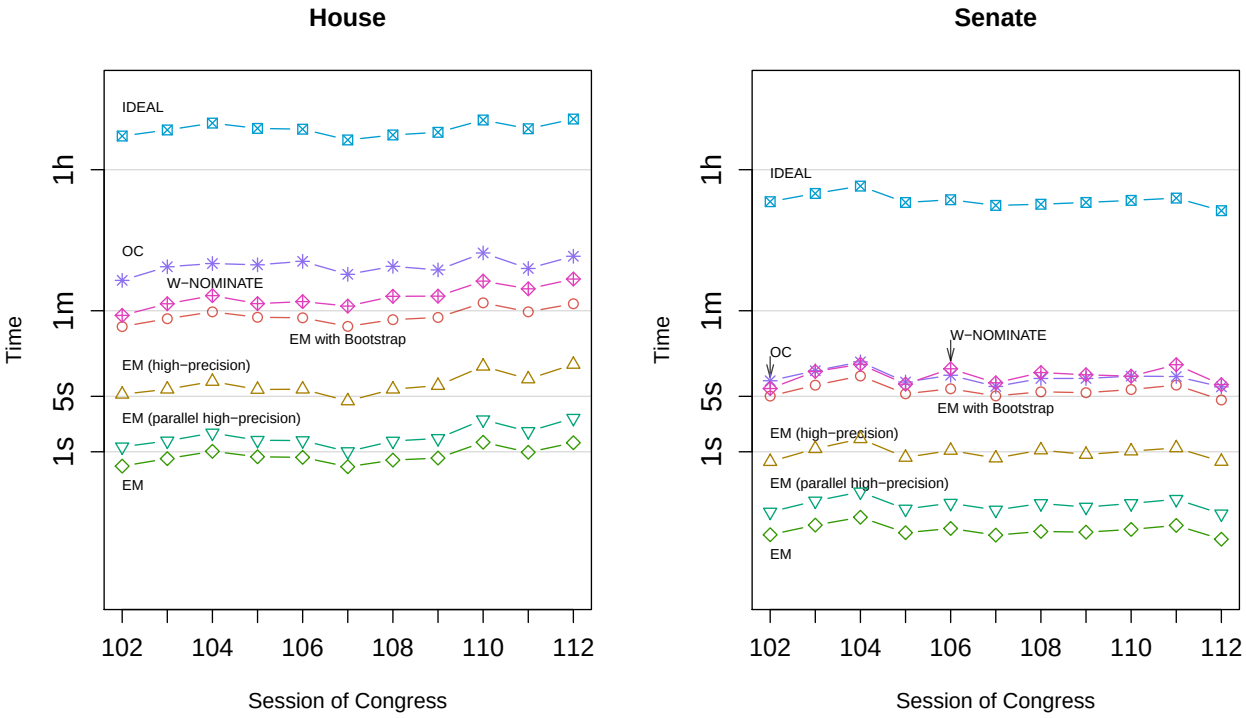

Figure 1: Comparison of Computational Performance across the Methods. Each point represents the length of time required to compute estimates where the spacing of time on the vertical axis is based on the log-scale. The proposed EM algorithm, indicated by 'EM', 'EM (high precision)', 'EM (parallel high precision)', and 'EM with Bootstrap' is compared with 'W-NOMINATE' (Poole et al., 2011), the MCMC algorithm 'IDEAL' (Jackman, 2012), and the non-parametric optimal classification estimator 'OC' (Poole et al., 2012). The EM algorithm is faster than the other approaches whether focused on point estimates or also estimation uncertainty.

We begin by examining the computational performance of the EM algorithm.[2] Figure 1 shows the time required for the ideal point estimation for each Congressional session in House (left panel) and Senate (right panel). Note that the vertical axis in on the log-scale. Although the results are only illustrative for the aforementioned reasons, it is clear that the EM algorithm is by far the fastest. For example, for the 102nd House of Representatives, the proposed algorithm, denoted by 'EM', takes less than one second to compute estimates, using the same convergence criteria as W-NOMINATE. Even with a much more stringent convergence criteria 'EM (high-precision)', the computational time is only six seconds. This contrasts with the other algorithms, which require much more time for estimation. Although the direct comparison is difficult, the MCMC algorithm is by far the slowest, taking more than 2.5 hours. Because the MCMC algorithm produces standard errors, we contrast the performance of 'IDEAL' with 'EM with Bootstrap' and find that obtaining

[2]All computation in this paper is completed on a cluster computer running Red Hat Linux with multiple 2.67 GHz Intel Xeon X5550 processors. Unless otherwise noted, however, the computation is done utilizing a single processor to emulate the computational performance of our algorithms on a personal computer.

100 bootstrap replicates requires just under one minute. Even if more iterations are desired, over 10,000 bootstrap iterations could be computed before approaching the time required by the MCMC algorithm.

The W-NOMINATE and optimal classification estimators are faster than the MCMC algorithm but take approximately one and 2.5 minutes, respectively. These methods do not provide measures of uncertainty, and all of the point-estimate EM variants are over ten times faster. Lastly, the EM algorithm is amenable to parallelization within each of the three update steps. The open-source implementation that we provide supports this on some platforms (Imai, Lo and Olmsted, 2015). And, the parallelized implementation performs well. For any of these roll-call matrices, using eight processor cores to estimate the parameters instead of just one core reduces the required time-to-completion to about one sixth of the single core time.

We next show that the computational gain of the proposed algorithm is achieved without sacrificing the quality of estimates. To do so, we directly compare individual-level ideal point estimates across the methods. Figure 2 shows, using the 112th Congress, that except for a small number of legislators, the estimates from the proposed EM algorithm are essentially identical to those from the MCMC algorithm (left column) and W-NOMINATE (right column). The within-party correlation remains high across the plots, indicating that a strong agreement among these estimates up to affine transformation. The agreement with the results based on the MCMC algorithm is hardly surprising given that both algorithms are based on the same posterior distribution. The comparability of estimates across methods holds for the other sessions of Congress considered.

For a small number of legislators, the deviation between results for the proposed EM algorithm and both the MCMC algorithm and W-NOMINATE is not negligible. However, this is not a coincidence — it results from the degree of missing votes associated with each legislator.[3] The individuals for whom the estimates differ significantly all have no position registered for more than 40% of the possible votes. Examples include President Obama, the late Congressman Donald Payne (Democrat, NJ), and Congressman Thomas Massie (Republican, KY). The source of missingness differs for all three. Observations for the president are generated by interpreting statements of support or opposition (including vetoes) as votes. While these are not uncommon, they are issued far less frequently than Congress takes roll calls. Congressman Payne, on the other hand, passed away in the middle of the 112th session of Congress due to complications of colon cancer. Finally, Congressman Massie

[3]The discrepancies in these estimates are present even when the MCMC algorithm is executed by imputing missing votes rather than simply dropping them.

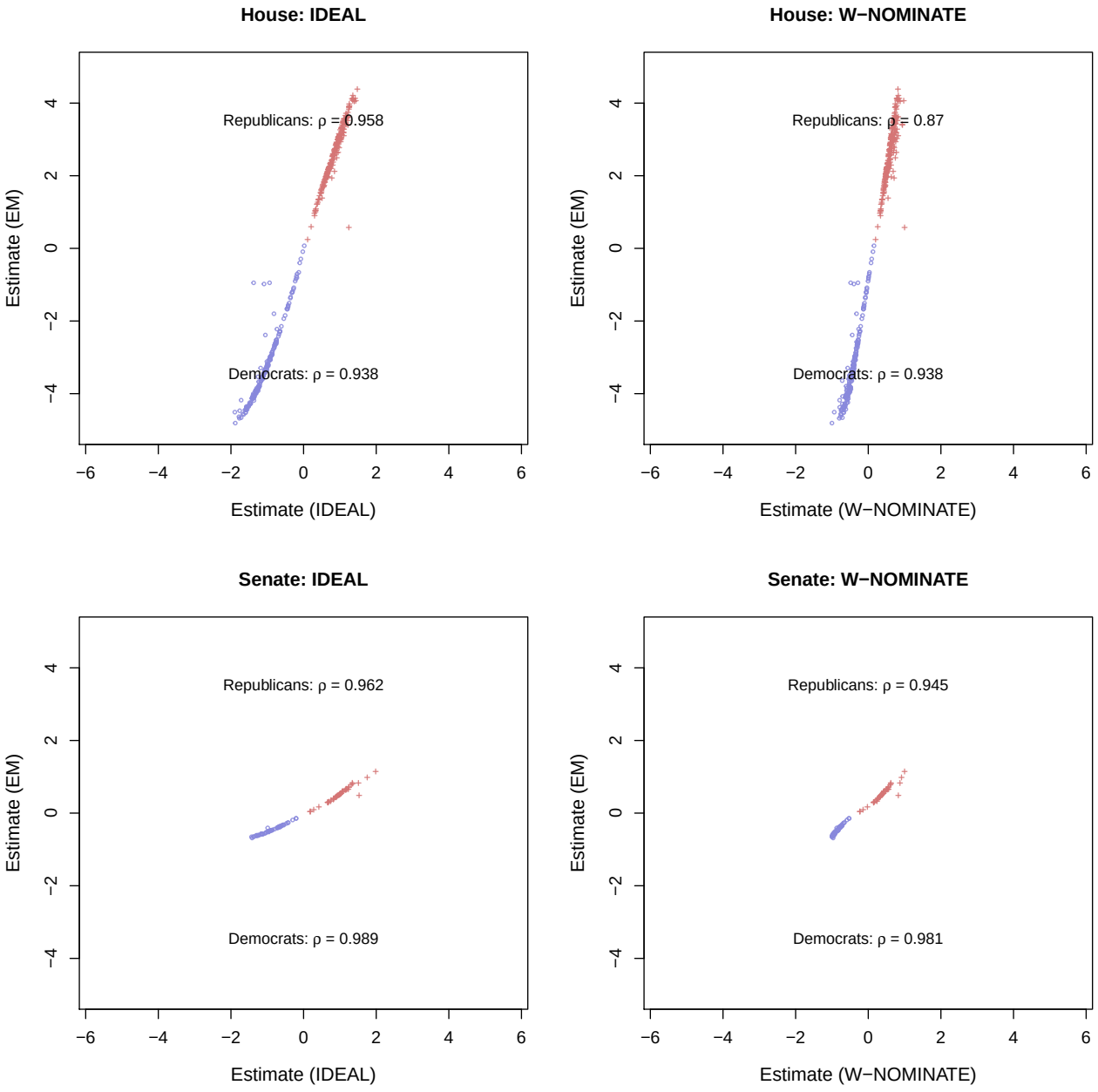

Figure 2: Comparison of Estimated Ideal Points across the Methods for the 112th Congress. Republicans are shown with red crosses while Democrats are indicated by blue hollow circles. The proposed EM algorithm is compared with the MCMC algorithm 'IDEAL' (left column; Jackman, 2012) and 'W-NOMINATE' (right column; Poole et al., 2011). Within-party Pearson correlation coefficients are also reported. The proposed algorithm yields the estimates that are essentially identical to those from the other two methods.

was sworn into office as the representative for Kentucky's 4th congressional district after winning a special election in November 2012 with just three months left in the 112th session.

We also compare the standard errors from the proposed EM algorithm using the parametric bootstrap with those from the MCMC algorithm. Because the output from the MCMC algorithm is rescaled to have a mean of zero and standard deviation of one, we rescale the EM estimates to

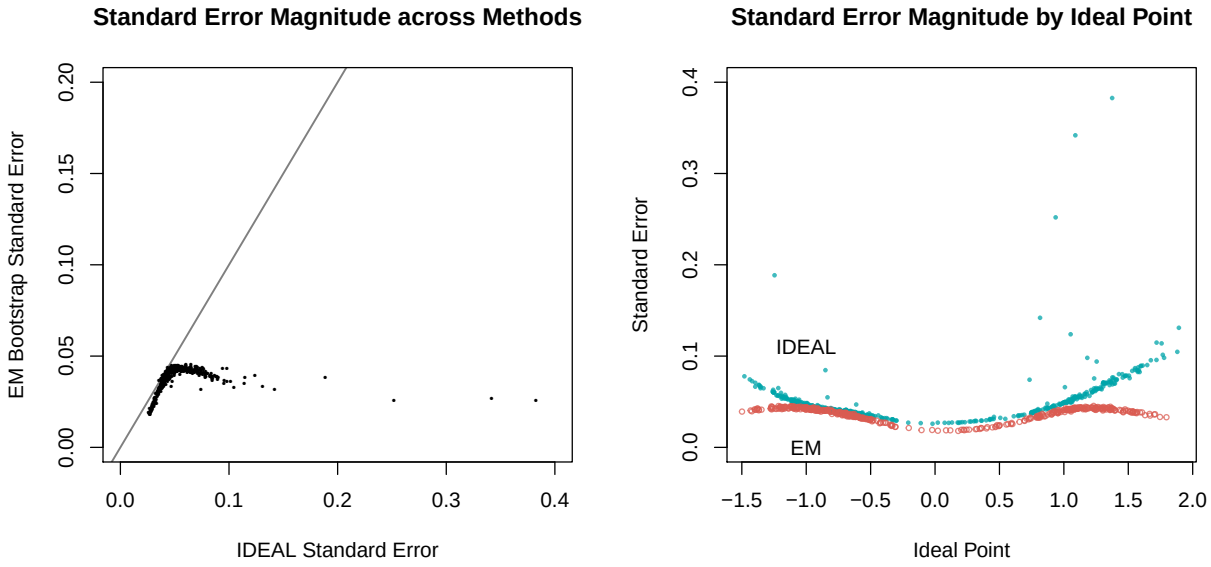

Figure 3: Comparison of Standard Errors between the Proposed EM Algorithm and the Bayesian MCMC Algorithm using the 112th House of Representatives. The standard errors from the EM algorithm are based on the parametric bootstrap of 1,000 replicates. The left plot shows that the proposed standard errors (the vertical axis) are similar to those from the MCMC algorithms (the horizontal axis) for most legislators. For some legislators, the MCMC standard errors are much larger. The right panel shows that these legislators tend to have extreme ideological preferences.

have the same sample moments. This affine transformation is applied to each bootstrap replicate so that the resulting bootstrap standard errors are on the same scale as those based on the MCMC algorithm. Figure 3 does this comparison using the estimates for the 112th House of Representatives. The left panel shows that the standard errors based on the EM algorithm with the bootstrap (the vertical axis) are only slightly smaller than those from the MCMC algorithm (the horizontal axis) for most legislators. However, for a few legislators, the standard errors from the EM algorithm are substantially smaller than those from the MCMC algorithm. The right panel of the figure shows that these legislators have extreme ideological preferences.

2.4 Simulation Evidence

Next, we use Monte Carlo simulation to assess the computational scalability of the proposed EM algorithm and the accuracy of the parametric bootstrap procedure. First, we examine the computational time of the proposed algorithm as the dimensions of the roll-call matrix get larger. The empirical application in the previous subsection demonstrated the poor scalability of the MCMC algorithm. Hence, we compare the performance of our algorithm with W-NOMINATE. Figure 4 shows the median performance of both the EM algorithm with high precision and 'W-NOMINATE' over 25 Monte Carlo trials for various numbers of legislators and bills. In the left panel, the num-

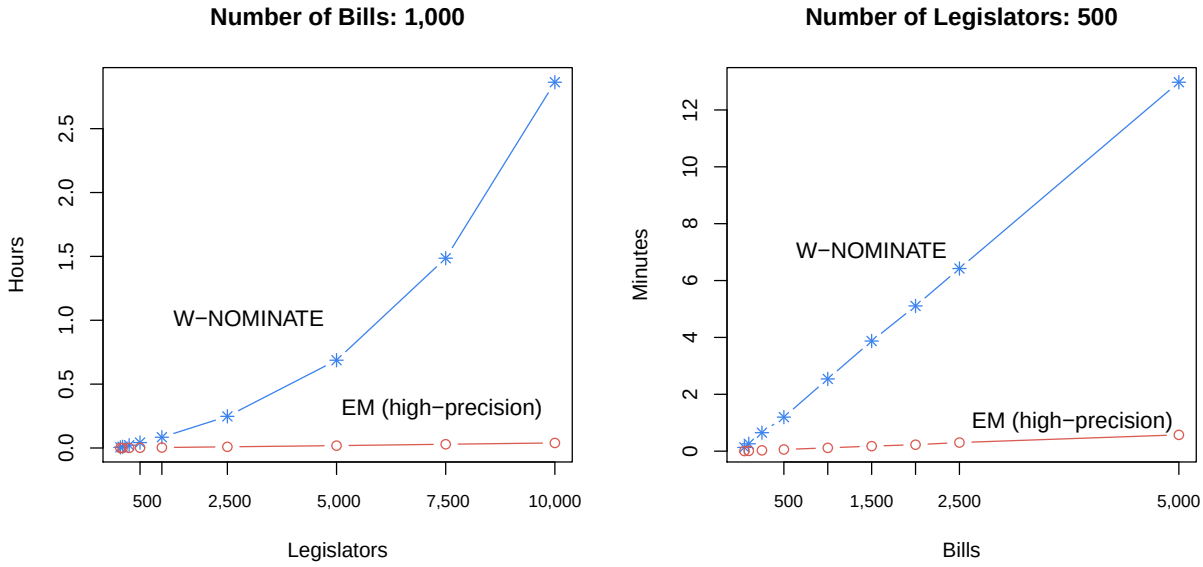

Figure 4: Comparison of Changing Performance across the Methods as the Dimensions of Roll-Call Matrix Increase. Estimation time is shown on the vertical axis as the number of legislators increases (left panel) and the number of bills increases (right panel). Values are the median times over 25 replications. 'EM (high precision)' is more computationally efficient than W-NOMINATE (Poole et al., 2011) especially when the roll-call matrix is large.

ber of bills is fixed at 1,000 and the number of legislators ranges from 50 to 10,000. In the right panel, the number of legislators is fixed at 500, and the number of bills ranges from 50 to 5,000. In both cases, the data-generating process follows the single dimensional ideal point model where ideal points are generated according to the standard normal distribution and the bill parameters follow from a normal distribution with mean zero and standard deviation of 10. These parameter values are chosen so that the true parameter values explain around 85% percent of the observed votes—a level of classification success comparable to the in-sample fit obtained in contemporary sessions of Congress.

The computational efficiency of the proposed algorithm can be seen immediately. Even with 10,000 legislators and 1,000 bills, convergence at high-precision is achieved in less than 15 minutes. The runtime of our algorithm increases only linearly as the number of ideal points increases. This contrasts with W-NOMINATE, whose required computation time grows exponentially as the number of legislators increases. For example, both the EM algorithm and W-NOMINATE require less than 5 minutes to estimate the parameters associated with a roll-call matrix with 1,000 legislators and 1,000 bills. However, when the number of legislators increases to 10,000, W-NOMINATE takes around 2.5 hours while the EM algorithm requires less than 15 minutes. The difference is less stark when the number of bills increase (right panel). Even here, however, the EM algorithm is more computationally efficient especially when the number of bills is large.

3 Ideal Point Model with Mixed Binary, Ordinal, and Continuous Outcomes

We extend the EM algorithm developed above to the ideal point model with mixed binary, ordinal, and continuous outcomes. Quinn (2004) develops an MCMC algorithm for fitting this model and implements it as the `MCMCmixfactanal()` function in the open-source R package `MCMCpack` (Martin, Quinn and Park, 2013). The EM algorithm for the ordinal probit model, which is closely related to this model, poses a special challenge because its E-step is not available in closed form. Perhaps, for this reason, to the best of our knowledge, the EM algorithm has not been developed for the ordinal probit model in the statistics literature.

In this section, we first show that the E-step can be derived in a closed-form so long as the outcome variable only has three ordinal categories. With a suitable transformation of parameters, we derive an EM algorithm that is analytically tractable. We then consider the cases where the number of categories in the outcome variable exceeds three and the outcome variable is a mix of binary, ordinal, and continuous variables. Finally, we apply the proposed algorithm to a survey of Japanese politicians and voters.

3.1 The Model with a Three-category Ordinal Outcome

We consider the same exact setup as in Section 2 with the exception that the outcome variable now takes one of the three ordered values, i.e., $y_{ij} \in \{0, 1, 2\}$. In this model, the probability of each observed choice is given as follows,

$$\Pr(y_{ij} = 0) = \Phi(\alpha_{1j} - \mathbf{x}_i^\top \boldsymbol{\beta}_j) \tag{9}$$

$$\Pr(y_{ij} = 1) = \Phi(\alpha_{2j} - \mathbf{x}_i^\top \boldsymbol{\beta}_j) - \Phi(\alpha_{1j} - \mathbf{x}_i^\top \boldsymbol{\beta}_j) \tag{10}$$

$$\Pr(y_{ij} = 2) = 1 - \Phi(\alpha_{2j} - \mathbf{x}_i^\top \boldsymbol{\beta}_j) \tag{11}$$

where $\alpha_{2j} > \alpha_{1j}$ for all $j = 1, 2, \ldots, J$. The model can be written using the latent propensity to agree y_{ij}^* for respondent i as,

$$y_{ij}^* = \mathbf{x}_i^\top \boldsymbol{\beta}_j + \epsilon_{ij} \tag{12}$$

where $\epsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and these latent propensities are connected to the observed outcomes through the following relationship,

$$y_{ij} = \begin{cases} 0 & \text{if } y^*_{ij} < \alpha_{1j} \\ 1 & \text{if } \alpha_{1j} \leq y^*_{ij} < \alpha_{2j} \\ 2 & \text{if } \alpha_{2j} \leq y^*_{ij} \end{cases} \tag{13}$$

As in the standard ideal point model, we treat abstention as missing at random.

Following the literature, we assume the same normal independent prior distribution on $(\{\boldsymbol{\beta}_j\}_{j=1}^J, \{\mathbf{x}_i\}_{i=1}^N)$ as the one used in Section 2.1. For $(\{\alpha_{1j}, \alpha_{2j}\}_{j=1}^J)$, we assume an improper uniform prior with the appropriate ordering restriction $\alpha_{1j} < \alpha_{2j}$. Then, the joint posterior distribution is given by,

$$\begin{aligned} & p(\mathbf{Y}^*, \{\mathbf{x}_i\}_{i=1}^N, \{\alpha_{1j}, \alpha_{2j}, \boldsymbol{\beta}_j\}_{j=1}^J \mid \mathbf{Y}) \\ \propto & \prod_{i=1}^N \prod_{j=1}^J \left[\mathbf{1}\{y^*_{ij} < \alpha_{1j}\}\mathbf{1}\{y_{ij}=0\} + \mathbf{1}\{\alpha_{1j} \leq y^*_{ij} < \alpha_{2j}\}\mathbf{1}\{y_{ij}=1\} + \mathbf{1}\{y^*_{ij} \geq \alpha_{2j}\}\mathbf{1}\{y_{ij}=2\}\right] \phi_1(y^*_{ij}; \mathbf{x}_i^\top \boldsymbol{\beta}_j, 1) \\ & \times \prod_{i=1}^N \phi_K(\mathbf{x}_i; \boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}}) \prod_{j=1}^J \phi_K(\boldsymbol{\beta}_j; \boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}) \end{aligned} \tag{14}$$

3.2 The Proposed Algorithm

To develop an EM algorithm that is analytically tractable, we employ the following one-to-one transformation of parameters,

$$\tau_j = \alpha_{2j} - \alpha_{1j} > 0 \tag{15}$$

$$\alpha^*_j = -\frac{\alpha_{1j}}{\tau_j} \tag{16}$$

$$\boldsymbol{\beta}^*_j = \frac{\boldsymbol{\beta}_j}{\tau_j} \tag{17}$$

$$z^*_{ij} = \frac{y^*_{ij} - \alpha_{1j}}{\tau_j} \tag{18}$$

$$\epsilon^*_{ij} = \frac{\epsilon_{ij}}{\tau_j} \tag{19}$$

Then, the simple algebra shows that the model can be rewritten as,

$$\Pr(y_{ij}=0) = \Phi(-\tau_j \alpha^*_j - \tau_j \mathbf{x}_i^\top \boldsymbol{\beta}^*_j) \tag{20}$$

$$\Pr(y_{ij}=1) = \Phi(-\tau_j \alpha^*_j + \tau_j - \tau_j \mathbf{x}_i^\top \boldsymbol{\beta}^*_j) - \Phi(-\tau_j \alpha^*_j - \tau_j \mathbf{x}_i^\top \boldsymbol{\beta}^*_j) \tag{21}$$

$$\Pr(y_{ij}=2) = 1 - \Phi(-\tau_j \alpha^*_j + \tau_j - \tau_j \mathbf{x}_i^\top \boldsymbol{\beta}^*_j) \tag{22}$$

where the latent variable representation is given by,

$$z^*_{ij} \;=\; \alpha^*_j + \mathbf{x}_i^\top \boldsymbol{\beta}^*_j + \epsilon^* \quad \text{with} \quad \epsilon^*_{ij} \overset{\text{indep.}}{\sim} \mathcal{N}(0,\ \tau_j^{-2}). \tag{23}$$

Under this parameterization, the relationship between the observed outcome y_{ij} and the latent variable z^*_{ij} is given by,

$$y_{ij} \;=\; \begin{cases} 0 & \text{if } \; z^*_{ij} < 0 \\ 1 & \text{if } \; 0 \leq z^*_{ij} < 1 \\ 2 & \text{if } \; 1 \leq z^*_{ij} \end{cases} \tag{24}$$

Thus, the consequence of this reparameterization is that the threshold parameters $(\alpha_{1j}, \alpha_{2j})$ are replaced with the intercept term α^*_j and the heterogenous variance parameter τ_j^{-2}.

To maintain conjugacy, we alter the prior distribution specified in equation (14). In particular, we use the following prior distribution,

$$p(\{\tilde{\boldsymbol{\beta}}_j\}_{j=1}^J, \{\tau_j^2\}_{j=1}^J, \{\mathbf{x}_i\}_{i=1}^N) \;=\; \prod_{j=1}^J \phi_{K+1}(\tilde{\boldsymbol{\beta}}_j; \boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}})\ \mathcal{G}\left(\tau_j^2; \frac{\nu_\tau}{2}, \frac{s_\tau}{2}\right) \prod_{i=1}^N \phi_K(\mathbf{x}_i; \boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}}) \tag{25}$$

where $\tilde{\boldsymbol{\beta}}_j = (\alpha^*_j, \boldsymbol{\beta}^*_j)$ and $\mathcal{G}(\tau_j^2; \nu_\tau/2, s_\tau/2)$ is the Gamma distribution with $\nu_\tau/2 > 0$ and $s_\tau/2 > 0$ representing the prior shape and rate parameters, respectively. This change in prior distribution alters the model but so long as the prior is diffuse and the size of data is large the resulting inference should not differ much.

Given this setup, we derive the EM algorithm to maximize the posterior distribution. We take the analytical strategy similar to the one used for the standard ideal point model in Section 2.2. Specifically, the "Q-function" of the EM algorithm is given by,

$$\begin{aligned}
& Q(\{\mathbf{x}_i\}_{i=1}^N, \{\tau_j, \tilde{\boldsymbol{\beta}}_j\}_{j=1}^J) \\
= \;& \mathbb{E}\left\{\log p(\mathbf{Z}^*, \{x_i\}_{i=1}^N, \{\tau_j^2, \tilde{\boldsymbol{\beta}}_j\}_{j=1}^J \mid \mathbf{Y}) \mid \{\mathbf{x}_i^{(t-1)}\}_{i=1}^N, \{\tau_j^{(t-1)}, \tilde{\boldsymbol{\beta}}_j^{(t-1)}\}_{j=1}^J, \mathbf{Y}\right\} \\
= \;& \sum_{i=1}^N \sum_{j=1}^J \mathbb{E}\left\{\log \phi_1(z^*_{ij}; \tilde{\mathbf{x}}_i^\top \tilde{\boldsymbol{\beta}}_j, \tau_j^{-2})\right\} + \sum_{j=1}^J \left\{\log \phi_{K+1}(\tilde{\boldsymbol{\beta}}_j; \boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}) + \log \mathcal{G}\left(\tau_j^2; \frac{\nu_\tau}{2}, \frac{s_\tau^2}{2}\right)\right\} \\
& + \sum_{i=1}^N \log \phi_K(\mathbf{x}_i; \boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}}) + \text{const.} \\
= \;& \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^J \left[\log \tau_j^2 - \tau_j^2 \left\{\tilde{\boldsymbol{\beta}}_j^\top \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top \tilde{\boldsymbol{\beta}}_j - 2\tilde{\boldsymbol{\beta}}_j^\top \tilde{\mathbf{x}}_i z^{*\,(t)}_{ij} + \left(z^{*\,2}_{ij}\right)^{(t)}\right\}\right] - \frac{1}{2} \sum_{i=1}^N \left(\mathbf{x}_i^\top \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \mathbf{x}_i - 2\mathbf{x}_i^\top \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \boldsymbol{\mu}_{\mathbf{x}}\right) \\
& - \sum_{j=1}^J \left[\frac{1}{2}\left(\tilde{\boldsymbol{\beta}}_j^\top \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}^{-1} \tilde{\boldsymbol{\beta}}_j - 2\tilde{\boldsymbol{\beta}}_j^\top \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}^{-1} \boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}}\right) - \left\{\left(\frac{\nu_\tau}{2} - 1\right) \log \tau_j^2 - \frac{s_\tau^2 \tau_j^2}{2}\right\}\right] + \text{const.}
\end{aligned} \tag{26}$$

The latent variable updates are equal to,

$$z_{ij}^{*\,(t)} \;=\; \mathbb{E}(z_{ij}^{*} \mid \mathbf{x}_i^{(t-1)}, \tau_j^{(t-1)}, \tilde{\boldsymbol{\beta}}_j^{(t-1)}, y_{ij}) \;=\; \begin{cases} m_{ij}^{(t-1)} - \frac{1}{\tau_j^{(t-1)}} \lambda\left(m_{ij}^{(t-1)},\ \tau_j^{(t-1)}\right) & \text{if } y_{ij} = 0 \\ m_{ij}^{(t-1)} + \frac{1}{\tau_j^{(t-1)}} \delta\left(1 - m_{ij}^{(t-1)},\ \tau_j^{(t-1)}\right) & \text{if } y_{ij} = 1 \\ m_{ij}^{(t-1)} + \frac{1}{\tau_j^{(t-1)}} \lambda\left(1 - m_{ij}^{(t-1)},\ \tau_j^{(t-1)}\right) & \text{if } y_{ij} = 2 \end{cases} \tag{27}$$

where $m_{ij}^{(t-1)} = \left(\tilde{\mathbf{x}}_i^{(t-1)}\right)^\top \tilde{\boldsymbol{\beta}}_j^{(t-1)}$, $\lambda(m,\tau) = \phi(m\tau)/\{1 - \Phi(m\tau)\}$ and $\delta(m,\tau) = \{\phi(m\tau) - \phi((1 - m)\tau)\}/\{\Phi((1-m)\tau) + \Phi(m\tau) - 1\}$. If y_{ij} is missing, then we set $z_{ij}^{*\,(t)} = m_{ij}^{(t-1)}$. The required second moment is given by,

$$\begin{aligned} &\left(z_{ij}^{*\,2}\right)^{(t)} \\ &= \mathbb{E}(z_{ij}^{*\,2} \mid \mathbf{x}_i^{(t-1)}, \tau_j^{(t-1)}, \tilde{\boldsymbol{\beta}}_j^{(t-1)}, y_{ij}) \\ &= \begin{cases} \left(z_{ij}^{*\,(t)}\right)^2 + \frac{1}{\left(\tau_j^{(t-1)}\right)^2}\left[1 + \tau_j^{(t-1)} m_{ij}^{(t-1)} \lambda\left(m_{ij}^{(t-1)},\ \tau_j^{(t-1)}\right) - \left\{\lambda\left(m_{ij}^{(t-1)},\ \tau_j^{(t-1)}\right)\right\}^2\right] & \text{if } y_{ij} = 0 \\ \left(z_{ij}^{*\,(t)}\right)^2 + \frac{1}{\left(\tau_j^{(t-1)}\right)^2}\left[1 - \frac{\tau_j^{(t-1)}\left\{m_{ij}^{(t-1)}\phi\left(m_{ij}^{(t-1)}\tau_j^{(t-1)}\right) + \left(1 - m_{ij}^{(t-1)}\right)\phi\left(\left(1 - m_{ij}^{(t-1)}\right)\tau_j^{(t-1)}\right)\right\}}{\Phi\left(\left(1 - m_{ij}^{(t-1)}\right)\tau_j^{(t-1)}\right) + \Phi\left(m_{ij}^{(t-1)}\tau_j^{(t-1)}\right) - 1} - \left\{\delta\left(1 - m_{ij}^{(t-1)},\ \tau_j^{(t-1)}\right)\right\}^2\right] & \text{if } y_{ij} = 1 \\ \left(z_{ij}^{*\,(t)}\right)^2 + \frac{1}{\left(\tau_j^{(t-1)}\right)^2}\left[1 - m_{ij}^{(t-1)} \lambda\left(1 - m_{ij}^{(t-1)},\ \tau_j^{(t-1)}\right)\left\{\lambda\left(1 - m_{ij}^{(t-1)},\ \tau_j^{(t-1)}\right) - \left(1 - m_{ij}^{(t-1)}\right)\tau_j^{(t-1)}\right\}\right] & \text{if } y_{ij} = 2 \end{cases} \end{aligned} \tag{28}$$

where if y_{ij} is missing, then we set $\left(z_{ij}^{*\,(t)}\right)^2 = \left(m_{ij}^{(t-1)}\right)^2 + \left(\tau_j^{(t-1)}\right)^{-2}$.

Finally, the M-step consists of the following conditional maximization steps,

$$\mathbf{x}_i^{(t)} \;=\; \left(\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} + \sum_{j=1}^{J}\left(\tau_j^{(t-1)}\right)^2 \boldsymbol{\beta}_j^{(t-1)} \boldsymbol{\beta}_j^{(t-1)\top}\right)^{-1} \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\mu}_{\mathbf{x}} + \sum_{j=1}^{J}\left(\tau_j^{(t-1)}\right)^2 \boldsymbol{\beta}_j^{(t-1)}(z_{ij}^{*\,(t)} - \alpha_j^{(t-1)}) \tag{29}$$

$$\tilde{\boldsymbol{\beta}}_j^{(t)} \;=\; \left(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}^{-1} + \left(\tau_j^{(t-1)}\right)^2 \sum_{i=1}^{N} \tilde{\mathbf{x}}_i^{(t)}\left(\tilde{\mathbf{x}}_i^{(t)}\right)^\top\right)^{-1} \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}^{-1}\boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}} + \left(\tau_j^{(t-1)}\right)^2 \sum_{i=1}^{N} \tilde{\mathbf{x}}_i^{(t)} z_{ij}^{*\,(t)} \tag{30}$$

$$\left(\tau_j^{(t)}\right)^2 \;=\; \frac{N + \nu_\tau - 2}{s_\tau^2 + \left(\tilde{\boldsymbol{\beta}}_j^{(t)}\right)^\top \left\{\sum_{i=1}^{N} \tilde{\mathbf{x}}_i^{(t)}\left(\tilde{\mathbf{x}}_i^{(t)}\right)^\top\right\} \tilde{\boldsymbol{\beta}}_j^{(t)} - 2\left(\tilde{\boldsymbol{\beta}}_j^{(t)}\right)^\top \sum_{i=1}^{N} \tilde{\mathbf{x}}_i^{(t)} z_{ij}^{*\,(t)} + \sum_{i=1}^{N}\left(z_{ij}^{*\,2}\right)^{(t)}} \tag{31}$$

3.3 Mixed Binary, Ordinal, and Continuous Outcomes

Here, we consider how to apply, possibly after some modification, the EM algorithm developed above to a more genereal case with mixed binary, ordinal, and continuous outcomes. If the number

of ordered categories in outcome exceeds three, we collapse them into three categories. For example, responses to a survey question on the 5-point Likert scale, i.e., `strongly disagree, disagree, neither agree nor disagree, agree, strongly agree`, may be converted into a 3-point scale by combining `strongly disagree` and `disagree` into a single category and `strongly agree` and `agree` into another category.

Next, suppose that y_{ij} is a binary outcome for a particular observation (i, j). Then, we consider the following relationship between this observed outcome and the latent propensity z^*_{ij}; $z_{ij} < 1 \Longleftrightarrow y_{ij} = 0$ and $z_{ij} \geq 0 \Longleftrightarrow y_{ij} = 1$. Under this assumption, the E-step becomes,

$$z_{ij}^{*\,(t)} = \begin{cases} m_{ij}^{(t-1)} - \frac{1}{\tau_j^{(t-1)}} \lambda\left(-m_{ij}^{(t-1)} + 1, \tau_j^{(t-1)}\right) & \text{if } y_{ij} = 0 \\ m_{ij}^{(t-1)} + \frac{1}{\tau_j^{(t-1)}} \lambda\left(m_{ij}^{(t-1)}, \tau_j^{(t-1)}\right) & \text{if } y_{ij} = 1 \end{cases} \tag{32}$$

and

$$\left(z_{ij}^{*\,2}\right)^{(t)} = \begin{cases} \left(z_{ij}^{*\,(t)}\right)^2 + \frac{1}{\left(\tau_j^{(t-1)}\right)^2}\left[1 - \frac{1-m_{ij}^{(t-1)}}{\tau_j^{(t-1)}}\lambda\left(m_{ij}^{(t-1)} - 1, \tau_j^{(t-1)}\right) - \left\{\lambda\left(m_{ij}^{(t-1)} - 1, \tau_j^{(t-1)}\right)\right\}^2\right] & \text{if } y_{ij} = 0 \\ \left(z_{ij}^{*\,(t)}\right)^2 + \frac{1}{\left(\tau_j^{(t-1)}\right)^2}\left[1 - \lambda\left(-m_{ij}^{(t-1)}, \tau_j^{(t-1)}\right)\left\{\lambda\left(-m_{ij}^{(t-1)}, \tau_j^{(t-1)}\right) + m_{ij}^{(t-1)}\tau_j^{(t-1)}\right\}\right] & \text{if } y_{ij} = 1 \end{cases} \tag{33}$$

where if y_{ij} is missing, we set $z_{ij}^{*\,(t)} = m_{ij}^{(t-1)}$ and $\left(z_{ij}^{*\,2}\right)^{(t)} = \left(z_{ij}^{*\,(t)}\right)^2 + \left(\tau_j^{(t-1)}\right)^{-2}$. Other than these modifications, the rest of the EM algorithm stays identical.

Finally, it is straightforward to extend this model to also include a continuous outcome as done by Quinn (2004). In that case, set the first and second moments of the latent propensity as $z_{ij}^{*\,(t)} = y_{ij}$ and $\left(z_{ij}^{*\,2}\right)^{(t)} = y_{ij}^2 + \left(\tau_j^{(t-1)}\right)^{-2}$ for this observation. The rest of the EM algorithm remain unchanged.

3.4 An Empirical Application

We apply the ordinal ideal point model to the survey data of the candidates and voters of Japanese Upper and Lower House elections. This Asahi-Todai Elite survey was conducted by the University of Tokyo in collaboration with a major national newspaper, the Asahi Shimbun, covering all candidates (both incumbents and challengers) for the eight elections that occurred between 2003 and 2013. Six out of eight waves, the survey was also administered to a nationally representative sample of voters with the sample size ranging from approximately 1,100 to about 2,000. The novel feature of the data is that there are a set of common policy questions, which can be used to scale both politicians and voters over time on the same dimension. Another important advantage of the data is a high response

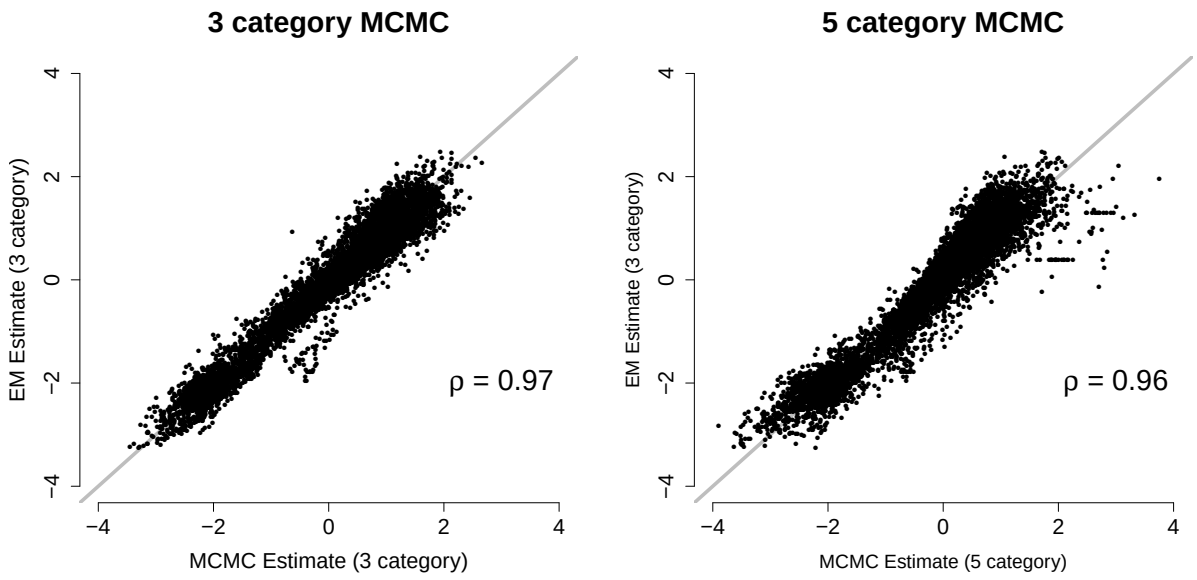

Figure 5: Comparison of Ideal Point Estimates from the EM and Markov chain Monte Carlo (MCMC) Algorithms for Japanese Politicians Using the Asahi-Todai Elite survey. The figures compare the EM estimates (horizontal axis) against MCMC estimates (vertical axis). The EM estimates use a coarsened three category response, which is compared against the MCMC estimates based on the same three category response (left panel) and the original five category response (right panel). Overall correlation between the EM and MCMC estimates are high, exceeding 0.95 in both cases.

rate among politicians, which exceeded 85%. Such a high response rate is obtained in large part because the survey results are published in the Asahi Shimbun whose circulation is approximately eight million (see Hirano et al., 2011, for more details).

All together, the data set we analyze contains a total of $N = 19,443$ respondents, including $7,734$ politicians and $11,709$ voters. Here, we count multiple appearances of the same politician separately because an ideal point will be estimated separately for each wave. There are $J = 98$ unique questions in the survey, most of which consisted of questions asking for responses on a 5-point Likert scale. We apply the proposed EM algorithm after coarsening each response into three categories (disagree, neutral, and agree). For the purpose of comparison, we have written a C code for implementing the standard MCMC algorithm because the data set is too large for `MCMCmixfactanal()` of the `MCMCpack` package to handle. One model uses the full range of categories found in the data without coarsening, and the other model uses the same coarsened responses as done for our algorithm. Obtaining 10,000 draws from the posterior distribution using the MCMC algorithm takes 4 hours and 24 minutes (5 category) or 3 hours and 54 minutes (3 category). In contrast, estimation using our proposed EM algorithm takes 164 iterations and only 68 seconds to complete where the algorithm is iterated until the correlation of parameter values between two consecutive iterations reaches $1 - 10^{-6}$.

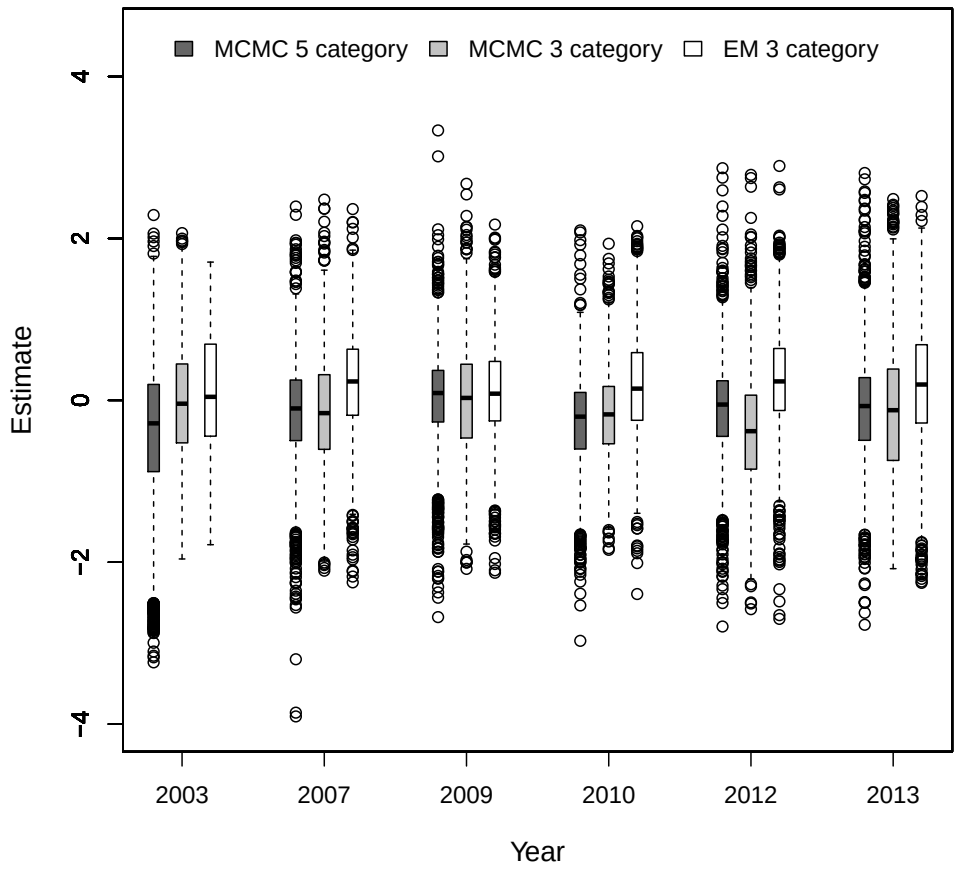

Figure 6: Comparing the Distributions of Estimated Ideal Points between the EM and Markov chain Monte Carlo (MCMC) Algorithms for Japanese Voters across Six Waves of the Asahi-Todai Elite survey. White boxplots describe the distribution of the EM estimates whereas light and dark grey boxplots represent the MCMC Estimates for the coarsened three category and original five category responses, respectively. Across all waves, these three algorithms produce similar estimates of ideal points.

Figure 5 compares the estimated ideal points of politicians based on our EM algorithm (vertical axis) against those obtained from the standard MCMC algorithm (horizontal axis). As explained earlier, the EM estimates are based on the coarsened three category response, while we present the MCMC estimates using the original five category response (right panel) as well as the same coarsened three category response (left panel). The plots show that these two algorithms produce essentially identical estimates, achieving the correlation greater than 0.95. In addition, for this data set, coarsening the original five category response into a three category response does not appear to have a significant impact on the degree of correlation between the two sets of ideal point estimates of Japanese politicians.

Figure 6 compares the estimated ideal points of voters for each wave of survey, obtained from our

EM algorithm (white boxplots) and the standard MCMC algorithm (light and dark grey boxplots for the coarsened three category and original five category responses, respectively). Across all six waves of the survey, the three algorithms give similar distribution of estimated ideal points. The differences across the algorithms lie mostly in their estimated ideal points for a small subset of voters who answer too few questions. For example, the 2003 survey included only two policy questions, and 306 respondents from this survey gave the same two responses. For these respondents, our EM algorithm produces an identical ideal point estimate of -0.89 whereas the MCMC algorithm gives a set of ideal points ranging from about -3 to 0, mainly due to the imprecise nature of posterior mean point estimates when votes are not informative. Overall, the results suggest that our EM algorithm recovers virtually identical estimates to those derived via the standard MCMC algorithm but with substantial savings in time.

4 Dynamic Ideal Point Model

We next consider the dynamic ideal point model of Martin and Quinn (2002) who characterized how the ideal points of supreme court justices change over time. The authors develop an MCMC algorithm for fitting this model and make it available as the `MCMCdynamicIRT1d()` function in the open-source R package `MCMCpack` (Martin, Quinn and Park, 2013). This methodology is based on the dynamic linear modeling approach and is more flexible than polynomial time trend models considered by other scholars (see e.g., DW-NOMINATE, Bailey, 2007).

Nevertheless, this flexibility comes at a significant computational cost. In particular, Martin and Quinn (2002) report that using a dedicated workstation it took over 6 days to estimate ideal points for US supreme court justices over 47 years (footnote 12). Because of this computational burden, Bailey (2007) resorts to a simpler parametric dynamic model (p. 441). In addition, unlike the two models we considered above, no closed-form EM algorithm is available for maximizing the posterior in this case. Therefore, we propose the variational inference that approximates the posterior inference by deriving the variational EM algorithm. We show that the proposed algorithm is orders of magnitude faster than the standard MCMC algorithm and scales to a large data set while yielding the estimates that are similar to those obtained from the standard MCMC algorithm.

4.1 The Model

Let y_{ijt} be an indicator variable representing the observed vote of legislator i on roll call j at time t where $y_{ijt} = 1$ ($y_{ijt} = 0$) represents "yea" ("nay"). There are a total of N unique legislators, i.e., $i = 1, \ldots, N$, and for any given time period t, there are J_t roll calls, i.e., $j = 1, \ldots, J_t$. Lastly, the

number of time periods is T, i.e., $t = 1, \ldots, T$. Then, the single-dimensional ideal point model is given by,

$$\Pr(y_{ijt} = 1) \;=\; \Phi(\alpha_{jt} + \beta_{jt} x_{it}) \;=\; \Phi(\tilde{\mathbf{x}}_{it}^{\top} \tilde{\boldsymbol{\beta}}_{jt}) \tag{34}$$

where x_{it} is justice i's ideal point at time t, and α_{jt} and β_{jt} represent the item difficulty and item discrimination parameters for roll call j in time t, respectively. Note that as before we use a vector notation $\tilde{\mathbf{x}}_{it} = (1, x_{it})$ and $\tilde{\boldsymbol{\beta}}_{jt} = (\alpha_{jt}, \beta_{jt})$. As before, the model can be rewritten with the latent propensity y^*_{ijt},

$$y^*_{ijt} \;=\; \tilde{\mathbf{x}}_{it}^{\top} \tilde{\boldsymbol{\beta}}_{jt} + \epsilon_{ijt} \tag{35}$$

where $\epsilon_{ijt} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ and $y_{ijt} = 1$ ($y_{ijt} = 0$) if $y^*_{ijt} > 0$ ($y^*_{ijt} \leq 0$).

As done in the standard dynamic linear modeling framework, the dynamic aspect of the ideal point estimation is specified through the following random walk prior for each legislator i,

$$x_{it} \mid x_{i,t-1} \;\overset{\text{indep.}}{\sim}\; \mathcal{N}(x_{i,t-1},\; \omega_x^2) \tag{36}$$

for $t = \underline{T}_i, \underline{T}_i + 1, \ldots, \overline{T}_i - 1, \overline{T}_i$ where $\underline{T}_i$ is the first time period legislator i enters the data and $\overline{T}_i$ is the last time period the legislator appears in the data, i.e., $1 \leq \underline{T}_i \leq \overline{T}_i \leq T$. In addition, we assume $x_{i,\underline{T}_i - 1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_x, \Sigma_x)$ for each legislator i.

Finally, given this setup, with the independent conjugate prior on $\tilde{\boldsymbol{\beta}}_{jt}$, we have the following joint posterior distribution,

$$\begin{aligned}
& p(\mathbf{Y}^*, \{\mathbf{x}_i\}_{i=1}^N, \{\tilde{\boldsymbol{\beta}}_j\}_{t=1}^T \mid \mathbf{Y}) \\
\propto\; & \prod_{i=1}^{N} \prod_{t=\underline{T}_i}^{\overline{T}_i} \prod_{j=1}^{J_t} \left(\mathbf{1}\{y^*_{ijt} > 0\} \mathbf{1}\{y_{ijt} = 1\} + \mathbf{1}\{y^*_{ijt} \leq 0\} \mathbf{1}\{y_{ijt} = 0\} \right) \; \phi_1\left(y^*_{ijt}; \tilde{\mathbf{x}}_{it}^{\top} \tilde{\boldsymbol{\beta}}_{jt}, 1 \right) \\
& \times \prod_{i=1}^{N} \left\{ \phi_1(x_{i,\underline{T}_i - 1}; \mu_x, \Sigma_x) \prod_{t=\underline{T}_i}^{\overline{T}_i} \phi_1(x_{it}; x_{i,t-1}, \omega_x^2) \right\} \prod_{t=1}^{T} \prod_{j=1}^{J_t} \phi_2(\tilde{\boldsymbol{\beta}}_{jt}; \boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}})
\end{aligned} \tag{37}$$

where $\mathbf{x}_i = (x_{i\underline{T}_i}, \ldots, x_{i\overline{T}_i})$ for $i = 1, \ldots, N$.

4.2 The Proposed Algorithm

We propose a variational EM algorithm for the dynamic ideal point model summarized above. The variational inference makes factorization assumptions and approximates the posterior inference by minimizing the Kullback-Leibler divergence between the true posterior distribution and the factorized

distribution (see Wainwright and Jordan (2008) for a review and Grimmer (2011) for an introductory article in political science). In the current context, we assume the following factorization assumption,

$$q(\mathbf{Y}^*, \{\mathbf{x}_i\}_{i=1}^N, \{\tilde{\boldsymbol{\beta}}_j\}_{t=1}^T) \;=\; \prod_{i=1}^{N} \prod_{t=\underline{T}_i}^{\overline{T}_i} q(y_{it}^*) \prod_{i=1}^{N} q(\mathbf{x}_i) \prod_{t=1}^{T} \prod_{j=1}^{J_t} q(\tilde{\boldsymbol{\beta}}_{jt}) \tag{38}$$

which basically assumes the independence across parameters. Importantly, we do not assume the independence between x_{it} and $x_{it'}$ so that we do not sacrifice our ability to model dynamics of ideal points. We also do not assume a family of approximating distributions. Rather, our results show that the optimal variational distribution belongs to a certain parametric family.

The detailed derivation of the variational EM algorithm is given in Appendix C. For completeness, we also derive the variational EM algorithms for the standard ideal point model (Appendix A) and the ideal point model with an ordinal outcome (Appendix B).

The proposed variational EM algorithm consists of three steps. First, the latent propensity update step is based on the following optimal approximating distribution,

$$q(y_{ijt}^*) \;=\; \begin{cases} \mathcal{TN}\left(m_{ijt}, 1, 0, \infty\right) & \text{if } y_{ijt} = 1 \\ \mathcal{TN}\left(m_{ijt}, 1, -\infty, 0\right) & \text{if } y_{ijt} = 0 \end{cases} \tag{39}$$

with $m_{ijt} = \mathbb{E}(\tilde{\mathbf{x}}_{it})^\top \mathbb{E}(\tilde{\boldsymbol{\beta}}_{jt})$. Then, the updated mean of y_{ijt}^* is given by,

$$\mathbb{E}(y_{ijt}^*) \;=\; \begin{cases} m_{ijt} + \frac{\phi(m_{ijt})}{\Phi(m_{ijt})} & \text{if } y_{ijt} = 1 \\ m_{ijt} - \frac{\phi(m_{ijt})}{1-\Phi(m_{ijt})} & \text{if } y_{ijt} = 0 \end{cases} \tag{40}$$

For abstention (i.e., missing y_{ijt}), we set $q(y_{ijt}^*) = \mathcal{N}(m_{ijt}, 1)$ and $\mathbb{E}(y_{ijt}^*) = m_{ijt}$.

Second, the variational distribution for $\tilde{\boldsymbol{\beta}}$ is given by,

$$q(\tilde{\boldsymbol{\beta}}_{jt}) \;=\; \mathcal{N}(\mathbf{B}_{jt}^{-1}\mathbf{b}_{jt},\; \mathbf{B}_{jt}^{-1}) \tag{41}$$

where $\mathbf{b}_{jt} = \Sigma_{\tilde{\boldsymbol{\beta}}}^{-1}\boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}} + \sum_{i\in\mathcal{I}_t} \mathbb{E}(\tilde{\mathbf{x}}_{it})\mathbb{E}(y_{ijt}^*)$ and $\mathbf{B}_{jt} = \Sigma_{\tilde{\boldsymbol{\beta}}}^{-1} + \sum_{i\in\mathcal{I}_t} \mathbb{E}(\tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}_{it}^\top)$ with $\mathcal{I}_t = \{i : \underline{T}_i \leq t \leq \overline{T}_i\}$. Note that the summation is taken over $\mathcal{I}_t$, the set of legislators who are present at time t.

Finally, we consider the variational distribution of dynamic ideal points. Here, we rely on the forward-backward algorithm derived for the variational Kalman filtering. Specifically, we first use the forward recursion to compute,

$$x_{it} \mid \ddot{y}_{i1}, \ldots, \ddot{y}_{it} \;\overset{\text{indep.}}{\sim}\; \mathcal{N}(c_{it},\; C_{it}) \tag{42}$$

where $\ddot{\beta}_t = \sqrt{\sum_{j=1}^{J_t} \mathbb{E}(\beta_{jt}^2)}$, $\ddot{y}_{it} \;=\; \{\sum_{j=1}^{J_t} \mathbb{E}(y_{ijt}^*)\mathbb{E}(\beta_{jt}) - \mathbb{E}(\beta_{jt}\alpha_{jt})\}/\ddot{\beta}_t$, $c_{it} = c_{i,t-1} + K_t(\ddot{y}_{it} - \ddot{\beta}_t c_{i,t-1})$ and $C_{it} = (1 - K_t\ddot{\beta}_t)\Omega_t$ with $\Omega_t = \omega_x^2 + C_{i,t-1}$, $K_t = \ddot{\beta}_t\Omega_t/S_t$ and $S_t = \ddot{\beta}_t^2\Omega_t + 1$. We recursively

compute these quantities by setting $c_{i0} = \mu_x$ and $C_{i0} = \Sigma_x$. Then, combined with the backward recursion, we can derive the following variational distribution,

$$x_{it} \mid \ddot{y}_{i\underline{T}_i}, \ldots, \ddot{y}_{i\overline{T}_i} \overset{\text{indep.}}{\sim} \mathcal{N}(d_{it},\ D_{it}) \tag{43}$$

where $d_{it} = c_t + J_t(d_{t+1} - c_{it})$ and $D_{it} = C_{it} + J_t^2(D_{t+1} - \Omega_{t+1})$ with $J_t = C_{it}/\Omega_{t+1}$. The recursive computation is done by setting $d_{i\overline{T}_i} = c_{i\overline{T}_i}$ and $D_{i\overline{T}_i} = C_{i\overline{T}_i}$. Thus, the required first and second moments of x_{it} can be easily obtained.

4.3 An Empirical Application

We apply the proposed variational EM algorithm for estimating the dynamic ideal point to the voting data from the U.S. Supreme court (from October 1937 to October 2013). The data set includes 5,164 votes on court cases by 45 distinct justices over 77 terms, resulting in the estimation of 697 unique ideal points for all justice-term combinations. The same data set was used to compute the ideal point estimates published as the well-known Martin-Quinn scores at `http://mqscores.berkeley.edu/` (July 23, 2014 Release version).

We set the prior parameters using the replication code, which was directly obtained from the authors. In particular, the key random-walk prior variance parameter ω_x^2 is set to be equal to 0.1 for all justices. Note that this choice differs from the specification in Martin and Quinn (2002) where Douglas' prior variance parameter was set as $\omega_x^2 = 0.001$ because of his ideological extremity and the small number of cases he heard towards the end of his career. This means that Douglas's ideal point estimate is fixed at his prior mean of -3.0, but in the results we report below this constraint is not imposed.

We use the same prior specification and apply the proposed variational EM algorithm as well as the standard MCMC algorithm implemented via the `MCMCdynamicIRT1d()` function from `MCMCpack`. For the MCMC algorithm, using the replication code, 1.2 million iterations took just over 5 days of computing time. In contrast, our variational EM algorithm took only four seconds. To obtain a measure of estimation uncertainty, we use the parametric bootstrap approach of Lewis and Poole (2004) to create 100 replicates and construct bias-corrected 95% bootstrap confidence intervals. Note that even with this bootstrap procedure, the computation is done within several minutes.

We begin by examining, for each term, the correlation of the resulting estimated ideal points for nine justices between the proposed variational inference algorithm and MCMC algorithm. Figure 7 presents both Pearson's correlations and Spearman's rank-order correlations. Overall, the correlations are high, exceeding 95% in most cases. In particular, for many terms, rank-order correlations

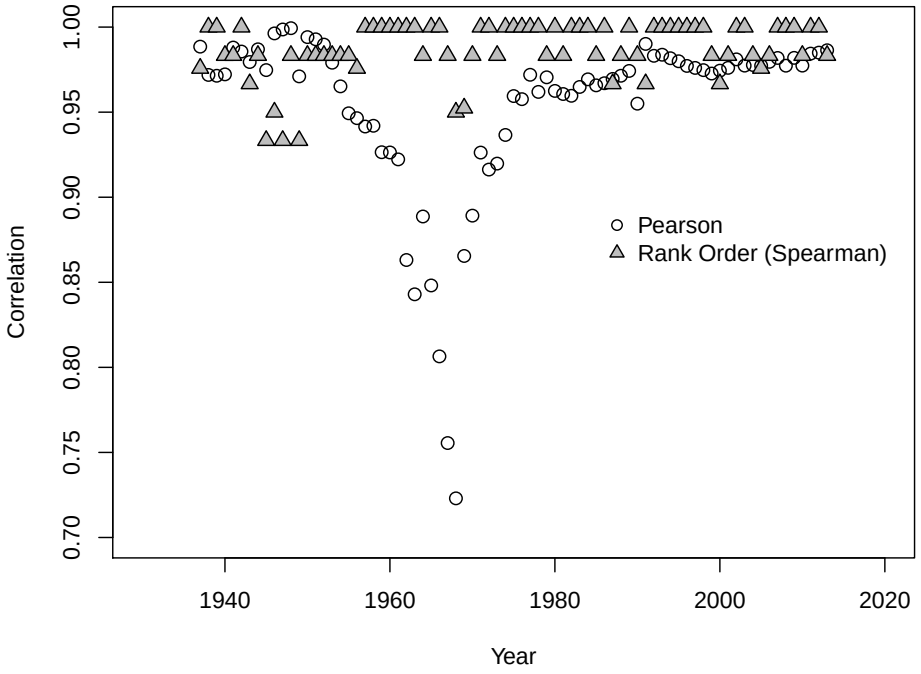

Figure 7: Correlation of the Estimated Ideal Points for each Term between the Variational EM and Markov chain Monte Carlo (MCMC) Algorithms. Open circles indicate Pearson correlations, while grey triangles represent Spearman's rank-order correlations. Overall, the correlations are high, exceeding 95% in most cases. The poorer Pearson correlations around 1969 are driven largely by Douglas' ideological extremity (see Figure 8).

are equal to unity, indicating that the two algorithms produce justices' estimated ideal points whose rank-order is identical. We note that a significant drop in Pearson correlation between 1960 and 1975 is driven almost entirely by the extreme MCMC estimates of Douglas' position in these years, which correspond to the final years of his tenure. And yet, even in these years, the rank-order correlations remain high.

Figures 8 present time series plots of the estimated ideal points for the 16 justices who served the longest periods of time in our study. Solid red lines indicate the variational estimates, while the dashed lines indicate their 95% confidence intervals based on the parametric bootstrap. The grey polygons represent the 95% credible intervals obtained from the MCMC algorithm. For almost all justices, the movement of estimated ideal points over time is similar between the two algorithms. Indeed, for most justices, the correlation between the two sets of estimates is high, often exceeding 95%. A notable exception to this is Douglas, where we see his ideal point estimates based on the MCMC algorithm becomes more extreme as time passes. The behavior observed here is consistent with earlier warnings about Douglas' ideological extremity and the fact that he cast only a small number of votes in the final years of his career (Martin and Quinn, 2002). The correlation across all

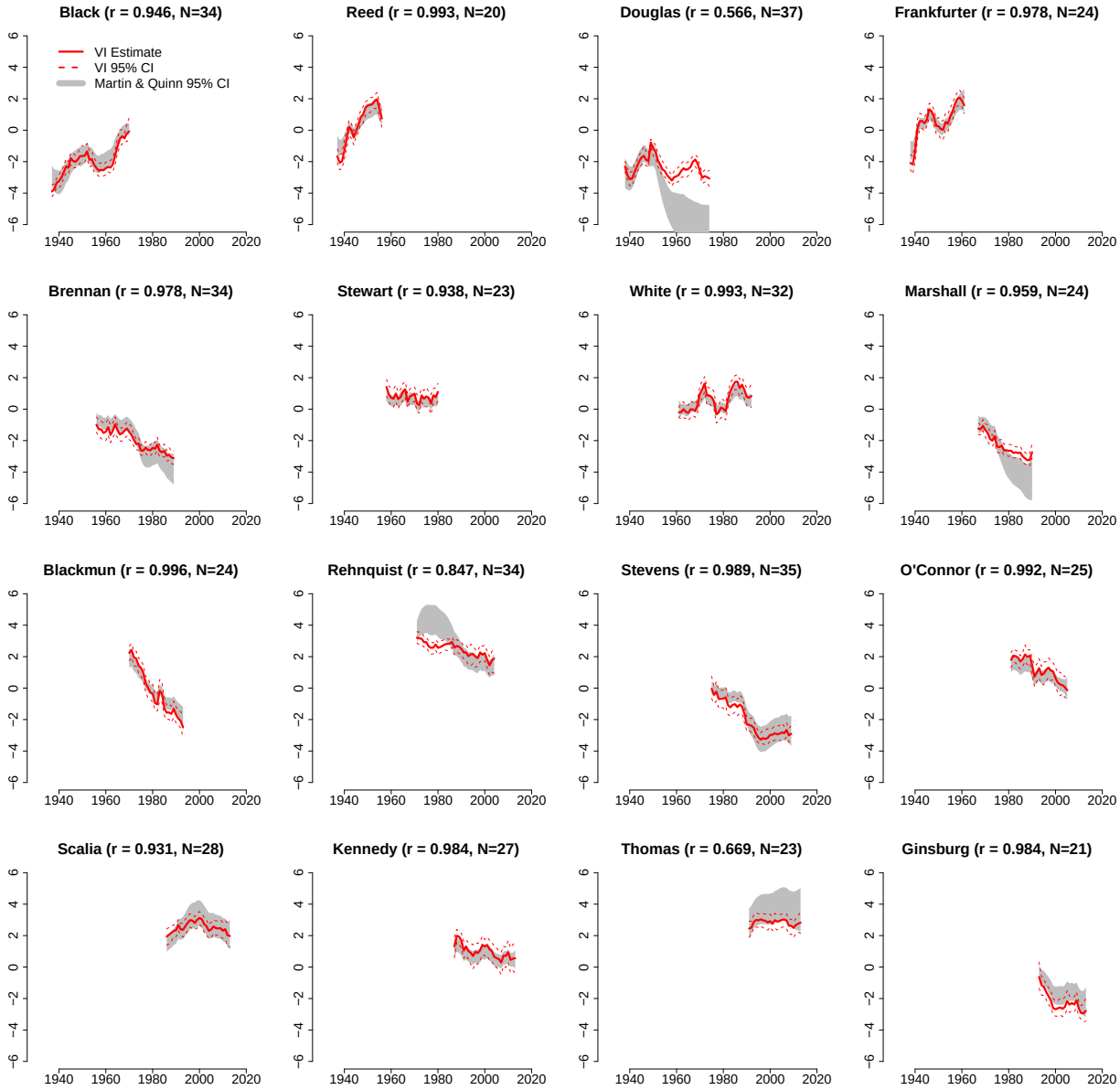

Figure 8: Ideal Point Estimates for 16 Longest-serving Justices based on the Variational Inference (VI) and Markov chain Monte Carlo (MCMC) Algorithm. The VI point estimates are indicated by solid red lines while the dashed lines indicate its 95% confidence intervals based on the parametric bootstrap. We also present the 95% Bayesian confidence intervals as grey polygons. The horizontal axis indicates year and the vertical axis indicates estimated ideal points. For each justice, we also compute the Pearson's correlation between the two sets of the estimates. Overall, the correlations between the two sets of estimates are high except Douglas who is ideologically extreme and have only a small number of votes in the final years of his career.

ideal points between the two sets of estimates increases from 0.93 to 0.96, once we exclude Douglas. Overall, our proposed variational inference algorithm produces the estimates of ideal points that are close to the Martin-Quinn score but with a significantly less amount of computing time.

4.4 Simulation Evidence

We further demonstrate the computational scalability of the proposed variational EM algorithm through a series of simulations. We generate a number of roll call matrices that vary in size. These include roll call matrices that have $N = 10$ legislators and $J = 100$ roll calls per session (roughly corresponding to the size of the U.S. Supreme Court), roll call matrices with $N = 100$ and $J = 500$ (roughly corresponding to the size of the U.S. Senate), and roll call matrices with $N = 500$ and $J = 1,000$ (roughly corresponding in size to the U.S. House). We also vary the total number of sessions, ranging from $T = 10$ to $T = 100$. Thus, the largest rollcall matrix represents a scenario that all members of US House vote on 1,000 bills during each of 100 consecutive sessions! As we show next, even in this extreme case, our algorithm runs in about 25 minutes, yielding the estimated ideal points that are close to the true values.

We then apply our variational EM algorithm and record the amount of time needed to estimate the model, as well as correlation between the true and recovered ideal points. In the simulation, all legislators serve throughout all periods, whose ideal points in the first period follow the standard normal distribution. Independence across legislators is assumed as done in the model, and their subsequent ideal points are generated as a random walk with $\omega_x^2 = 0.1$ for all legislators. Item difficulty and discrimination parameters in all sessions were drawn from uniform $(-1.5, 1.5)$ and $(-5.5, 5.5)$ distributions respectively. While parallelization of the algorithm is trivial and would further reduce runtimes, we do not implement it for this calculation. As before, convergence is assumed to be achieved when correlation across all parameters across consecutive iterations is greater than $1 - 10^{-6}$.[4]

The left panel of Figure 9 displays the amount of time needed for each simulation, with the total number of sessions T given on the horizontal axis. As a benchmark comparison, MCMC replication code provided by Martin and Quinn (2002) took over five days to estimate ideal points for US supreme court justices over 77 years ($N = 45$, $T = 77$, and $J = 5,164$). For the scenario with $N = 10$ legislators and $J = 100$ roll calls per session, estimation is completed under a minute regardless of the number of sessions. Similarly, for the scenarios with 100 legislators and 500 roll calls per session, computation is completed in a matter of minutes regardless of the number of sessions. Computation only begins to significantly increase with our largest scenario of 500 legislators and $1,000$ roll calls per session. But even here, for 100 sessions, the variational EM algorithm converges

[4]To reduce Monte Carlo error, estimates for cases where $N = 10$ are repeated 25 times, with median runtimes and correlations reported in the figure.

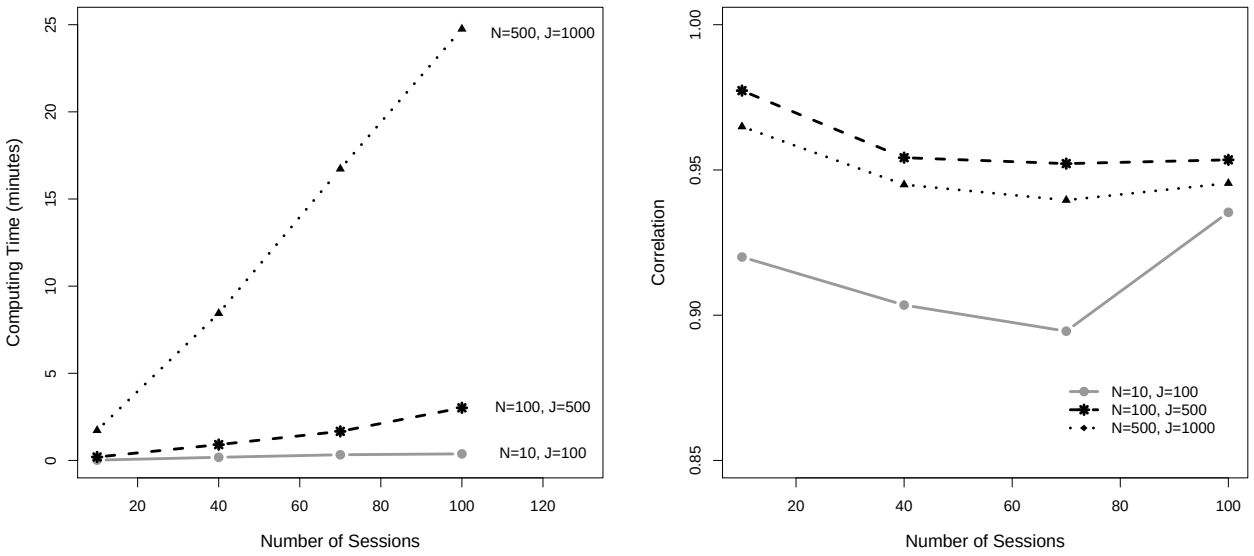

Figure 9: Scalability and Accuracy of the Proposed Variational Inference for the Dynamic Ideal Point Model. The left panel presents runtimes of the proposed variational EM algorithm for fitting the dynamic ideal point model. We consider three different simulation scenarios where the number of legislators N varies from 10 to 500 and the number of roll calls per session J ranges from 100 to 1,000. The number of sessions T is shown on the horizontal axis, with all N legislators assumed to vote on all J bills in every session. The vertical axis indicates the time necessary to fit the dynamic ideal point model for each data set through the proposed algorithm. Even with the largest data set we consider ($N = 500$, $J = 1,000$, and $T = 100$), the algorithm can estimate a half million ideal points in about two hours. The right panel shows the (Pearson) correlation between the estimated ideal points and their true values. In almost all cases, the correlation exceeds 0.95.

in approximately 25 minutes.

The right panel of the figure presents, for each simulation scenario, the correlation between the variational estimates of ideal points and their true values across all legislators and sessions. The plot demonstrates that the correlation exceeds 95% throughout all the simulations except the case where the size of roll call matrix is small. Even in this case, the correlation is about 90%, which suggests the reasonable accuracy of the variational estimates under the dynamic ideal point model.

5 Hierarchical Ideal Point Model

Finally, we consider the hierarchical ideal point model where the ideal points are modeled as a linear function of covariates (Bafumi et al., 2005). Like the dynamic ideal point model, there is no closed-form EM algorithm that directly maximizes the posterior distribution. Therefore, we apply variational inference to approximate the posterior distribution. We derive the variational EM algorithm and demonstrate its computational efficiency and the accuracy of approximation through empirical and simulation studies.

5.1 The Model

Let each district vote denoted by a binary random variable y_l where there exist a total of L such votes, i.e., $\ell \in \{1, \ldots, L\}$. Each vote y_ℓ represents a vote cast by legislator $i[\ell]$ on bill $j[\ell]$ ($y_\ell = 1$ and $y_\ell = 0$ representing 'yea' and 'nay,' respectively) where $i[\ell] \in \{1, \ldots, N\}$ and $j[\ell] \in \{1, \ldots, J\}$. Thus, there are a total of N legislators and J bills. Finally, let $g[i[\ell]]$ represent the group membership of legislator $i[\ell]$ where $g[i[\ell]] \in \{1, \ldots, G\}$ and G indicates the total number of groups.

The hierarchical model we consider has the following latent variable structure with the observed vote written as $y_\ell = \mathbf{1}\{y_\ell^* > 0\}$ as before,

$$y_\ell^* = \alpha_{j[\ell]} + \beta_{j[\ell]} x_{i[\ell]} + \epsilon_\ell \quad \text{where} \quad \epsilon_\ell \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1) \tag{44}$$

$$x_{i[\ell]} = \boldsymbol{\gamma}_{g[i[\ell]]}^\top \mathbf{z}_{i[\ell]} + \eta_{i[\ell]} \quad \text{where} \quad \eta_{i[\ell]} \overset{\text{indep.}}{\sim} \mathcal{N}(0, \sigma^2_{g[i[\ell]]}) \tag{45}$$

where $\boldsymbol{\gamma}_{g[i[\ell]]}$ is an M dimensional vector of group-specific coefficients, $\mathbf{z}_{i[\ell]}$ is an M dimensional vector of legislator-specific covariates, which typically includes one for an intercept, and $\sigma^2_{g[i[\ell]]}$ is the group-specific variance.

One important special case of this model is a dynamic ideal point model with a parametric time trend, an approach used to compute the DW-NOMINATE score (Poole and Rosenthal, 1997) and adopted by some scholars (e.g., Bailey, 2007). In this case, the $i[\ell]$ represents a legislator-session, e.g., John Kerry in 2014, and $g[i[\ell]]$ indicates the legislator, John Kerry, whereas $\mathbf{z}_{i[\ell]}$ may include the number of sessions since the legislator took office as well as one for an intercept. Then, the ideal points are modeled with a linear time trend. In addition, including the square term will allow one to model ideal points using a quadratic time trend. Note that in this setting the time trend is estimated separately for each legislator.

The model is completed with the following conjugate prior distribution,

$$\tilde{\boldsymbol{\beta}}_{j[\ell]} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}},\ \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}) \tag{46}$$

$$\boldsymbol{\gamma}_{g[i[\ell]]} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\gamma}},\ \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}) \tag{47}$$

$$\sigma^2_{g[i[\ell]]} \overset{\text{i.i.d.}}{\sim} \mathcal{IG}\left(\frac{\nu_\sigma}{2}, \frac{s^2_\sigma}{2}\right) \tag{48}$$

where $\tilde{\boldsymbol{\beta}}_{j[\ell]} = (\alpha_{j[\ell]}, \boldsymbol{\beta}_{j[\ell]})$ and $\mathcal{IG}(\nu, s^2)$ represents the inverse-gamma distribution with scale and shape parameters equal to ν and s^2, respectively.

It is convenient to rewrite the model in the following reduced form,

$$y_\ell^* = \alpha_{j[\ell]} + \beta_{j[\ell]} \boldsymbol{\gamma}_{g[i[\ell]]}^\top \mathbf{z}_{i[\ell]} + \beta_{j[\ell]} \eta_{i[\ell]} + \epsilon_\ell \tag{49}$$

Then, the joint posterior distribution is given by,

$$
\begin{aligned}
&p(\mathbf{Y}^*, \{\tilde{\boldsymbol{\beta}}_k\}_{k=1}^J, \{\boldsymbol{\gamma}_m\}_{m=1}^G, \{\eta_n\}_{n=1}^N \mid \mathbf{Y}) \\
&\propto \prod_{\ell=1}^{L}\prod_{k=1}^{J}\prod_{n=1}^{N}\prod_{m=1}^{G} \left(\mathbf{1}\{y_{ij}^* \geq 0, y_{ij} = 1\} + \mathbf{1}\{y_{ij}^* < 0, y_{ij} = 0\}\right) \phi_1(y_\ell^*; \alpha_k + \beta_k \boldsymbol{\gamma}_m^\top \mathbf{z}_n + \beta_k \eta_n, 1)^{\mathbf{1}\{j[\ell]=k, i[\ell]=n, g[i[\ell]]=m\}} \\
&\quad \times \prod_{k=1}^{J} \phi_2(\tilde{\boldsymbol{\beta}}_k; \boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}) \prod_{n=1}^{N}\prod_{m=1}^{G} \phi_1(\eta_n; 0, \sigma_m^2)^{\mathbf{1}\{g[n]=m\}} \prod_{m=1}^{G} \mathcal{IG}\left(\sigma_m^2; \frac{\nu_\sigma}{2}, \frac{s_\sigma^2}{2}\right)
\end{aligned} \tag{50}
$$

5.2 The Proposed Algorithm

For this hierarchical model, there is no closed-form EM algorithm that can directly maximize the posterior distribution given in equation (50). Therefore, as done in the case of the dynamic model, we seek for the variational approximation. The factorization assumption we invoke is given by the following,

$$
q(\mathbf{Y}^*, \{\tilde{\boldsymbol{\beta}}_k\}_{k=1}^J, \{\boldsymbol{\gamma}_m, \sigma_m^2\}_{m=1}^G, \{\eta_n\}_{n=1}^N) \;=\; \prod_{\ell=1}^{L} q(y_\ell^*) \prod_{k=1}^{J} q(\tilde{\boldsymbol{\beta}}_k) \prod_{m=1}^{G} q(\boldsymbol{\gamma}_m) q(\sigma_m^2) \prod_{n=1}^{N} q(\eta_n) \tag{51}
$$

Under this factorization assumption, we can derive the variational EM algorithm that approximates the joint posterior distribution by maximizing the lower bound. Note that aside from the factorization assumption no additional assumption is made to derive the proposed algorithm. The detailed derivation is given in Appendix D and we summarize the results below.

The proposed EM algorithm cycles through the following updating steps until convergence. First, we update the variational distribution for the latent propensities y_ℓ^* for all $\ell = 1, \ldots, L$.

$$
q(y_\ell^*) \;=\; \begin{cases} \mathcal{TN}(m_\ell, 1, -\infty, 0) & \text{if} \quad y_\ell = 1 \\ \mathcal{TN}(m_\ell, 1, 0, \infty,) & \text{if} \quad y_\ell = 0 \\ \mathcal{N}(m_\ell, 1) & \text{if} \quad y_\ell \text{ is missing} \end{cases} \tag{52}
$$

where $m_\ell = \mathbb{E}(\alpha_{j[\ell]}) + \mathbb{E}(\beta_{j[l]})\mathbb{E}(\boldsymbol{\gamma}_{g[i[\ell]]})^\top \mathbf{z}_{i[\ell]} + \mathbb{E}(\eta_{i[\ell]})\mathbb{E}(\beta_{j[\ell]})$. The required moment update step is given by,

$$
\mathbb{E}(y_\ell^*) \;=\; \begin{cases} m_\ell + \frac{\phi(m_\ell)}{\Phi(m_\ell)} & \text{if} \quad y_\ell = 1 \\ m_\ell - \frac{\phi(m_\ell)}{1-\Phi(m_\ell)} & \text{if} \quad y_\ell = 0 \\ m_\ell & \text{if} \quad y_\ell \text{ is missing} \end{cases} \tag{53}
$$

Next, we update the first and second moments of the ideal point error term η_n using the following variational distribution,

$$
q(\eta_n) \;=\; \mathcal{N}(A_n^{-1} a_n,\; A_n^{-1}) \tag{54}
$$

where $A_n = \mathbb{E}(\sigma^{-2}_{g[n]}) + \sum_{\ell=1}^{L} \mathbf{1}\{i[\ell] = n\}\mathbb{E}(\beta^2_{j[\ell]})$ and $a_n = \sum_{\ell=1}^{L} \mathbf{1}\{i[\ell] = n\}\{\mathbb{E}(y^*_\ell)\mathbb{E}(\beta_{j[\ell]}) - \mathbb{E}(\alpha_{j[\ell]}\beta_{j[\ell]}) - \mathbb{E}(\beta^2_{j[\ell]})\mathbb{E}(\boldsymbol{\gamma}_{g[n]})^\top \mathbf{z}_n\}$. Thus, the required moments are given by $\mathbb{E}(\eta_n) = A_n^{-1}a_n$ and $\mathbb{E}(\eta_n^2) = A_n^{-1} + (A_n^{-1}a_n)^2$.

Third, we derive the variational distribution for the item parameters. This distribution is equal to,

$$q(\tilde{\boldsymbol{\beta}}_k) \quad = \quad \mathcal{N}(\mathbf{B}_k^{-1}\mathbf{b}_k, \mathbf{B}_k^{-1}) \tag{55}$$

where $\mathbf{B}_k = \boldsymbol{\Sigma}^{-1}_{\tilde{\boldsymbol{\beta}}} + \sum_{\ell=1}^{L} \mathbf{1}\{j[\ell] = k\}\mathbb{E}(\tilde{\mathbf{x}}_{i[\ell]}\tilde{\mathbf{x}}^\top_{i[\ell]})$ and $\mathbf{b}_k = \boldsymbol{\Sigma}^{-1}_{\tilde{\boldsymbol{\beta}}}\boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}} + \sum_{\ell=1}^{L} \mathbf{1}\{j[\ell] = k\}\mathbb{E}(y^*_\ell)\mathbb{E}(\tilde{\mathbf{x}}_{i[\ell]})$ with $\tilde{\mathbf{x}}_{i[\ell]} = (1, \boldsymbol{\gamma}^\top_{g[i[\ell]]}\mathbf{z}_{i[\ell]} + \eta_{i[\ell]})^\top$. We note that $\mathbb{E}(\tilde{\mathbf{x}}_{i[\ell]}) = (1, \mathbb{E}(\boldsymbol{\gamma}_{g[i[\ell]]})^\top \mathbf{z}_{i[\ell]} + \mathbb{E}(\eta_{i[\ell]}))^\top$ and

$$\mathbb{E}\left(\tilde{\mathbf{x}}_{i[\ell]}\tilde{\mathbf{x}}^\top_{i[\ell]}\right) = \begin{pmatrix} 1 & \mathbb{E}(\boldsymbol{\gamma}_{g[i[\ell]]})^\top \mathbf{z}_{i[\ell]} + \mathbb{E}(\eta_{i[\ell]}) \\ \mathbb{E}(\boldsymbol{\gamma}_{g[i[\ell]]})^\top \mathbf{z}_{i[\ell]} + \mathbb{E}(\eta_{i[\ell]}) & \mathbf{z}^\top_{i[\ell]}\mathbb{E}(\boldsymbol{\gamma}_{g[i[\ell]]}\boldsymbol{\gamma}^\top_{g[i[\ell]]})\mathbf{z}_{i[\ell]} + 2\mathbb{E}(\boldsymbol{\gamma}_{g[i[\ell]]})^\top \mathbf{z}_{i[\ell]}\mathbb{E}(\eta_{i[\ell]}) + \mathbb{E}(\eta^2_{i[\ell]}) \end{pmatrix}. \tag{56}$$

This gives the required moment update, $\mathbb{E}(\tilde{\boldsymbol{\beta}}_k) = \mathbf{B}_k^{-1}\mathbf{b}_k$.

Fourth, the variational distribution for the group-level coefficients is given by,

$$q(\boldsymbol{\gamma}_m) \quad = \quad \mathcal{N}(\mathbf{C}_m^{-1}\mathbf{c}_m,\ \mathbf{C}_m^{-1}) \tag{57}$$

where $\mathbf{C}_m = \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\gamma}} + \sum_{\ell=1}^{L} 1\{g[i[\ell]] = m\}\mathbb{E}(\beta^2_{j[\ell]})\mathbf{z}_{i[\ell]}\mathbf{z}^\top_{i[\ell]}$ and $\mathbf{c}_m = \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\gamma}}\boldsymbol{\mu}_{\boldsymbol{\gamma}} + \sum_{\ell=1}^{L} 1\{g[i[\ell]] = m\}\mathbf{z}_{i[\ell]}[\mathbb{E}(\beta_{j[\ell]})\{\mathbb{E}(y^*_\ell) - \mathbb{E}(\alpha_{j[\ell]})\} - \mathbb{E}(\beta^2_{j[\ell]})\mathbb{E}(\eta_{i[\ell]})]$. Thus, the required moment updates are given by $\mathbb{E}(\boldsymbol{\gamma}_m) = \mathbf{C}_m^{-1}\mathbf{c}_m$ and $\mathbb{E}(\boldsymbol{\gamma}_m\boldsymbol{\gamma}_m^\top) = \mathbf{C}_m^{-1} + \mathbf{C}_m^{-1}\mathbf{c}_m\mathbf{c}_m^\top\mathbf{C}_m^{-1}$.

Finally, we derive the variational distribution for the group-level variance parameters. This distribution is equal to,

$$q(\sigma^2_m) \quad = \quad \mathcal{IG}\left(\frac{\nu_\sigma + \sum_{n=1}^{N} \mathbf{1}\{g[n] = m\}}{2},\ \frac{1}{2}\left(s^2_\sigma + \sum_{n=1}^{N} \mathbf{1}\{g[n] = m\}\mathbb{E}(\eta^2_n)\right)\right) \tag{58}$$

where the desired moment update is given by $\mathbb{E}(\sigma^{-2}_m) = [\nu_\sigma + \sum_{n=1}^{N} \mathbf{1}\{g[n] = m\}]/[s^2_\sigma + \sum_{n=1}^{N} \mathbf{1}\{g[n] = m\}\mathbb{E}(\eta^2_n)]$. These updating steps are repeated until convergence.

5.3 Simulation Evidence

We conduct a simulation study to demonstrate the computational scalability and accuracy of the proposed variational EM algorithm. To do this, we generate roll call matrices that vary in size following the simulation study for dynamic models in Section 4.4 where the number of legislators is now replaced by the number of groups G instead. Each group has N different ideal points to be estimated, and three covariates $\mathbf{z}_{i[\ell]}$ are observed for each ideal point, i.e., $M = 3$. Finally, we

construct the simulation such that each group votes on the same set of J bills but within each group different members vote on different subsets of the bills.

The intercepts for ideal points follow another uniform distribution with $(-1, 1)$, while item difficulty and discrimination parameters were both drawn uniformly from $(-1.5, 1.5)$. The group-level variance parameters $\sigma^2_{g[i[\ell]]}$ were set to 0.01 for all groups. We use diffuse priors for item difficulty and discrimination parameters as well as for group-level coefficients. Specifically, the prior distribution for these parameters is the independent normal distribution with a mean of zero and a standard deviation of five. For group-level variance parameters, we use a semi-informative prior such that they follow the inverse-gamma distribution with $\nu_0 = 2$ and $s^2 = 0.02$.

When compared to the other models considered in this paper, we find the hierarchical model to be computationally more demanding. To partially address this issue, we parallelize the algorithm wherever possible and implement this parallelized code using eight cores through OpenMP in this simulation study. We also use a slightly less stringent convergence criteria than in the other cases where we check whether the correlations for bill parameters and group-level coefficients across their consecutive iterations is greater than $1 - 10^{-4}$. We find that applying a stricter convergence criteria does not significantly improve the quality of the resulting estimates.

We consider three different sets of simulation scenarios where the number of groups G varies from 10 to 500 and the number of bills (per group) J ranges from 100 to 1,000. Figure 10 shows the results. In the left plot, the vertical axis represents the runtime of our algorithm in hours, while the horizontal axis shows the size of each group N, i.e., the number of ideal points to be estimated per group. Our variational EM algorithm scales well to a large data set. In the largest data set we consider ($N = 100$, $J = 1,000$, and $G = 500$), for example, the proposed algorithm can estimate hundred thousand ideal points in only about 22 hours.

In the right plot of Figure 10, we plot the correlation between the estimated ideal points and their true values for each simulation scenario.[5] The quality of estimates appear to depend on the number of groups with the simulations with a larger number of groups yielding almost perfect correlation. When the number of groups is small, however, we find that the correlations are weaker and the results are highly dependent on prior specification. This is a well-known feature of Bayesian hierarchical models (Gelman, 2006) and the ideal point models appear to be no exception in this regard.

[5]To reduce Monte Carlo error, estimates for cases where $N = 10$ are repeated 25 times, with median runtimes and correlations reported in the figure.

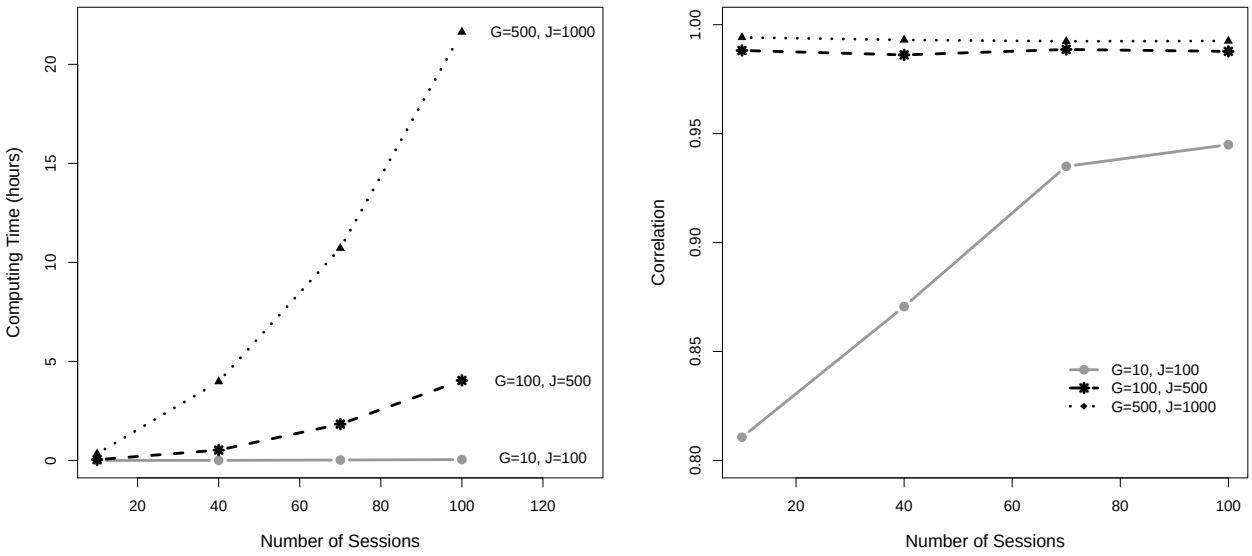

Figure 10: Scalability and Accuracy of the Proposed Variational Inference for the Hierarchical Ideal Point Model. The left panel presents runtimes of the proposed variational EM algorithm for fitting the hierarchical ideal point model. We consider three different simulation scenarios where the number of groups G varies from 10 to 500 and the number of bills (per group) J ranges from 100 to 1,000. The number of ideal points to be estimated (per group) N is shown on the horizontal axis, with all G groups assumed to vote on all J bills but within each group different legislators vote on different subsets of the bills. In the largest data set we consider ($N = 100$, $J = 1,000$, and $G = 500$), our algorithm can estimate hundred thousand ideal points in about 22 hours. The right panel shows the (Pearson) correlation between the estimated ideal points and their true values.

5.4 An Empirical Illustration

As noted earlier, DW-NOMINATE scores adopt a linear time trend model for legislators. A model essentially equivalent to this model can be estimated as a special case of our general hierarchical model, in which the covariate $z_{i[l]}$ is the term served by a particular legislator and the ideal point noise parameter $\eta_{i[l]}$ is fixed at 0.[6] We analyze the roll call data from the 1st – 100th U.S. House and show empirically that the proposed variational EM algorithm for this model produces the ideal point estimates essentially similar to DW-NOMINATE scores. We specify the prior parameters as $\nu_\sigma = 10^8$ and $s_\sigma^2 = 10^{-8}$, which effectively fix the noise parameter as desired, and use the same starting values as those used in DW-NOMINATE. An additional constraint we impose that is consistent with DW-NOMINATE is that legislators who serve less than four terms do not shift ideal points over time.

Our model includes $G = 10,474$ groups (i.e. legislators) with $I = 36,177$ different ideal points, estimated using $J = 48,381$ bills. Estimation of the model using 8 threads required just under

[6]There are other minor differences relating to different utility functions (Carroll, Lewis, Lo, Poole and Rosenthal, 2009; Carroll et al., 2013).

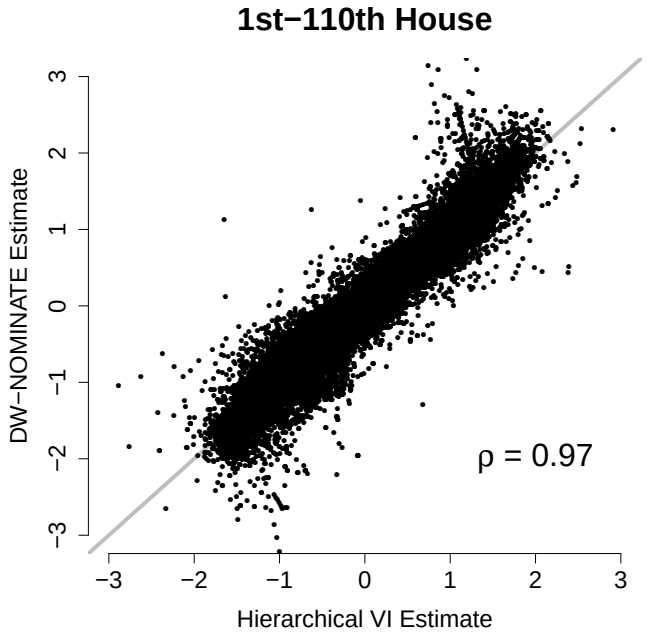

Figure 11: Correlation between DW-NOMINATE Estimates and the Proposed Hierarchical Ideal Point Estimates for the 1st – 110th Congress. These ideal point estimates are quite similar, with the correlation reaching 0.97.

5 hours of computing time. This runtime could be considerably reduced, for example, by not updating $\eta_{i[\ell]}$ and σ_m^{-2}, which are fixed at zero. Figure 11 shows the estimated ideal points from the hierarchical model, plotted against the corresponding DW-NOMINATE estimates. The two sets of ideal points correlate at 0.97, thus validating the ability of the hierarchical model to reproduce DW-NOMINATE's linear time trend ideal point model.

6 Concluding Remarks

In this paper, we develop fast estimation algorithms for ideal points with massive data. Over the last decade, political science, like other social science disciplines, witnessed the "big data revolution" where empirical researchers are collecting increasingly large data sets of diverse types. When analyzing these data, traditional statistical methods are often ill-suited due to their computational inefficiency. While the available computational power is steadily increasing, the amount of data available to social scientists is growing even at a faster rate. As a result, researchers are unable to estimate the models of their choice within a reasonable amount of time and are often forced to make a compromise by adopting a feasible and yet undesirable statistical procedure.

We demonstrate, in the specific context of ideal point estimation, how to overcome this com-

putational bottleneck by developing new fast estimation algorithms. Specifically, we develop the Expectation-Maximization (EM) algorithms that maximize the posterior distribution. When such an algorithm is not available in a closed-form, we derive a variational EM algorithm that approximates posterior inference. Through empirical and simulation studies, we show that the proposed methodology improves the computational efficiency by orders of magnitude without sacrificing the statistical properties of the resulting estimates. With this new methodology, researchers can estimate ideal points from massive data on their laptop within minutes rather than running other estimation algorithms for days on a high-performance computing cluster.

We predict that this line of methodological research will become essential for the next generation of empirical political science research. The political science data now come in a variety of form – textual data, network data, and spatial-temporal data to name a few – and in a large quantity. To efficiently extract useful information from these data will require the development of scalable statistical estimation techniques like the ones proposed in this paper.

References

Bafumi, Joseph, Andrew Gelman, David K. Park and Noah Kaplan. 2005. "Practical Issues in Implementing and Understanding Bayesian Ideal Point Estimation." *Political Analysis* 13:171–187.

Bafumi, Joseph and Michael Herron. 2010. "Leapfrog Representation and Extremism: A Study of American Voters and Their Members in Congress." *American Political Science Review* 104:519–542.

Bailey, Michael. 2007. "Comparable Preferences across Time and Institutions for the Court, Congress, and Presidency." *American Journal of Political Science* 51:433–448.

Bailey, Michael A. 2013. "Is Today's Court the Most Conservative in Sixty Years? Challenges and Opportunities in Measuring Judicial Preferences." *Journal of Politics* 75:821–834.

Bailey, Michael A., Brian Kamoie and Forrest Maltzman. 2005. "Signals from the Tenth Justice: The Political Role of the Solicitor General in the Supreme Court Decision Making." *American Journal of Political Science* 49:72–85.

Bailey, Michael and Kelly H. Chang. 2001. "Comparing Presidents, Senators, and Justices: Interinstitutional Preference Estimation." *The Journal of Law, Economics, and Organization* 17:477–506.

Battista, James Coleman, Michael Peress and Jesse Richman. 2013. "Common-Space IdealPoints, Committee Assignments, and Financial Interests in the State Legislatures." *State Politics & Policy Quarterly* 13:70–87.

Bock, R. Darrell and Murray Aitkin. 1981. "Marginal maximum likelihood estimation of item parameters: Application of an EM Algorithm." *Psychometrika* 46:443–459.

Bonica, Adam. 2014. "Mapping the Ideological Marketplace." *American Journal of Political Science* 58:367–387.

Carroll, Royce, Jeffrey B. Lewis, James Lo and Keith T. Poole. 2009. "Measuring Bias and Uncertainty in DW-NOMINATE Ideal Point Estimates via the Parametric Bootstrap." *Political Analysis* 17:261–275.

Carroll, Royce, Jeffrey B Lewis, James Lo, Keith T Poole and Howard Rosenthal. 2009. "Comparing NOMINATE and IDEAL: Points of difference and Monte Carlo tests." *Legislative Studies Quarterly* 34:555–591.

Carroll, Royce, Jeffrey B Lewis, James Lo, Keith T Poole and Howard Rosenthal. 2013. "The Structure of Utility in Spatial Models of Voting." *American Journal of Political Science* 57:1008–1028.

Clinton, Joshua D. and Adam Meirowitz. 2003. "Integrating Voting Theory andRoll Call Analysis: A Framework." *Political Analysis* 11:381–396.

Clinton, Joshua D., Anthony Bertelli, Christian R. Grose, David E. Lewis and David C. Nixon. 2012. "Separated Powers in the United States: The Ideology of Agencies, Presidents, and Congress." *American Journal of Political Science* 56:341–354.

Clinton, Joshua D. and David E. Lewis. 2008. "Expert Opinion, Agency Characteristics, and Agency Preferences." *Political Analysis* 16:3–20.

Clinton, Joshua, Simon Jackman and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98:355–370.

Dempster, Arthur P., Nan M. Laird and Donald B. Rubin. 1977. "Maximum Likelihood from Incomplete Data Via the EM Algorithm (with Discussion)." *Journal of the Royal Statistical Society, Series B, Methodological* 39:1–37.

Gelman, Andrew. 2006. "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis* 1:515–533.

Gerber, Elisabeth R. and Jeffrey B. Lewis. 2004. "Beyond the median: Voter preferences, district heterogeneity, and political representation." *Journal of Political Economy* 112:1364–1383.

Gerrish, Sean M. and David M. Blei. 2012. "How They Vote: Issue-Adjusted Models of Legislative Behavior." *Neural Information Processing Systems.*

Grimmer, Justin. 2011. "An Introduction to Bayesian Inference via Variational Approximations." *Political Analysis* 19:32–47.

Hirano, Shigeo, Kosuke Imai, Yuki Shiraito and Masaaki Taniguchi. 2011. "Policy Positions in Mixed Member Electoral Systems:Evidence from Japan." Working Paper available at `http://imai.princeton.edu/research/japan.html`.

Ho, Daniel E. and Kevin M. Quinn. 2010. "Did a Switch in Time Save Nine?" *Journal of Legal Analysis* 2:1–45.

Imai, Kosuke, James Lo and Jonathan Olmsted. 2015. "emIRT: EM Algorithms for Estimating Item Response Theory Models." available at the Comprehensive R Archive Network (CRAN). `http://CRAN.R-project.org/package=emIRT`.

Jackman, Simon. 2001. "Multidimensional Analysis of Roll Call Data via Bayesian Simulation: Identification, Estimation, Inference, and Model Checking." *Political Analysis* 9:227–241.

Jackman, Simon. 2012. *pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University*. Department of Political Science, Stanford University, Stanford, California: Stanford University. R package version 1.04.4.

Lewis, Jeffrey B. and Keith T. Poole. 2004. "Measuring Bias and Uncertainty in Ideal Point Estimates via the Parametric Boostrap." *Political Analysis* 12(2):105–127.

Londregan, John B. 1999. "Estimating Legislators' Preferred Points." *Political Analysis* 8:35–56.

Londregan, John B. 2007. *Legislative Institutions and Ideology in Chile*. Cambridge: Cambridge University Press.

Martin, Andrew D. and Kevin M. Quinn. 2002. "Dynamic Ideal Point Estimation via Markov chain Monte Carlo for the U.S. Supreme Court, 1953–1999." *Political Analysis* 10:134–153.

Martin, Andrew D., Kevin M. Quinn and Jong Hee Park. 2013. *MCMCpack: Markov chain Monte Carlo MCMC Package*.
URL: *http://cran.r-project.org/web/packages/MCMCpack*

McCarty, Nolan, Keith T. Poole and Howard Rosenthal. 2006. *Polarized America: The Dance of Ideology and Unequal Riches*. Cambridge: MIT Press.

Morgenstern, Scott. 2004. *Patterns of Legislative Politics: Roll-Call Voting in Latin America and the United States*. Cambridge: Cambridge University Press.

Peress, Michael. 2013. "Estimating Proposal and Status Quo LocationsUsing Voting and Cosponsorship Data." *Journal of Politics* 75:613–631.

Poole, Keith, Jeffrey Lewis, James Lo and Royce Carroll. 2011. "Scaling Roll Call Votes with wnominate in R." *Journal of Statistical Software* 42:1–21.
URL: *http://www.jstatsoft.org/v42/i14/*

Poole, Keith, Jeffrey Lewis, James Lo and Royce Carroll. 2012. *oc: OC Roll Call Analysis Software.* R package version 0.93.
URL: *http://CRAN.R-project.org/package=oc*

Poole, Keith T. 2000. "Nonparametric Unfolding of Binary Choice Data." *Political Analysis* 8:211–237.

Poole, Keith T. and Howard Rosenthal. 1997. *Congress: A Political Economic History of Roll Call Voting.* Oxford University Press.

Poole, Keither T. and Howard Rosenthal. 1991. "Patterns of Congressional Voting." *American Journal of Political Science* 35:228–278.

Quinn, Kevin M. 2004. "Bayesian Factor Analysis for Mixed Ordinal and Continuous Responses." *Political Analysis* 12:338—353.

Rosas, Guillermo and Yael Shomer. 2008. "Models of Nonresponse in Legislative Politics." *Legislative Studies Quarterly* 33:573–601.

Shor, Boris, Christopher Berry and Nolan McCarty. 2011. "A Bridge to Somewhere: Mapping State and Congressional Ideology on a Cross-institutional Common Space." *Legislative Studies Quarterly* 35:417–448.

Shor, Boris and Nolan McCarty. 2011. "The Ideological Mapping of American Legislatures." *American Political Science Review* 105:530–551.

Spirling, Arthur and Iain McLean. 2007. "UK OC OK? Interpreting Optimal Classification Scores for the U.K. House of Commons." *Political Analysis* 15:85–96.

Tausanovitch, Chris and Christopher Warshaw. 2013. "Measuring Constituent Policy Preferences in Congress, State Legislatures, and Cities." *Journal of Politics* 75:330–342.

Voeten, Erik. 2000. "Clashes in the Assembly." *International Organization* 54:185–215.

Wainwright, Martin J. and Michael I. Jordan. 2008. "Graphical Models, Exponential Families, and Variational Inference." *Foundations and Trends in Machine Learning* 1:1–310.

Supplementary Appendix for "Fast Estimation of Ideal Points with Massive Data"

Kosuke Imai James Lo Jonathan Olmsted

In this appendix, we derive the variational Expectation-Maximization (EM) algorithm for the Bayesian ideal point models. For completeness, we begin by describing variational inference for the standard ideal point model (Appendix A) and the model with an ordinal outcome (Appendix B). In Appendix A, we also briefly explain variational inference in the context of the standard ideal point model for the readers who are not familiar with it. Finally, we derive the variational EM algorithms for the dynamic and hierarchical ideal point models (Appendices C and D), which represent the main contributions of the paper.

A Variational Inference for the Standard Ideal Point Model

We begin by deriving the variational EM algorithm for the standard ideal point model described in Section 2.1. The key idea of variational inference is to come up with the best approximation to the posterior distribution under a certain factorization assumption. Under the standard ideal point model, we consider the approximating distribution that satisfies the following independence relationship,

$$q(\mathbf{Y}^*, \{\mathbf{x}_i\}_{i=1}^N, \{\tilde{\boldsymbol{\beta}}_j\}_{j=1}^J) \;=\; q(\mathbf{Y}^*)\; q(\mathbf{x}_1, \ldots, \mathbf{x}_N)\; q(\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_J). \tag{59}$$

Under this assumption, we find the optimal variational distribution that best approximates the true posterior distribution. We do this by minimizing the following Kullback-Leibler divergence, which is a measure of similarity of two distributions,

$$KL(q||p) \;=\; \mathbb{E}_q \left\{ \log \frac{q(\mathbf{Y}^*, \{\mathbf{x}_i\}_{i=1}^N, \{\tilde{\boldsymbol{\beta}}_j\}_{j=1}^J)}{p(\mathbf{Y}^*, \{\mathbf{x}_i\}_{i=1}^N, \{\tilde{\boldsymbol{\beta}}_j\}_{j=1}^J \mid \mathbf{Y})} \right\} \tag{60}$$

where the expectation is taken with respect to the approximating distribution q.

It is well known that minimizing the Kullback-Leibler divergence given in equation (60) is equivalent to maximizing the lower bound of the marginal log-likelihood function, which is called the evidence lower bound in the literature. This can be shown by the following equality,

$$\log p(\mathbf{Y}) \;=\; \underbrace{\mathbb{E}_q \left[\log \left\{ \frac{p(\mathbf{Y}, \mathbf{Y}^*, \{\mathbf{x}_i\}_{i=1}^N, \{\tilde{\boldsymbol{\beta}}_j\}_{j=1}^J)}{q(\mathbf{Y}^*, \{\mathbf{x}_i\}_{i=1}^N, \{\tilde{\boldsymbol{\beta}}_j\}_{j=1}^J)} \right\} \right]}_{\text{evidence lower bound}} + KL(q||p) \tag{61}$$

Therefore, we can use the Expectation-Maximization (EM) algorithm of Dempster, Laird and Rubin (1977) to maximize the evidence lower bound. It is important to note that the proposed algorithm is derived without making an additional assumption other than the factorization assumption given in equation (59). For example, we do not assume that q belongs to a certain family of distributions.

A.1 Variational Distribution

We outline the proposed variational EM algorithm. As shown below, each step resembles the Gibbs sampler algorithm used to estimate the standard Bayesian ideal point model. There are a total of three steps in this EM algorithm (latent propensities, ideal points, and item parameters) and we repeat these steps until convergence.

Latent Propensities. We expand the joint density given in equation (4) and apply the logarithm to the product. Collecting the terms that involve $\mathbf{Y}^*$, we obtain,

$$\begin{aligned}
\log q(\mathbf{Y}^*) &= \mathbb{E}_{\tilde{\boldsymbol{\beta}},\mathbf{x}}\left[\log p(\mathbf{Y} \mid \mathbf{Y}^*) + \log p(\mathbf{Y}^* \mid \{\mathbf{x}_i\}_{i=1}^N, \{\tilde{\boldsymbol{\beta}}_j\}_{j=1}^J)\right] + \text{const.} \\
&= \sum_{i=1}^N \sum_{j=1}^J \left[\log\left(\mathbf{1}\{y_{ij}^* > 0\}\mathbf{1}\{y_{ij} = 1\} + \mathbf{1}\{y_{ij}^* \leq 0\}\mathbf{1}\{y_{ij} = 0\}\right) - \frac{1}{2}\left\{{y_{ij}^*}^2 - 2y_{ij}^* \mathbb{E}(\tilde{\mathbf{x}}_i)^\top \mathbb{E}(\tilde{\boldsymbol{\beta}}_j)\right\}\right] + \text{const.}
\end{aligned}$$

We recognize this as the product of $N \times J$ truncated Normal distributions, which is given as,

$$q(\mathbf{Y}^*) = \prod_{i=1}^N \prod_{j=1}^J q(y_{ij}^*) \quad \text{where} \quad q(y_{ij}^*) = \begin{cases} \mathcal{TN}(m_{ij}, 1, 0, \infty) & \text{if } y_{ij} = 1 \\ \mathcal{TN}(m_{ij}, 1, -\infty, 0) & \text{if } y_{ij} = 0 \end{cases}$$

where $m_{ij} = \mathbb{E}(\tilde{\mathbf{x}}_i)^\top \mathbb{E}(\tilde{\boldsymbol{\beta}}_j)$. Abstention is treated as missing at random and so the variational distribution in that case is $q(y_{ij}^*) = \mathcal{N}(m_{ij}, 1)$. Given this result, we update the mean of y_{ij}^* as,

$$\mathbb{E}(y_{ij}^*) = \begin{cases} m_{ij} + \frac{\phi(m_{ij})}{\Phi(m_{ij})} & \text{if } y_{ij} = 1 \\ m_{ij} - \frac{\phi(m_{ij})}{1-\Phi(m_{ij})} & \text{if } y_{ij} = 0 \end{cases}$$

Ideal Points. For the ideal points, we have,

$$\begin{aligned}
\log q(\mathbf{x}_1, \ldots, \mathbf{x}_N) &= \mathbb{E}_{\tilde{\boldsymbol{\beta}},\mathbf{y}^*}[\log p(\mathbf{Y}^* \mid \{\mathbf{x}_i\}_{i=1}^N, \{\tilde{\boldsymbol{\beta}}_j\}_{j=1}^J) + \log p(\mathbf{x}_1, \ldots, \mathbf{x}_N)] + \text{const.} \\
&= \sum_{i=1}^N \sum_{j=1}^J \mathbb{E}_{\tilde{\boldsymbol{\beta}}_j, y_{ij}^*}\left[\log \phi_1(y_{ij}^*; \tilde{\mathbf{x}}_i^\top \tilde{\boldsymbol{\beta}}_j, 1)\right] + \sum_{j=1}^J \log \phi_K(\mathbf{x}_i; \boldsymbol{\mu}_\mathbf{x}, \boldsymbol{\Sigma}_\mathbf{x}) + \text{const.}
\end{aligned}$$

We apply the standard result of the Bayesian linear regression to obtain,

$$q(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \prod_{i=1}^N q(\mathbf{x}_i) \quad \text{where} \quad q(\mathbf{x}_i) = \mathcal{N}(\mathbf{A}^{-1}\mathbf{a}_i,\ \mathbf{A}^{-1})$$

where $\mathbf{A} = \boldsymbol{\Sigma}_\mathbf{x}^{-1} + \sum_{j=1}^J \mathbb{E}(\boldsymbol{\beta}_j \boldsymbol{\beta}_j^\top)$ and $\mathbf{a}_i = \boldsymbol{\Sigma}_\mathbf{x}^{-1}\boldsymbol{\mu}_\mathbf{x} + \sum_{j=1}^J \mathbb{E}(\boldsymbol{\beta}_j)\mathbb{E}(y_{ij}^*) - \mathbb{E}(\boldsymbol{\beta}_j \alpha_j)$. Thus, we update the required moments as $\mathbb{E}(\mathbf{x}_i) = \mathbf{A}^{-1}\mathbf{a}_i$, and

$$\mathbb{E}(\tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top) = \begin{bmatrix} 1 & \mathbf{a}_i^\top \mathbf{A}^{-1} \\ \mathbf{A}^{-1}\mathbf{a}_i & \mathbf{A}^{-1}\mathbf{a}_i\mathbf{a}_i^\top\mathbf{A}^{-1} + \mathbf{A}^{-1} \end{bmatrix}$$

Item Parameters. The derivation for item parameters proceeds as follows,

$$\begin{aligned}\log q(\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_J) &= \mathbb{E}_{\mathbf{x},\mathbf{y}^*}[\log p(\mathbf{Y}^* \mid \{\mathbf{x}_i\}_{i=1}^N, \{\tilde{\boldsymbol{\beta}}_j\}_{j=1}^J) + \log p(\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_J)] + \text{const.} \\ &= \sum_{j=1}^{J}\sum_{i=1}^{N} \mathbb{E}_{\mathbf{x}_i, y_{ij}^*}\left[\log \phi_1(y_{ij}^* \mid \tilde{\mathbf{x}}_i^\top \tilde{\boldsymbol{\beta}}_j, 1)\right] + \sum_{j=1}^{J} \log p(\tilde{\boldsymbol{\beta}}_j) + \text{const.}\end{aligned}$$

Again, using the standard Bayesian linear regression result, we obtain,

$$q(\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_J) = \prod_{j=1}^{J} q(\tilde{\beta}_j) \quad \text{where} \quad q(\tilde{\boldsymbol{\beta}}_j) = \mathcal{N}(\mathbf{B}^{-1}\mathbf{b}_j,\ \mathbf{B}^{-1})$$

where $\mathbf{B} = \Sigma_{\tilde{\boldsymbol{\beta}}}^{-1} + \sum_{i=1}^N \mathbb{E}(\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top)$ and $\mathbf{b}_j = \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}^{-1}\boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}} + \sum_{i=1}^N \mathbb{E}(\tilde{\mathbf{x}}_i)\mathbb{E}(y_{ij}^*)$. This variational distribution implies the following required moments, i.e., $\mathbb{E}(\tilde{\boldsymbol{\beta}}_j) = \mathbf{B}^{-1}\mathbf{b}_j$, $\mathbb{E}(\tilde{\boldsymbol{\beta}}_j\tilde{\boldsymbol{\beta}}_j^\top) = \mathbf{B}^{-1}\mathbf{b}_j\mathbf{b}_j^\top\mathbf{B}^{-1} + \mathbf{B}^{-1}$ whose lower-right $K \times K$ block matrix and lower-left K dimensional column vector are equal to $\mathbb{E}(\boldsymbol{\beta}_j\boldsymbol{\beta}_j^\top)$ and $\mathbb{E}(\boldsymbol{\beta}_j\alpha_j)$, respectively.

A.2 Evidence Lower Bound

We can also derive the expression for the evidence lower bound, $\mathcal{L}(q)$ given in equation (61). This lower bound can be decomposed as

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}\left[\log p(\mathbf{Y}, \mathbf{Y}^*, \{\mathbf{x}_i\}_{i=1}^N, \{\tilde{\boldsymbol{\beta}}_j\}_{j=1}^J)\right] - \mathbb{E}\left[\log q(\mathbf{Y}^*, \{\mathbf{x}_i\}_{i=1}^N, \{\tilde{\boldsymbol{\beta}}_j\}_{j=1}^J)\right] \\ &= \sum_{i=1}^{N}\sum_{j=1}^{J} \mathbb{E}\left[\log p(y_{ij} \mid y_{ij}^*) + \log p(y_{ij}^* \mid \mathbf{x}_i, \tilde{\boldsymbol{\beta}}_j)\right] + \sum_{i=1}^{N} \mathbb{E}\left[\log p(\mathbf{x}_i)\right] + \sum_{j=1}^{J} \mathbb{E}\left[\log p(\tilde{\boldsymbol{\beta}}_j)\right] \\ &\quad - \sum_{i=1}^{N}\sum_{j=1}^{J} \mathbb{E}\left[\log q(y_{ij}^*)\right] - \sum_{i=1}^{N} \mathbb{E}\left[\log q(\mathbf{x}_i)\right] - \sum_{j=1}^{J} \mathbb{E}\left[\log q(\tilde{\boldsymbol{\beta}}_j)\right]\end{aligned}$$

First, we need to compute the entropy for each variational distribution, which is given below,

$$\begin{aligned}\mathbb{E}[\log q(y_{ij}^*)] &= \frac{1}{2}\log(2\pi e) + y_{ij}\log\Phi(m_{ij}) + (1-y_{ij})\log\Phi\left(-m_{ij}\right) + \left(\frac{-m_{ij}\phi(-m_{ij})}{2\Phi(m_{ij})}\right)^{y_{ij}}\left(\frac{m_{ij}\phi(-m_{ij})}{2\Phi(-m_{ij})}\right)^{1-y_{ij}} \\ \mathbb{E}\left[\log q(\mathbf{x}_i)\right] &= \frac{K}{2}\log(2\pi e) - \frac{1}{2}\log|\mathbf{A}| \\ \mathbb{E}[\log q(\tilde{\boldsymbol{\beta}}_j)] &= \frac{K+1}{2}\log(2\pi e) - \frac{1}{2}\log|\mathbf{B}|\end{aligned}$$

Second, we need to evaluate the expectation of log-likelihood and log prior density using the variational distribution. We begin by noting $\mathbb{E}[\log p(y_{ij} \mid y_{ij}^*)] = 0$. For the log-likelihood, therefore, we have,

$$\mathbb{E}\left[\log p(y_{ij}^* \mid \mathbf{x}_i, \tilde{\boldsymbol{\beta}}_j)\right] = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\left\{\mathbb{E}\left[(y_{ij}^*)^2\right] - 2\mathbb{E}(y_{ij}^*)\mathbb{E}(\tilde{\mathbf{x}}_i)^\top\mathbb{E}(\tilde{\boldsymbol{\beta}}_j) + \mathbb{E}\left[(\tilde{\mathbf{x}}_i^\top\tilde{\boldsymbol{\beta}}_j)^2\right]\right\}$$

where $\mathbb{E}[(\tilde{\mathbf{x}}_i^\top\tilde{\boldsymbol{\beta}}_j)^2] = \mathbb{E}[\text{tr}(\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top\tilde{\boldsymbol{\beta}}_j\tilde{\boldsymbol{\beta}}_j^\top)] = \text{tr}[\mathbb{E}(\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top)\mathbb{E}(\tilde{\boldsymbol{\beta}}_j\tilde{\boldsymbol{\beta}}_j^\top)]$. The expectation of the log prior density is given by,

$$\mathbb{E}[\log p(\mathbf{x}_i)] = -\frac{K}{2}\log(2\pi) - \frac{1}{2}|\boldsymbol{\Sigma}_{\mathbf{x}}| - \frac{1}{2}\left\{\mathbb{E}(\mathbf{x}_i^\top\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{x}_i) - 2\boldsymbol{\mu}_{\mathbf{x}}^\top\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbb{E}(\mathbf{x}_i) + \boldsymbol{\mu}_{\mathbf{x}}^\top\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\mu}_{\mathbf{x}}\right\}$$

where $\mathbb{E}\left[\mathbf{x}_i^\top \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{x}_i\right] = \text{tr}\{\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbb{E}(\mathbf{x}_i\mathbf{x}_i)^\top\}$. Moreover,

$$\mathbb{E}[\log(\tilde{\boldsymbol{\beta}}_j)] \;=\; -\frac{K+1}{2}\log(2\pi) - \frac{1}{2}\left\{\mathbb{E}\left(\tilde{\boldsymbol{\beta}}_j^\top \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}_j}^{-1}\tilde{\boldsymbol{\beta}}_j\right) - 2\boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}_j}^\top \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}_j}^{-1}\mathbb{E}(\tilde{\boldsymbol{\beta}}_j) + \boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}_j}^\top \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}_j}^{-1}\boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}_j}\right\}$$

where $\mathbb{E}(\tilde{\boldsymbol{\beta}}_j^\top \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}^{-1}\tilde{\boldsymbol{\beta}}_j) = \text{tr}\{\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}^{-1}\mathbb{E}(\tilde{\boldsymbol{\beta}}_j\tilde{\boldsymbol{\beta}}_j^\top)\}$. Finally, we note that $\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^\top) = \mathbf{A}^{-1}\mathbf{a}_i\mathbf{a}_i^\top\mathbf{A}^{-1} + \mathbf{A}^{-1}$ and $\mathbb{E}(\tilde{\boldsymbol{\beta}}_j\tilde{\boldsymbol{\beta}}_j^\top) = \mathbf{B}^{-1}\mathbf{b}_j\mathbf{b}_j^\top\mathbf{B}^{-1} - \mathbf{B}^{-1}$.

B Variational Inference for the Model with an Ordinal Outcome

We derive the variational distribution for the ideal point with a three-category ordinal outcome. We use the reparameterized model given in Section 3.2. The joint posterior of the reparameterized model is given by,

$$\begin{aligned}
&p(\mathbf{Z}^*, \{\mathbf{x}_i\}_{i=1}^N, \{\tau_j^2, \tilde{\boldsymbol{\beta}}_j\}_{j=1}^J \mid \mathbf{Y}) \\
&\propto \prod_{i=1}^N\prod_{j=1}^J \left[\mathbf{1}\{z_{ij}^* < 0\}\mathbf{1}\{y_{ij}=0\} + \mathbf{1}\{0 \le z_{ij}^* < 1\}\mathbf{1}\{y_{ij}=1\} + \mathbf{1}\{z_{ij}^* \ge 1\}\mathbf{1}\{y_{ij}=2\}\right]\phi_1(z_{ij}^*; \tilde{\mathbf{x}}_i^\top\tilde{\boldsymbol{\beta}}_j, \tau_j^{-2}) \\
&\quad\times \prod_{j=1}^J \phi_{K+1}(\tilde{\boldsymbol{\beta}}_j; \boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}})\; \mathcal{G}\left(\tau_j^2; \frac{\nu_\tau}{2}, \frac{s_\tau^2}{2}\right)\prod_{i=1}^N \phi_K(\mathbf{x}_i; \mu_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}})
\end{aligned}$$

We derive the variational EM algorithm in a manner similar to the one used for the standard item response model with a binary outcome variable. The key difference is that the current model has an additional parameter τ_j.

B.1 Variational Distribution

Latent Propensities. We begin by deriving the variational distribution for the latent propensities as follows,

$$\begin{aligned}
\log q(z_{ij}^*) \;&=\; \mathbb{E}_{\tilde{\boldsymbol{\beta}}_j, \mathbf{x}_i, \tau_j}[\log p(y_{ij} \mid z_{ij}^*) + \log p(z_{ij}^* \mid \mathbf{x}_i, \tilde{\boldsymbol{\beta}}, \tau_j^2)] + \text{const.} \\
&=\; \log\left[\mathbf{1}\{z_{ij}^* < 0\}\mathbf{1}\{y_{ij}=1\} + \mathbf{1}\{0 \le z_{ij}^* < 1\}\mathbf{1}\{y_{ij}=2\} + \mathbf{1}\{z_{ij}^* \ge 1\}\mathbf{1}\{y_{ij}=3\}\right] \\
&\quad -\; \frac{\mathbb{E}(\tau_j^2)}{2}\left\{(z_{ij}^*)^2 - 2z_{ij}^*\mathbb{E}(\tilde{\mathbf{x}}_i)^\top\mathbb{E}(\tilde{\boldsymbol{\beta}}_j)\right\} + \text{const.}
\end{aligned}$$

We recognize this as the following truncated normal distribution,

$$q(\mathbf{Z}^*) \;=\; \prod_{i=1}^N\prod_{j=1}^J q(z_{ij}^*) \quad \text{where} \quad q(z_{ij}^*) \;=\; \begin{cases} \mathcal{TN}(m_{ij},\; w_j^{-2},\; -\infty,\; 0) & \text{if } y_{ij}=0 \\ \mathcal{TN}(m_{ij},\; w_j^{-2},\; 0,\; 1) & \text{if } y_{ij}=1 \\ \mathcal{TN}(m_{ij},\; w_j^{-2},\; 1,\; \infty) & \text{if } y_{ij}=2 \end{cases}$$

where $m_{ij} = \mathbb{E}(\tilde{\mathbf{x}}_i)^\top\mathbb{E}(\tilde{\boldsymbol{\beta}}_j)$ and $w_j^2 = \mathbb{E}(\tau_j^2)$. We then update the mean of z_{ij}^* as,

$$\mathbb{E}(z_{ij}^*) \;=\; \begin{cases} m_{ij} - \frac{\phi(m_{ij}w_j)}{1-\Phi(m_{ij}w_j)}w_j^{-1} & \text{if } y_{ij}=0 \\ m_{ij} + \frac{\phi(m_{ij}w_j)-\phi((1-m_{ij})w_j)}{\Phi((1-m_{ij})w_j)+\Phi(m_{ij}w_j)-1}w_j^{-1} & \text{if } y_{ij}=1 \\ m_{ij} + \frac{\phi((1-m_{ij})w_j)}{1-\Phi((1-m_{ij})w_j)}w_j^{-1} & \text{if } y_{ij}=2 \end{cases}$$

For abstention, we have $q(z_{ij}^*) = \mathcal{N}(m_{ij},\; w_j^{-2})$ and hence $\mathbb{E}(z_{ij}^*) = m_{ij}$.

Ideal Points. We expand the joint density p and apply the logarithm to the product.

$$\begin{aligned}\log q(\mathbf{x}_i) &= \mathbb{E}_{z^*,\tilde{\boldsymbol{\beta}}^*,\tau}\left[\sum_{j=1}^{J}\log p(z_{ij}^* \mid \mathbf{x}_i, \tilde{\boldsymbol{\beta}}_j^*, \tau_j^2)\right] + \log p(\mathbf{x}_i) + \text{const.} \\ &= -\sum_{j=1}^{J}\frac{\mathbb{E}(\tau_j^2)}{2}\left\{\mathbf{x}_i^\top \mathbb{E}\left(\boldsymbol{\beta}_j^*\boldsymbol{\beta}_j^{*\top}\right)\mathbf{x}_i - 2\boldsymbol{\beta}_j^{*\top}(z_{ij}^* - \alpha_j^*)\mathbf{x}_i\right\} - \frac{1}{2}\left(\mathbf{x}_i^\top \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{x}_i - 2\boldsymbol{\mu}_{\mathbf{x}}^\top\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{x}_i\right) + \text{const.}\end{aligned}$$

Since this is a normal distribution density, we have,

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \prod_{i=1}^{N} q(\mathbf{x}_i) \quad \text{where} \quad q(\mathbf{x}_i) = \mathcal{N}(\mathbf{A}^{-1}\mathbf{a}_i,\ \mathbf{A}^{-1})$$

with $\mathbf{A} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} + \sum_{j=1}^{J}\mathbb{E}(\tau_j^2)\mathbb{E}\left(\boldsymbol{\beta}_j^*\boldsymbol{\beta}_j^{*\top}\right)$ and $\mathbf{a}_i = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\mu}_{\mathbf{x}} + \sum_{j=1}^{J}\mathbb{E}(\tau_j^2)\left\{\mathbb{E}(\boldsymbol{\beta}_j^*)^\top\mathbb{E}(z_{ij}^*) - \mathbb{E}\left(\boldsymbol{\beta}_j^{*\top}\alpha_j^*\right)\right\}$. Given this result, we update the required moments, i.e., $\mathbb{E}(\mathbf{x}_i) = \mathbf{A}^{-1}\mathbf{a}_i$ and $\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^\top) = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{a}_i\mathbf{a}_i^\top\mathbf{A}^{-1}$.

Item Parameters. For the item parameters, the variational distribution can be derived as follows,

$$\begin{aligned}\log q(\tilde{\boldsymbol{\beta}}_j) &= \mathbb{E}_{\mathbf{z}_i^*,\mathbf{x},\tau_j}\left[\sum_{i=1}^{N}\log p(z^* \mid \mathbf{x}_i, \tilde{\boldsymbol{\beta}}_j, \tau_j)\right] + \log p(\tilde{\boldsymbol{\beta}}_j))] + \text{const.} \\ &= -\frac{\mathbb{E}(\tau_j^2)}{2}\sum_{i=1}^{N}\left\{\tilde{\boldsymbol{\beta}}_j^\top\mathbb{E}(\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top)\tilde{\boldsymbol{\beta}}_j - 2z_{ij}^*\tilde{\mathbf{x}}_i^\top\tilde{\boldsymbol{\beta}}_j\right\} - \frac{1}{2}\left(\tilde{\boldsymbol{\beta}}_j^\top\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}^{-1}\tilde{\boldsymbol{\beta}}_j - \boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}}^\top\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}^{-1}\tilde{\boldsymbol{\beta}}_j\right) + \text{const.}\end{aligned}$$

This is another Normal distribution, which is given by,

$$q(\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_J) = \prod_{j=1}^{J} q(\tilde{\boldsymbol{\beta}}_j) \quad \text{where} \quad q(\tilde{\boldsymbol{\beta}}_j) = \mathcal{N}(\mathbf{B}_j^{-1}\mathbf{b}_j,\ \mathbf{B}_j^{-1})$$

where $\mathbf{B}_j = \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}^{-1} + \mathbb{E}(\tau_j^2)\sum_{i=1}^{N}\mathbb{E}(\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top)$ and $\mathbf{b}_j = \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}^{-1}\boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}} + \mathbb{E}(\tau_j^2)\sum_{i=1}^{N}\mathbb{E}(\tilde{\mathbf{x}}_i)^\top\mathbb{E}(z_{ij}^*)$. We update the required moment as $\mathbb{E}(\tilde{\boldsymbol{\beta}}_j\tilde{\boldsymbol{\beta}}_j^\top) = \mathbf{B}_j^{-1} + \mathbf{B}_j^{-1}\mathbf{b}_j\mathbf{b}_j^\top\mathbf{B}_j^{-1}$.

Variance Parameters. Finally, we derive the variational distribution for the variance parameters,

$$\begin{aligned}\log q(\tau_j^2) &= \mathbb{E}_{\mathbf{z}_j^*,\mathbf{x},\tilde{\boldsymbol{\beta}}_j^*}\left[\sum_{i=1}^{N}\log p(z_{ij}^* \mid \mathbf{x}_i, \tilde{\boldsymbol{\beta}}_j, \tau_j^2)\right] + \log p(\tau_j^2) + \text{const.} \\ &= \frac{N}{2}\log\tau_j^2 - \frac{\tau_j^2}{2}\sum_{i=1}^{N}\left[\mathbb{E}({z_{ij}^*}^2) - 2\mathbb{E}(z_{ij}^*)\mathbb{E}(\tilde{\mathbf{x}}_i)^\top\mathbb{E}(\tilde{\boldsymbol{\beta}}_j) + \text{tr}\{\mathbb{E}(\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top)\mathbb{E}(\tilde{\boldsymbol{\beta}}_j\tilde{\boldsymbol{\beta}}_j^\top)\}\right] \\ &\quad + \left(\frac{\nu_\tau}{2} - 1\right)\log\tau_j^2 - \frac{s_\tau\tau_j^2}{2} + \text{const.}\end{aligned}$$

We recognize this as a Gamma distribution and thus the variational distribution is given by,

$$q(\tau_1^2, \ldots, \tau_J^2) = \prod_{j=1}^{J} q(\tau_j) \quad \text{where} \quad q(\tau_j^2) = \mathcal{G}\left(\frac{c_j}{2},\ \frac{d_j}{2}\right)$$

with $c_j = N + \nu_\tau$ and $d_j = s_\tau + \sum_{i=1}^{N}\mathbb{E}({z_{ij}^*}^2) - 2\sum_{i=1}^{N}\mathbb{E}(z_{ij}^*)\mathbb{E}(\tilde{\mathbf{x}}_i)^\top\mathbb{E}(\tilde{\boldsymbol{\beta}}_j) + \sum_{i=1}^{N}\text{tr}\{\mathbb{E}(\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top)\mathbb{E}(\tilde{\boldsymbol{\beta}}_j\tilde{\boldsymbol{\beta}}_j^\top)\}$. The required moment is given by $\mathbb{E}(\tau_j^2) = c_j/d_j$.

B.2 Evidence Lower Bound

The lower bound is given by,

$$\begin{aligned}
\mathcal{L}(q) &= \mathbb{E}\left[\log p(\mathbf{Y}, \mathbf{Z}^*, \{\mathbf{x}_i\}_{i=1}^N, \{\tau_j^2, \tilde{\boldsymbol{\beta}}_j\}_{j=1}^J)\right] - \mathbb{E}\left[\log q(\mathbf{Z}^*, \{\mathbf{x}_i\}_{i=1}^N, \{\tau_j^2, \tilde{\boldsymbol{\beta}}_j\}_{j=1}^J)\right] \\
&= \sum_{i=1}^N \sum_{j=1}^J \mathbb{E}\left[\log p(y_{ij} \mid z_{ij}^*) + \log p(z_{ij}^* \mid \mathbf{x}_i, \tilde{\boldsymbol{\beta}}_j, \tau_j^2)\right] + \sum_{i=1}^N \mathbb{E}[\log p(\mathbf{x}_i)] + \sum_{j=1}^J \mathbb{E}[\log p(\tilde{\boldsymbol{\beta}}_j)] + \sum_{j=1}^J \mathbb{E}[\log p(\tau_j^2)] \\
&\quad - \sum_{i=1}^N \sum_{j=1}^J \mathbb{E}\left[\log q(z_{ij}^*)\right] - \sum_{i=1}^N \mathbb{E}\left[\log q(\mathbf{x}_i)\right] - \sum_{j=1}^J \mathbb{E}\left[\log q(\tilde{\boldsymbol{\beta}}_j)\right] - \sum_{j=1}^J \mathbb{E}\left[\log q(\tau_j^2)\right]
\end{aligned}$$

We begin by noting $\mathbb{E}[\log p(y_{ij} \mid z_{ij}^*)] = 0$. Next, we compute the entropy of variational distribution for the latent propensity as follows,

$$\begin{aligned}
&\mathbb{E}[\log q(z_{ij}^*)] \\
&= \frac{1}{2}\{\log(2\pi e) - \log \mathbb{E}(\tau_j^2)\} + \mathbf{1}\{y_{ij} = 0\} \log \Phi\left(m_{ij}^*\right) + \mathbf{1}\{y_{ij} = 1\} \log\{\Phi(\tilde{m}_{ij}) - \Phi(m_{ij}^*)\} + \mathbf{1}\{y_{ij} = 2\} \log \Phi(-\tilde{m}_{ij}) \\
&\quad + \left(-\frac{m_{ij}^* \phi(m_{ij}^*)}{2\Phi(m_{ij}^*)}\right)^{\mathbf{1}\{y_{ij}=0\}} \left(\frac{m_{ij}^* \phi(m_{ij}^*) - \tilde{m}_{ij}\phi(\tilde{m}_{ij})}{2\{\Phi(\tilde{m}_{ij}) - \Phi(m_{ij}^*)\}}\right)^{\mathbf{1}\{y_{ij}=1\}} \left(\frac{\tilde{m}_{ij}\phi(\tilde{m}_{ij})}{2\Phi(-\tilde{m}_{ij})}\right)^{\mathbf{1}\{y_{ij}=2\}}
\end{aligned}$$

where $m_{ij}^* = -\mathbb{E}(\tilde{\mathbf{x}}_i)^\top \mathbb{E}(\tilde{\boldsymbol{\beta}}_j)/\sqrt{\mathbb{E}(\tau_j^2)}$ and $\tilde{m}_{ij} = \{1 - \mathbb{E}(\tilde{\mathbf{x}}_i)^\top \mathbb{E}(\tilde{\boldsymbol{\beta}}_j)\}/\sqrt{\mathbb{E}(\tau_j^2)}$. In addition, we have the following entropies,

$$\begin{aligned}
\mathbb{E}[\log q(\mathbf{x}_i)] &= \frac{K}{2}\log(2\pi e) - \frac{1}{2}\log|\mathbf{A}| \\
\mathbb{E}[\log q(\tilde{\boldsymbol{\beta}}_j)] &= \frac{K+1}{2}\log(2\pi e) - \frac{1}{2}\log|\mathbf{B}_j| \\
\mathbb{E}[\log q(\tau_j^2)] &= \frac{c_j}{2} - \log\frac{d_j}{2} + \log\left[\Gamma\left(\frac{c_j}{2}\right)\right] + \left(1 - \frac{c_j}{2}\right)\psi\left(\frac{c_j}{2}\right)
\end{aligned}$$

where $\Gamma(\cdot)$ and $\psi(\cdot)$ are the gamma and digamma functions, respectively.

Next, the expected log-likelihood can be calculated as,

$$\mathbb{E}[\log p(z_{ij}^* \mid \mathbf{x}_i, \tilde{\boldsymbol{\beta}}_j, \tau_j^2)] = -\frac{1}{2}\left\{\log(2\pi) - \mathbb{E}(\log \tau_j^2)\right\} - \frac{\mathbb{E}(\tau_j^2)}{2}\left\{\mathbb{E}[(z_{ij}^*)^2] - 2\mathbb{E}(z_{ij}^*)\mathbb{E}(\tilde{\mathbf{x}}_i)^\top \mathbb{E}(\tilde{\boldsymbol{\beta}}_j) + \mathbb{E}(\tilde{\boldsymbol{\beta}}_j^\top \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top \tilde{\boldsymbol{\beta}}_j)\right\}$$

where $\mathbb{E}(\log \tau_j^2) = \psi(c_j/2) - \log(d_j/2)$ and $\mathbb{E}(\tilde{\boldsymbol{\beta}}_j^\top \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top \tilde{\boldsymbol{\beta}}_j) = \mathrm{tr}\{\mathbb{E}(\tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top)\mathbb{E}(\tilde{\boldsymbol{\beta}}_j \tilde{\boldsymbol{\beta}}_j^\top)\}$.

Finally, we compute the expected log-prior as follows,

$$\begin{aligned}
\mathbb{E}[\log p(\mathbf{x}_i)] &= -\frac{K}{2}\{\log(2\pi) + \log|\boldsymbol{\Sigma}_{\mathbf{x}}|\} - \frac{1}{2}\{\mathbb{E}(\mathbf{x}_i^\top \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{x}_i) - 2\mathbb{E}(\mathbf{x}_i)^\top \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\mu}_{\mathbf{x}} + \boldsymbol{\mu}_{\mathbf{x}}^\top \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\mu}_{\mathbf{x}}\} \\
\mathbb{E}[\log p(\tilde{\boldsymbol{\beta}}_i)] &= -\frac{K+1}{2}\{\log(2\pi) + \log|\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}|\} - \frac{1}{2}\{\mathbb{E}(\tilde{\boldsymbol{\beta}}_j^\top \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}^{-1}\tilde{\boldsymbol{\beta}}_j) - 2\mathbb{E}(\tilde{\boldsymbol{\beta}}_j)^\top \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}^{-1}\boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}} + \boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}}^\top \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}^{-1}\boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}}\} \\
\mathbb{E}[\log p(\tau_j^2)] &= \frac{c_j}{2}\log\frac{d_j}{2} - \log\Gamma\left(\frac{c_j}{2}\right) + \left(\frac{c_j}{2} - 1\right)\mathbb{E}(\log \tau_j^2) - \frac{d_j}{2}\mathbb{E}(\tau_j^2)
\end{aligned}$$

where $\mathbb{E}(\mathbf{x}_i^\top \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{x}_i) = \mathrm{tr}\{\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^\top)\}$ and $\mathbb{E}(\tilde{\boldsymbol{\beta}}_j^\top \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}^{-1}\tilde{\boldsymbol{\beta}}_j) = \mathrm{tr}\{\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}^{-1}\mathbb{E}(\tilde{\boldsymbol{\beta}}_j\tilde{\boldsymbol{\beta}}_j^\top)\}$.

C Variational Inference for the Dynamic Ideal Point Model

For the dynamic model, the joint posterior distribution is given by equation (37). We derive the variational EM algorithm under the factorization assumption given in equation (38).

Latent Propensities. For the latent propensities, we derive the variational distribution as follows,

$$\begin{aligned}\log q(y_{ijt}) &= \mathbb{E}_{x_i,\tilde{\boldsymbol{\beta}}_j}[\log p(y_{ijt} \mid y^*_{ijt}) + \log p(y_{ijt} \mid x_{it}, \tilde{\boldsymbol{\beta}}_{jt})] + \text{const.} \\ &= \log[\mathbf{1}\{y_{ijt}=1\}\mathbf{1}\{y^*_{ijt}>0\} + \mathbf{1}\{y_{ijt}=0\}\mathbf{1}\{y_{ijt}\leq 0\}] - \frac{1}{2}\left({y^*_{ijt}}^2 - 2y^*_{ijt}\mathbb{E}(\tilde{\mathbf{x}}_{it})^\top \mathbb{E}(\tilde{\boldsymbol{\beta}}_{jt})\right) + \text{const.}\end{aligned}$$

This is a truncated normal distribution, and hence the approximating distribution is given by equation (39).

Item Parameters. The variational distribution for item parameters is given by,

$$\begin{aligned}\log(\tilde{\boldsymbol{\beta}}_{jt}) &= \sum_{i\in\mathcal{I}_t} \mathbb{E}_{y^*_{ijt},\mathbf{x}}[\log p(y^*_{ijt} \mid x_{it}, \tilde{\boldsymbol{\beta}}_{jt})] + \log p(\tilde{\boldsymbol{\beta}}) + \text{const.} \\ &= -\frac{1}{2}\sum_{i\in\mathcal{I}_t}\left\{\tilde{\boldsymbol{\beta}}_{jt}^\top \mathbb{E}(\tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}_{it}^\top)\tilde{\boldsymbol{\beta}}_{jt} - 2\mathbb{E}(y^*_{ijt})\mathbb{E}(\tilde{\mathbf{x}}_{it})^\top\tilde{\boldsymbol{\beta}}_{jt}\right\} - \frac{1}{2}(\tilde{\boldsymbol{\beta}}_{jt}^\top \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}^{-1}\tilde{\boldsymbol{\beta}}_{jt} - 2\boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}}^\top\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}^{-1}\tilde{\boldsymbol{\beta}}_{jt}) + \text{const.}\end{aligned}$$

where $\mathcal{I}_t = \{i : \underline{T}_i \leq t \leq \overline{T}_i\}$. Recognizing this as a Normal distribution, we have the resulting variational distribution given in equation (41).

Ideal Points. For notational simplicity and without loss of generality, let $\underline{T}_i = 1$ and $\overline{T}_i = T$. Then, we have,

$$\begin{aligned}&\log q(x_{i1},\ldots,x_{iT}) \\ &= \sum_{t=1}^{T}\sum_{j=1}^{J_t}\mathbb{E}_{y^*_{ijt},\tilde{\boldsymbol{\beta}}}[\log p(y^*_{ijt} \mid x_{it}, \tilde{\boldsymbol{\beta}}_{jt})] + \sum_{t=2}^{T}\log p(x_{it} \mid x_{i,t-1}) + \log p(x_{it}) + \text{const.} \\ &= -\frac{1}{2}\sum_{t=1}^{T}\sum_{j=1}^{J_t}\left[\mathbb{E}(\beta_{jt}^2)x_{it}^2 - 2\{\mathbb{E}(y^*_{ijt})\mathbb{E}(\beta_{jt}) - \mathbb{E}(\beta_{jt}\alpha_{jt})\}x_{it}\right] - \frac{1}{2\omega_x^2}\sum_{t=2}^{T}(x_{it}-x_{i,t-1})^2 - \frac{1}{2\Sigma_x}(x_i-\mu_x)^2 + \text{const.}\end{aligned}$$

We recognize that this expression resembles the posterior distribution of the following dynamic linear model, i.e., $p(x_{i1},\ldots,x_{iT} \mid \ddot{y}_{i1},\ldots,\ddot{y}_{iT},\ddot{\beta}_1,\ldots,\ddot{\beta}_T)$,

$$\begin{aligned}\ddot{y}_{it} &= \ddot{\beta}_t x_{it} + \ddot{\epsilon}_{it} \\ x_{it} &= x_{i,t-1} + \eta_{it}\end{aligned}$$

where $\ddot{\beta}_t = \sqrt{\sum_{j=1}^{J_t}\mathbb{E}(\beta_{jt}^2)}$, $\ddot{y}_{it} = \{\sum_{j=1}^{J_t}\mathbb{E}(y^*_{ijt})\mathbb{E}(\beta_{jt}) - \mathbb{E}(\beta_{jt}\alpha_{jt})\}/\ddot{\beta}_t$, $\ddot{\epsilon}_{it} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $\eta_{it} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\omega_x^2)$, and $x_{i1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_x,\Sigma_x)$.

Thus, we can apply the standard Kalman filtering results to obtain the necessary moments. We begin by applying the forward recursion relationship,

$$p(x_{it} \mid \ddot{y}_{i1},\ldots,\ddot{y}_{it}) = \phi(x_{it}; c_{it}, C_{it})$$

where $c_{it} = c_{i,t-1} + K_t(\ddot{y}_{it} - \ddot{\beta}_t c_{i,t-1})$ and $C_{it} = (1 - K_t\ddot{\beta}_t)\Omega_t$ with $\Omega_{it} = \omega_{x,i}^2 + C_{i,t-1}$, $K_t = \ddot{\beta}_t\Omega_{it}/S_{it}$ and $S_t = \ddot{\beta}_t^2\Omega_{it} + 1$. We recursively compute these quantities by setting $c_{i0} = \mu_{x,i}$ and $C_{i0} = \Sigma_{x,i}$. We then use the backward recursion to obtain,

$$p(x_{it} \mid \ddot{y}_{i1}, \ldots, \ddot{y}_{iT}) \quad = \quad \phi_1(x_{it}; d_{it}, D_{it})$$

where $d_{it} = c_{it} + J_{it}(d_{i,t+1} - c_{i,t+1})$ and $D_{it} = C_{it} + J_{it}^2(D_{i,t+1} - \Omega_{i,t+1})$ with $J_{it} = C_{it}/\Omega_{i,t+1}$. Again, the recursive computation is done by setting $d_{iT} = c_{iT}$ and $D_{iT} = C_{iT}$.

D Variational Inference for the Hierarchical Ideal Point Model

For the hierarchical model, the joint posterior distribution is given by equation (50). We derive the variational EM algorithm under the factorization assumption given in equation (51).

Latent Propensities. The latent propensities can be derived as,

$$\begin{aligned}
\log q(y_\ell^*) &= \mathbb{E}_{\tilde{\boldsymbol{\beta}},\boldsymbol{\gamma},\eta}[\log p(y_\ell \mid y_\ell^*) + \log p(y_\ell^* \mid \tilde{\boldsymbol{\beta}}_{j[\ell]}, \boldsymbol{\gamma}_{g[i[\ell]]}, \eta_{i[\ell]}, \mathbf{z}_{i[\ell]})] + \text{const.} \\
&= \log[\mathbf{1}\{y_\ell^* > 0, y_\ell = 1\} + \mathbf{1}\{y_\ell^* \leq 0, y_\ell = 0\}] \\
&\quad - \frac{1}{2}\mathbb{E}_{\tilde{\boldsymbol{\beta}},\boldsymbol{\gamma},\eta}[(y_\ell^* - \alpha_{j[\ell]} - \beta_{j[\ell]}\boldsymbol{\gamma}_{g[i[\ell]]}^\top \mathbf{z}_{i[\ell]} - \beta_{j[\ell]}\eta_{i[\ell]})^2] + \text{const.} \\
&= \log[\mathbf{1}\{y_\ell^* > 0, y_\ell = 1\} + \mathbf{1}\{y_\ell^* \leq 0, y_\ell = 0\}] \\
&\quad - \frac{1}{2}\left[y_\ell^{*2} - 2y_\ell^*\left\{\mathbb{E}(\alpha_{j[\ell]}) + \mathbb{E}(\beta_{j[l]})\mathbb{E}(\boldsymbol{\gamma}_{g[i[\ell]]})^\top \mathbf{z}_{i[\ell]} + \mathbb{E}(\eta_{i[\ell]})\mathbb{E}(\beta_{j[\ell]})\right\}\right] + \text{const.}
\end{aligned}$$

We recognize that this is a truncated normal distribution, and therefore, we obtain the variational distribution given in equation (52).

Ideal Point Error Terms. We derive the variational distribution for the ideal point error terms as follows,

$$\begin{aligned}
&\log q(\eta_n) \\
&= \mathbb{E}_{y^*,\tilde{\boldsymbol{\beta}},\boldsymbol{\gamma},\sigma^2}\left[\log p(\eta_n \mid \sigma_{g[n]}^2) + \sum_{\ell=1}^{L} \mathbf{1}\{i[\ell] = n\} \log p(y_\ell^* \mid \tilde{\boldsymbol{\beta}}_{j[\ell]}, \boldsymbol{\gamma}_{g[n]}, \mathbf{z}_n, \eta_n)\right] + \text{const.} \\
&= -\frac{1}{2}\mathbb{E}_{y^*,\tilde{\boldsymbol{\beta}},\boldsymbol{\gamma},\sigma^2}\left[\frac{\eta_n^2}{\sigma_{g[n]}^2} + \sum_{\ell=1}^{L} \mathbf{1}\{i[\ell] = n\}(y_\ell^* - \alpha_{j[\ell]} - \beta_{j[\ell]}\boldsymbol{\gamma}_{g[n]}^\top \mathbf{z}_n - \beta_{j[\ell]}\eta_n)^2\right] + \text{const.} \\
&= -\frac{1}{2}\mathbb{E}_{y^*,\tilde{\boldsymbol{\beta}},\boldsymbol{\gamma},\sigma^2}\left[\left(\sigma_{g[n]}^{-2} + \sum_{\ell=1}^{L} \mathbf{1}\{i[\ell] = n\}\beta_{j[\ell]}^2\right)\eta_n^2 - 2\sum_{\ell=1}^{L} \mathbf{1}\{i[\ell] = n\}(y_\ell^*\beta_{j[\ell]} - \alpha_{j[\ell]}\beta_{j[\ell]} - \beta_{j[\ell]}^2\boldsymbol{\gamma}_{g[n]}^\top \mathbf{z}_n)\eta_n\right] \\
&\quad + \text{const.}
\end{aligned}$$

Thus, the approximating distribution is again a normal distribution and is given by equation (54).

Item Parameters. The variational distribution for item parameters can be derived as follows,

$$\log q(\tilde{\boldsymbol{\beta}}_k)$$

$$
\begin{aligned}
&= \mathbb{E}_{y^*,\eta,\boldsymbol{\gamma}}\left[\log p(\tilde{\boldsymbol{\beta}}_k) + \sum_{\ell=1}^{L} \mathbf{1}\{j[\ell]=k\}\log p(y_\ell^* \mid \tilde{\boldsymbol{\beta}}_k, \boldsymbol{\gamma}_{g[i[\ell]]}, \eta_{i[\ell]})\right] + \text{const.} \\
&= -\frac{1}{2}\mathbb{E}_{y^*,\eta,\boldsymbol{\gamma}}\left[(\tilde{\boldsymbol{\beta}}_k - \boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}})^\top \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}^{-1}(\tilde{\boldsymbol{\beta}}_k - \boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}}) + \sum_{\ell=1}^{L}\mathbf{1}\{j[\ell]=k\}\{y_\ell^* - \alpha_k - \beta_k(\boldsymbol{\gamma}_{g[i[\ell]]}^\top \mathbf{z}_{i[\ell]} + \eta_{i[\ell]})\}^2\right] + \text{const.} \\
&= -\frac{1}{2}\mathbb{E}_{y^*,\eta,\boldsymbol{\gamma}}\left[(\tilde{\boldsymbol{\beta}}_k - \boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}})^\top \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}^{-1}(\tilde{\boldsymbol{\beta}}_k - \boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}}) + \sum_{\ell=1}^{L}\mathbf{1}\{j[\ell]=k\}(y_\ell^* - \tilde{\boldsymbol{\beta}}_k^\top \tilde{\mathbf{x}}_{i[\ell]})^2\right] + \text{const.} \\
&= -\frac{1}{2}\mathbb{E}_{y^*,\eta,\boldsymbol{\gamma}}\left[\tilde{\boldsymbol{\beta}}_k^\top\left(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}^{-1} + \sum_{\ell=1}^{L}\mathbf{1}\{j[\ell]=k\}\tilde{\mathbf{x}}_{i[\ell]}\tilde{\mathbf{x}}_{i[\ell]}^\top\right)\tilde{\boldsymbol{\beta}}_k - 2\tilde{\boldsymbol{\beta}}_k^\top\left(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}^{-1}\boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}} + \sum_{\ell=1}^{L}\mathbf{1}\{j[\ell]=k\}y_\ell^*\tilde{\mathbf{x}}_{i[\ell]}\right)\right] + \text{const.}
\end{aligned}
$$

where $\tilde{\mathbf{x}}_{i[\ell]} = (1, \boldsymbol{\gamma}_{g[i[\ell]]}^\top \mathbf{z}_{i[\ell]} + \eta_{i[\ell]})^\top$ is a 2 dimensional vector. Thus, the variational distribution is given by equation (55).

Group-level Coefficients. The derivation of the variational distribution for group-level coefficients is given as follows,

$$
\begin{aligned}
&\log q(\boldsymbol{\gamma}_m) \\
&= \mathbb{E}_{y^*,\eta,\tilde{\boldsymbol{\beta}}}\left[\log p(\boldsymbol{\gamma}_m) + \sum_{\ell=1}^{L}\mathbf{1}\{g[i[\ell]]=m\}\log p(y_\ell^* \mid \tilde{\boldsymbol{\beta}}_{j[\ell]}, \boldsymbol{\gamma}_m, \eta_{i[\ell]})\right] + \text{const.} \\
&= -\frac{1}{2}\mathbb{E}_{y^*,\eta,\tilde{\boldsymbol{\beta}}}\left[(\boldsymbol{\gamma}_m - \boldsymbol{\mu}_{\boldsymbol{\gamma}})^\top \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}(\boldsymbol{\gamma}_m - \boldsymbol{\mu}_{\boldsymbol{\gamma}}) + \sum_{\ell=1}^{L} 1\{g[i[\ell]]=m\}\left\{y_\ell^* - \alpha_{j[\ell]} - \beta_{j[\ell]}(\boldsymbol{\gamma}_m^\top \mathbf{z}_{i[\ell]} + \eta_{i[\ell]})\right\}^2\right] + \text{const.} \\
&= -\frac{1}{2}\mathbb{E}_{y^*,\eta,\tilde{\boldsymbol{\beta}}}\left[\boldsymbol{\gamma}_m^\top\left(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} + \sum_{\ell=1}^{L} 1\{g[i[\ell]]=m\}\beta_{j[\ell]}^2 \mathbf{z}_{i[\ell]}\mathbf{z}_{i[\ell]}^\top\right)\boldsymbol{\gamma}_m\right. \\
&\qquad\qquad \left. -2\boldsymbol{\gamma}_m^\top\left\{\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\gamma}} + \sum_{\ell=1}^{L} 1\{g[i[\ell]]=m\}\mathbf{z}_{i[\ell]}\beta_{j[\ell]}(y_\ell^* - \alpha_{j[\ell]} - \beta_{j[\ell]}\eta_{i[\ell]})\right\}\right] + \text{const.}
\end{aligned}
$$

Thus, the variational distribution is given by equation (57).

Group-level Variances. Finally, we derive the variational distribution for the group-level variance parameter σ_m^2.

$$
\begin{aligned}
\log q(\sigma_m^2) &= \mathbb{E}_\eta\left[\log p(\sigma_m^2) + \sum_{n=1}^{N}\mathbf{1}\{g[n]=m\}\log p(\eta_n \mid \sigma_m^2)\right] + \text{const.} \\
&= \mathbb{E}_\eta\left[-\left(\frac{\nu_\sigma + \sum_{n=1}^{N}\mathbf{1}\{g[n]=m\}}{2} + 1\right)\log\sigma_m^2 - \frac{1}{2\sigma_m^2}\left(s_\sigma^2 + \sum_{n=1}^{N}\mathbf{1}\{g[n]=m\}\eta_n^2\right)\right] + \text{const.}
\end{aligned}
$$

Thus, the approximating distribution is the inverse-gamma distribution given in equation (58).