Fast and Accurate Estimation of Hierarchical Models using Variational Algorithms

Max Goplerud*

September 7, 2019

Abstract

Hierarchical models are commonly used in political science to address unobserved heterogeneity and model dependencies between observations. Inference for these models is typically conducted by maximum likelihood estimation or fully Bayesian analysis. Both methods can be slow on even moderately sized datasets, especially for non-linear models. This paper addresses this by deriving new algorithms for finding the maximum likelihood estimate using fast variational Expectation Maximization (EM), leveraging recent advances in data augmentation. These algorithms outperform existing variational methods on simulated data and closely recover the point estimates of "gold standard" methods on both simulated and actual datasets. I also provide a method for calculating approximate standard errors that has reasonable performance. I re-examine two papers in detail and show that simple extensions (e.g. adding an additional random effect) dramatically increase the computational cost for existing methods but only lead to modest increases for the variational approach.

Key Words: hierarchical models, variational inference, data augmentation, Expectation Maximization

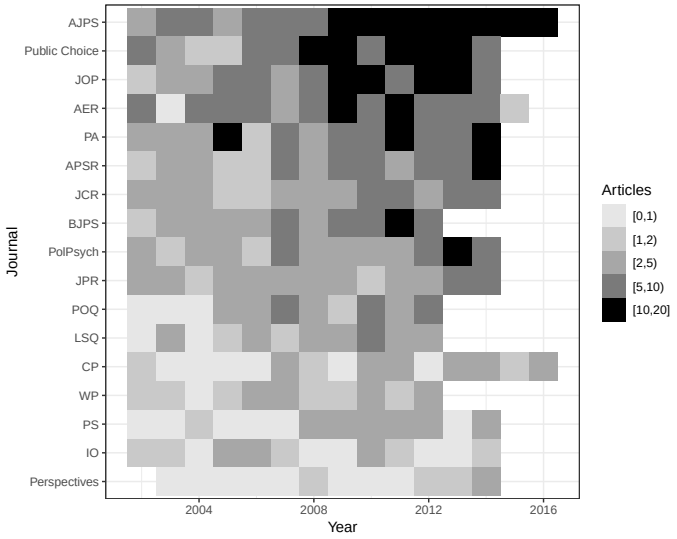
Hierarchical models—often known as multilevel models or "random effects" models—are ubiquitous in the social sciences. While they can take on many forms, this paper focuses on hierarchical models that place a mean-zero (multivariate) normal distribution on the random effects (Gelman and Hill 2006). Some common uses in political science are to address unobserved heterogeneity (Clark and Linzer 2015; Bell and Jones 2015), allow effects to vary across units (e.g. country; Stegmueller 2013) explicitly incorporate dependence between observations (Steenbergen and Jones 2002), and extrapolate survey data to smaller geographic units (Park, Gelman, and Bafumi 2004; Lax and Phillips 2009; Ghitza and Gelman 2013). Figure 1 shows a rough estimate of the number of articles per year in top political science journals that discuss or employ hierarchical models.[1]

*Draft. Current version available at `http://mgoplerud.com/papers/Goplerud_VI.pdf`. I thank Florian Stoeckel and Joscha Legewie for sharing their replication data. Naoki Egami, Shusei Eshima, June Hwang, Kosuke Imai, Gary King, Shiro Kuriwaki, Sun Young Park, Casey Petroff, Marc Ratkovic, Tyler Simko, Diana Stanescu, Dustin Tingley, and Soichiro Yamauchi all gave helpful comments on earlier versions of this paper. All remaining errors are my own.

[1]These results come from a search of the full text of articles on JSTOR for (a) the mention of a key textbook or

It shows clearly the (increasing) popularity of these methods across journals that span political science.

Figure 1: Number of Articles Discussing Hierarchical Models



Note: Each cell represents the number of articles containing a mention of hierarchical models, see Appendix A for the exact set of keywords examined. The journals are ordered from most articles overall to the fewest. Journals are listed by their standard abbreviations, e.g. "AJPS" stands for the "American Journal of Political Science", listed in Appendix A.

Unfortunately, inference for these models—especially in the non-linear case—becomes relatively challenging as the likelihood function contains an intractable integral. There are two popular methods ("gold standards") for solving this problem (Stegmueller 2013): First, one can approximate the integral numerically as done by software such as `lme4` (Bates et al. 2015). Second, one can use a fully Bayesian approach and sample from the joint distribution of all of the parameters of the model. This is easily done in software such as `STAN` (Carpenter et al. 2017) with a user-friendly interface from software such as `brms` (Bürkner 2017).[2] The key downside of these methods is that

resource on hierarchical models (Steenbergen and Jones 2002; Gelman and Hill 2006; Rabe-Hesketh, Skrondal, and Pickles 2004; Rabe-Hesketh and Skrondal 2008), (b) the word "hierarchical" or "multilevel" followed by a word such as "regression", "model", "logit', "probit", etc. or (c) a phrase such as "random effect", "random slope", or "random intercept". Appendix A provides full details.

[2]This paper does not dive into the difference between these methods, see Stegmueller (2013) for a detailed discussion. There are also estimation methods not considered here, see Steenbergen and Jones (2002) for an overview in the case of the linear model.

they can be very slow. As shown in the empirical applications, existing estimation algorithms take hours to run on only moderately sized datasets and may take many hours or even days to run if researchers include simple theoretically-justified extensions of the model. This constrains researchers insofar as computational costs prohibit estimating models that test their theories and the robustness of their results to alternative specifications.

This paper addresses this problem by deriving a set of new algorithms for estimating (non-linear) hierarchical models. It does this by bringing together two strands of work in statistics; first, it notes that for a *linear* hierarchical model, one can use a simple iterative algorithm to find the maximum likelihood estimate by Expectation Maximization (EM) (e.g. Dempster, Laird, and Rubin 1977; Laird and Ware 1982; Meng and Van Dyk 1998). Such an algorithm is very fast and scalable insofar as it only relies on iterative least squares. Second, it notes that for non-linear *non-hierarchical* models, recent advances from statistics allow for a simple and tractable form of data augmentation for binary, count, and multinomial models using "Polya-Gamma" data augmentation (Polson, Scott, and Windle 2013). For those models, it allows one to again use simple EM algorithms to find the maximum likelihood estimate. For a fully Bayesian approach, these algorithms can be combined together easily. However, this again inherits the slowness of existing "gold standard" methods.

I thus derive a set of new algorithms that rely on variational EM, i.e. EM with a variational approximation, to estimate these models. This approach conjectures that an approximation in the algorithm can bring dramatic gains in speed at limited cost in performance versus a slow, exact, method; see Grimmer (2011), Blei, Kucukelbir, and McAuliffe (2017), and Wang and Blei (2018) for reviews of variational inference in general. In contrast to existing work on variational inference for hierarchical models (e.g. Hall, Ormerod, and Wand 2011; Kucukelbir et al. 2017), relying on Polya-Gamma data augmentation allows for a less restrictive variational assumption (mean-field) while still allowing tractable inference as iterative least squares.

Relying on a variational approximation pays off handsomely for non-linear hierarchical models. Using this new variational algorithm, I show the following points on simulated data: (i) it is much faster than existing "gold standard" approaches; it outperforms existing variational approximations insofar as it is (ii) closer to the truth in simulated data and (iii) closer to the "gold standard" methods; (iv) using older results from the EM literature (Louis 1982), I derive standard errors

using the variational approximation that have good (although somewhat variable) coverage.

I then turn to a set of six recent papers that use non-linear hierarchical models. Each paper specifies a relatively parsimonious model (one or two random effects; usually no more than thirty thousand observations; around ten-to-twenty covariates) and estimates it using existing software—almost exclusively `STATA`'s default estimation tools (Rabe-Hesketh and Skrondal 2008) or `lme4` in `R` (Bates et al. 2015). Estimation is, however, rather slow in some cases with some models taking hours to estimate. I show that the variational algorithm derived in this paper provides considerable gains in speed cases with a speed-up ranging from a factor of about 3 to about 60. This comes at a limited cost in terms of the accuracy of the point estimates; in terms of standardized effects, the variational algorithm estimates are, on average, no more than 0.02 standard deviations away, and are often much closer. The associated approximate standard errors, where estimable, also quite close. I then focus on two papers (Stoeckel 2013; Legewie and Schaeffer 2016) and show that extensions of their original results (e.g. adding an additional random effect) dramatically increases the computational time of the original analysis, whereas it only moderately increases the run-time of the variational method. I then use those variational estimates to explore the results in the original papers.

Overall, the results demonstrate that for preliminary exploration, variational EM based on Polya-Gamma augmentation allow for a fast way to fit relatively complex models and recover very similar point estimates and approximate standard errors. As these methods are, however, approximate, it is worth stressing that for final results, the researcher should rely upon an "exact" method if possible. The estimation of said exact method, however, can be likely dramatically sped up by relying on the variational EM estimates as initial values.

1 Data Augmentation for Tractable Inference

The core of this paper relies on leverages recent results in data augmentation by Polson, Scott, and Windle (2013). Data augmentation can be concisely summarized as follows: Assume a data generating process where some outcome y_i is generated given a vector of parameters $\boldsymbol{\theta}$, i.e. $p(y_i|\boldsymbol{\theta})$; in most cases, estimating $\boldsymbol{\theta}$ is difficult given the non-normality of the data generating process. Data augmentation suggests that there may sometimes be a different variable z_i that is unobserved

but, if known, would make inference tractable (Wei and Tanner 1990). Specifically, if $p(y_i, z_i|\boldsymbol{\theta})$ is tractable, i.e. usually its logarithm is quadratic in terms of $\boldsymbol{\theta}$, and $p(z_i|y_i, \boldsymbol{\theta})$ is tractable, then there is a strategy for performing inference that is fast and stable (guaranteed to converge). This can be either via an Expectation-Maximization algorithm (Dempster, Laird, and Rubin 1977) to find the maximum likelihood estimate or via Markov Chain Monte Carlo to sample from the full posterior in a Bayesian framework.

To set up the discussion of Polya-Gamma augmentation, it is worth briefly recalling a very similar type of data augmentation for probit regression.[3] For this problem, the maximum likelihood solution can be cast as follows, where there are N observations with covariates $\boldsymbol{x}_i$ for each observation i. The outcome $y_i \in \{0, 1\}$.

$$\hat{\boldsymbol{\beta}}_{MLE} = \arg\max_{\boldsymbol{\beta}} \ln L(\boldsymbol{\beta}); \quad L(\boldsymbol{\beta}) = \prod_{i=1}^{N} \left[\Phi(\boldsymbol{x}_i^T\boldsymbol{\beta})\right]^{I(y_i=1)} \left[1 - \Phi(\boldsymbol{x}_i^T\boldsymbol{\beta})\right]^{1-I(y_i=1)} \tag{1}$$

This can be easily solved by a variety of numerical methods, e.g. Fisher Scoring. However, this can be recast in terms of latent variables z_i. Each z_i can be thought of as the latent "propensity" or "utility" for a positive outcome and if $z_i > 0$, then y_i is observed to be one. If $z_i < 0$, then y_i is observed to be zero. The "augmented" or "complete" data likelihood reflects the probability of seeing *both* y_i and z_i given the parameters. Equation 2 shows this for the probit case:

$$L_c(\boldsymbol{\beta}) = \sqrt{2\pi}\left[\exp\left(\frac{-(z_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2}{2}\right) I(z_i > 0)\right]^{I(y_i=1)} \left[\exp\left(\frac{-(z_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2}{2}\right) I(z_i \leq 0)\right]^{1-I(y_i=1)} \tag{2a}$$

$$= \sqrt{2\pi}\exp\left(\frac{-(z_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2}{2}\right) [I(z_i > 0)]^{I(y_i=1)} [I(z_i \leq 0)]^{1-I(y_i=1)} \tag{2b}$$

This re-arrangement has the following desirable interpretation: In the augmented likelihood, z_i is normally distributed given the linear predictor ($\boldsymbol{x}_i^T\boldsymbol{\beta}$). Thus, if z_i were known, one could simply find the maximum likelihood of the augmented likelihood by linear regression! The problem is that

[3] See Liu, Rubin, and Wu (1998) and Albert and Chib (1993) for a discussion of this model in a maximum likelihood and Bayesian context, respectively. Goplerud et al. (2018) provides a similar discussion to this section in the context of multi-level regression with post-stratification.

z_i is fundamentally unobservable. The crux of the EM algorithm is to note, however, that we can find the maximum likelihood by the following iterative procedure (Liu, Rubin, and Wu 1998):

Algorithm 1 Probit Regression via EM

Set: $\boldsymbol{\beta}^{(0)}$, T

For t in $1, \cdots, T$

E-Step: Calculate the conditional distribution of $q(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}^{(t)})$. This factorizes such that z_i are conditionally independent truncated normals with the following mean, where $\phi(.)$ and $\Phi(.)$ denote the normal PDF and CDF, respectively:

$$z_i^* = E[z_i|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}^{(t)}] = \begin{cases} \boldsymbol{x}_i^T \boldsymbol{\beta}^{(t-1)} + \frac{\phi(\boldsymbol{x}_i^T \boldsymbol{\beta}^{(t)})}{1-\Phi(-\boldsymbol{x}_i^T \boldsymbol{\beta}^{(t)})} & \text{if} \quad y_i = 1 \\ \boldsymbol{x}_i^T \boldsymbol{\beta}^{(t)} - \frac{\phi(\boldsymbol{x}_i^T \boldsymbol{\beta}^{(t)})}{\Phi(-\boldsymbol{x}_i^T \boldsymbol{\beta}^{(t)})} & \text{if} \quad y_i = 0 \end{cases}$$

M-Step: Maximize the expectation of log of the *complete* (augmented) likelihood with respect to $\boldsymbol{\beta}$, where the expectation is taken with respect to the distribution found in the E-Step. $\boldsymbol{z}^*$ stacks the z_i^*.

$$\boldsymbol{\beta}^{(t)} = \arg\max_{\boldsymbol{\beta}} E_{p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}^{(t-1)})} \left[\ln L_c(\boldsymbol{\beta})\right] = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{z}^*$$

Recasting probit regression in this form has a number of benefits. First, it gives an interpretation of the model in terms of least squares conducted on some latent variable. This permits the maximum likelihood estimate to be obtained via iterative least squares. This procedure is fast, scalable and stable insofar as each step of the EM algorithm is guaranteed to increase the (log-)likelihood and, if the MLE exists and is unique, will converge to the optimal solution(McLachlan and Krishnan 2008).[4]

1.1 Polya-Gamma Augmentation

Polya-Gamma data augmentation applies the principles above to the logistic link (Polson, Scott, and Windle 2013). In addition to standard regression analysis (Polson, Scott, and Windle 2013; Pillow and Scott 2012; Zhou et al. 2012), it is often used in more complex models where part of the likelihood contains a logistic link. Such applications include topic models (Chen et al. 2013; Linderman, Johnson, and Adams 2015; Glynn et al. 2018), ideal point estimation (Goplerud 2018), sparse models with non-linear outcomes (Makalic and Schmidt 2015; Betancourt, Rodríguez, and Boyd 2017; Goplerud et al. 2018), among others. As the name suggests, the data augmentation

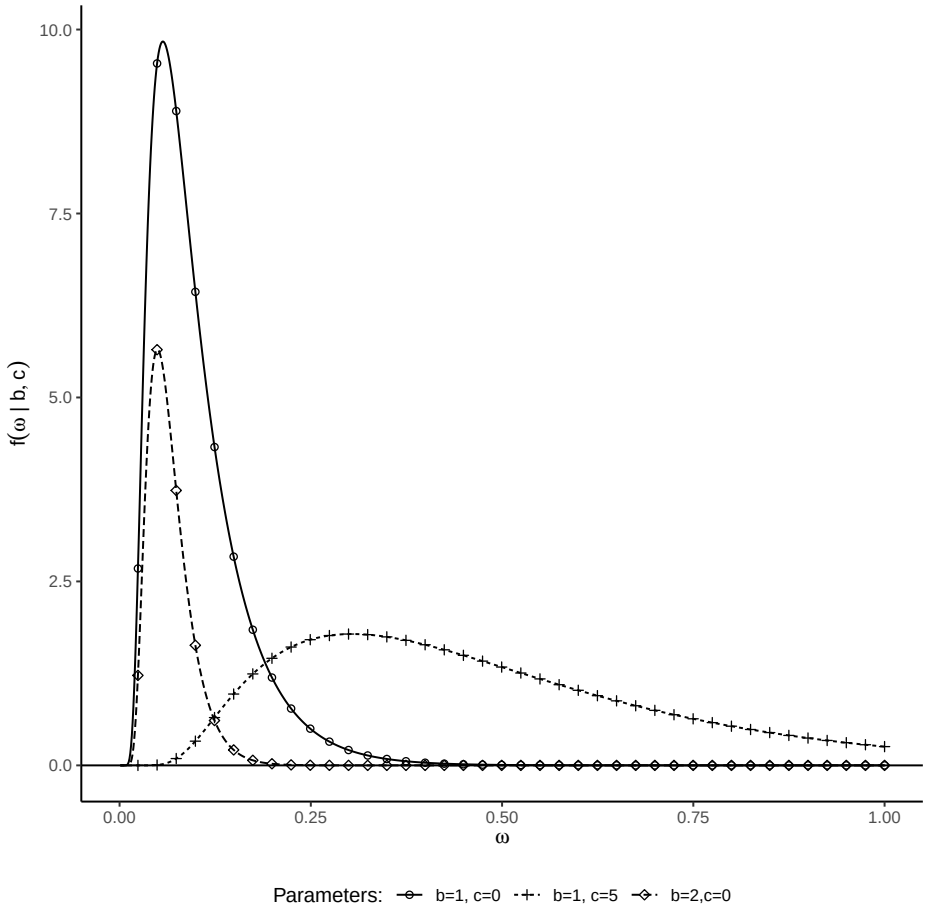
[4] A wide literature has discussed a number of techniques to further accelerate the EM algorithm, e.g. Neal and Hinton (1998). McLachlan and Krishnan (2008) provides a comprehensive review.

scheme relies on the "Polya-Gamma" random variable developed in Polson, Scott, and Windle (2013). The most intuitive definition considers it as a scaled sum of independent random variables in the way that a χ^2 random variable is the sum of squared independent standard normals. However, it is unusual in that (a) the sum is weighted in a very particular way and (b) that it is an *infinite* sum of random variables. Formally, a Polya-Gamma variable, $\omega \sim PG(b, c)$, with $b > 0$ and $c \in \mathbb{R}$, can be defined as follows.

$$\omega = \frac{1}{2\pi^2} \sum_{i=1}^{\infty} \frac{Z_i}{(k - 1/2)^2 + c^2/(4\pi^2)}; \quad Z_i \sim^{i.i.d.} Gamma(b, 1) \tag{3}$$

The variable itself looks similar to a Gamma random variable and, as b increases, it becomes approximately normal (Glynn et al. 2018). Figure 2 plots some density curves.

Figure 2: Polya-Gamma Densities



Note: This figure plots the probability density function for a number of Polya-Gamma random variables whose parameters are shown in the legend below the figure. For simplicity, the infinite sum is truncated after 100 terms.

ω has a density formula that is also equally challenging as it can only be expressed as an infinite sum:

$$p(\omega|b,c) = \cosh^b(c/2)\frac{2^{b-1}}{\Gamma(b)}\sum_{n=0}^{\infty}(-1)^n\frac{\Gamma(n+b)(2n+b)}{\Gamma(n+1)\sqrt{2\pi\omega^3}}\exp\left(\frac{-(2n+b)^2}{8\omega} - c^2/2\omega\right) \tag{4}$$

However, the definition of this variable leads to key results that facilitate a data augmentation scheme (Polson, Scott, and Windle 2013):

- For any real ψ, the follow identity holds:

$$\frac{\exp(\psi)^a}{(1+\exp(\psi))^b} = 2^{-b}\exp(s\psi)\int_0^{\infty}\exp(-\omega\psi^2/2)p(\omega)d\omega \quad p(\omega) \sim PG(b,0); \quad s = a - b/2 \tag{5}$$

- The expectation of a Polya-Gamma random variable has a simple closed form:

$$E(\omega) = \frac{b}{2c}\tanh(c/2); \quad \omega \sim PG(b,c); \quad \tanh(x) = \frac{\exp(2x)-1}{\exp(2x)+1} \tag{6}$$

- The joint and conditional distributions of ψ and ω can be characterized as follows:

$$p(\psi,\omega) = 2^{-b}\exp(s\psi - \omega\psi^2/2)p(\omega) \tag{7a}$$

$$p(\psi|\omega) \sim N\left(\frac{s}{\omega}, \frac{1}{\omega}\right) \tag{7b}$$

$$p(\omega|\psi) \sim PG(b,\psi) \tag{7c}$$

From these properties, it is possible to create a data augmentation scheme for an EM algorithm or a fully Bayesian Gibbs Sampler.[5] With these results, Polson, Scott, and Windle (2013) derived an EM algorithm for (non-hierarchical) logistic regression. Equation 8 shows the likelihood for a

[5]The later fully Bayesian approach requires a further point: That it is possible to easily sample from Polya-Gamma random variables. Polson, Scott, and Windle (2013) derive an efficient sampler for doing so that is implemented in two `R` packages (`helloPG` and `BayesLogit`). Some recent work notes that this sampler is slow when b is large (Glynn et al. 2018) and fully Bayesian schemes based on such variables exhibit poor mixing (Johndrow et al. 2018). Glynn et al. (2018) suggests mitigating the former by noting that if b is large, the Polya-Gamma is well approximated by a normal random variable. The later is not a problem for EM algorithms, but might affect the fully Bayesian implementation of this scheme for binomial regression with large N_i, see Table 1.

logistic regression and the augmented likelihood with Polya-Gammas (L_c).

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N} \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})^{I(y_i=1)}}{1+\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})} \tag{8a}$$

$$L_c(\boldsymbol{\beta}) = \prod_{i=1}^{N} 2^{-1} \exp\left([I(y_i=1)-1/2](\boldsymbol{x}_i^T \boldsymbol{\beta}) - \omega_i(\boldsymbol{x}_i^T \boldsymbol{\beta})^2/2\right) \tag{8b}$$

As in the probit case, the fact that the full conditional of $\omega_i|\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{y}$ has a tractable distribution means that a simple EM algorithm exists to find the maximum likelihood estimate. The algorithm can be written as a simple two-step procedure:

Algorithm 2 Logistic Regression via EM

Set: $\boldsymbol{\beta}^{(0)}, T$

For t in $1, \cdots, T$

E-Step: Calculate the conditional distribution of $q(\boldsymbol{\omega}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}^{(t)})$. In this case, it simplifies to conditionally independent Polya-Gammas with parameters $(1, \boldsymbol{x}_i^T \boldsymbol{\beta}^{(t-1)})$. The relevant expectation is shown below.

$$\omega_i^* = E[\omega_i] = \frac{1}{2(\boldsymbol{x}_i^T \boldsymbol{\beta}^{(t-1)})} \tanh\left(\frac{\boldsymbol{x}_i^T \boldsymbol{\beta}^{(t-1)}}{2}\right)$$

{**M-Step**: $\boldsymbol{\beta}$, where $\boldsymbol{s}$ is a vector such that $s_i = I(y_i = 1) - 1/2$ and $\boldsymbol{\Omega}^*$ is a matrix that diagonally stacks each $E[\omega_i]$.

$$\boldsymbol{\beta}^{(t)} = \left(\boldsymbol{X}^T \boldsymbol{\Omega}^* \boldsymbol{X}\right) \boldsymbol{X}^T \boldsymbol{s}$$

In addition to logistic regression, there are two other relevant scenarios for this paper: negative binomial and multinomial regression. Table 1 writes the generative model and the likelihood for each case.

The first two models (binomial and negative binomial) clearly contain a logistic link. One can also re-arrange the multinomial likelihood to have a similar functional form for some set of coefficients $\boldsymbol{\beta}_k$ by conditioning on all other coefficients (Polson, Scott, and Windle 2013).[7] In the

[7] Specifically,

$$p_{ik} = \frac{\exp(\psi_{ik})}{\sum_k \exp(\psi_{ik})} = \frac{\exp(\psi_{ik})}{C_{ik}+\exp(\psi_{ik})} = \frac{\exp(\psi_{ik}-\ln C_{ik})}{1+\exp(\psi_{ik}-\ln C_{ik})}; \quad C_{ik} = \sum_{l \neq k} \exp(\psi_{il})$$

Table 1: Generative Models

1. Binomial

$$Y_i \sim \text{Binomial}(N_i, p_i); \quad p_i = \frac{\exp(\psi_i)}{1+\exp(\psi_i)}; \quad \psi_i = \boldsymbol{x}_i^T\boldsymbol{\beta} \tag{9a}$$

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N} \binom{N_i}{y} p_i^y (1-p_i)^{N_i-y} = \prod_{i=1}^{N} \binom{N_i}{y} \frac{\exp(\psi_i)^y}{[1+\exp(\psi_i)]^{N_i}} \tag{9b}$$

2. Count (Negative Binomial)[6]

$$Y_i \sim \text{NB}(r, 1-\gamma_i); \quad \gamma_i = \frac{\exp(\psi_i - \ln r)}{1+\exp(\psi_i - \ln r)}; \quad \psi_i = \boldsymbol{x}_i^T\boldsymbol{\beta} \tag{10a}$$

$$L(\boldsymbol{\beta}, r) = \prod_{i=1}^{N} \frac{\Gamma(k+r)}{\Gamma(r)\Gamma(y+1)} \gamma_i^y (1-\gamma_i)^r = \prod_{i=1}^{N} \frac{\Gamma(k+r)}{\Gamma(r)\Gamma(y+1)} \frac{\exp(\psi_i - \ln r)^y}{[1+\exp(\psi_i - \ln r)]^{r+y}} \tag{10b}$$

3. Multinomial with K categories,

$$Y_i \sim \text{Multinomial}(1, \boldsymbol{p}_i); \quad [\boldsymbol{p}_i]_k = p_{ik} = \frac{\exp(\psi_{ik})}{\sum_k \exp(\psi_{ik})}; \quad \psi_{ik} = \boldsymbol{x}_i^T\boldsymbol{\beta}_k \tag{11a}$$

$$L\left(\{\boldsymbol{\beta}\}_{k=1}^{K-1}\right) = \prod_{i=1}^{N}\prod_{k=1}^{K} p_{ik}^{I(Y_i=k)} \tag{11b}$$

case of a model without random effects, an exact EM algorithm using Polya-Gamma augmentation can be used to find the maximum likelihood estimate for all three scenarios.[8] This also means that a fully Bayesian analysis can be conducted using a Gibbs Sampler insofar as the fully conditionals of the coefficients ($\boldsymbol{\beta}$) are multivariate normally distributed given the Polya-Gamma variables.[9]

2 Estimating Hierarchical Models by Variational Inference

A key extension of these likelihood models is to add hierarchical or structure via random effects (Gelman and Hill 2006). For exposition, I consider a logistic regression with a single random intercept. Appendix B derives full results for all three models with arbitrary random effect structures (cross-classified, random slopes, etc.). Following Gelman and Hill (2006), the J random effects are

[8] Negative binomial and multinomial models rely on the AECM algorithm as outlined in Appendix B and studied theoretically in a general case by Meng and Van Dyk 1997.

[9] The one exception is updating the dispersion parameter r in the negative binomial regression, see Appendix B for details.

assumed to be drawn independently from a normal distribution: $\alpha_j \sim N(0, \sigma_\alpha^2)$. I use $\boldsymbol{\alpha}$ to denote the vector that stacks all of the random effects together. I further assume that each observation i is assigned to a single group j denoted by $j[i]$. As before, the goal is to find the maximum likelihood over *both* $\boldsymbol{\beta}$ and σ_α^2, integrating away the random effects. Equation 12 shows the likelihood:

$$L(\boldsymbol{\beta}, \sigma_\alpha^2) = \prod_{j=1}^{J} \left[\int \left[\prod_{i:j[i]=j} \frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta} + \alpha_j)^{I(y_i=1)}}{1 + \exp(\boldsymbol{x}_i^T\boldsymbol{\beta} + \alpha_j)} \right] f(\alpha_j|\sigma_\alpha^2) d\alpha_j \right] \tag{12}$$

$$f(\alpha_j|\sigma_\alpha^2) = (2\pi\sigma_\alpha^2)^{-1/2} \exp\left(-\alpha_j^2/[2\sigma_\alpha^2]\right)$$

Unfortunately, this is challenging to optimize. Traditional approaches either rely on approximating the integral numerically (e.g. Bates et al. 2015) or on a fully Bayesian approach. In theory, however, one could apply an EM algorithm! This would require augmenting on *both* the random effects ($\boldsymbol{\alpha}$) and the Polya-Gamma random variables ($\{\omega_i\}_{i=1}^N$). This yields a complete data likelihood as follows, where $f(\omega_i|1,0)$ is the density of a Polya-Gamma variable with parameters $(1,0)$.

$$L_c(\boldsymbol{\beta}, \sigma_\alpha^2) = \prod_{j=1}^{J} f(\alpha_j|\sigma_\alpha^2) \prod_{i=1}^{N} \frac{1}{2} \exp\left([I(y_i = 1) - 1/2] \left[\boldsymbol{x}_i^T\boldsymbol{\beta} + \alpha_{j[i]}\right] - \omega_i(\boldsymbol{x}_i^T\boldsymbol{\beta} + \alpha_{j[i]})^2/2 \right) f(\omega_i|1,0) \tag{13}$$

A key result is that the maximum of the log of Equation 13 with respect to $\boldsymbol{\beta}$ is a simple least squares solution and the maximum with respect to σ_α^2 has a similarly simple closed form. Thus, if one knew $\boldsymbol{\alpha}$ and $\{\omega_i\}$, the question reduces to (effectively) a least squares problem. As discussed previously, therefore, an EM algorithm could be used to solve this problem by iterating between calculating the conditional distribution of the latent variables given the parameters, i.e. $q(\boldsymbol{\alpha}, \{\omega_i\}|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\beta}, \sigma_\alpha^2)$, and maximizing the expected log of Equation 13. Unfortunately, that key distribution is intractable. While each of the conditionals are tractable (see Appendix B),[10] neither has a marginal distribution that has a simple closed form. An "exact", although computationally burdensome, strategy would be to use a Monte Carlo EM algorithm (Wei and Tanner 1990) where one samples from $q(\boldsymbol{\alpha}, \{\omega_i\}|\boldsymbol{\beta}, \sigma_\alpha^2, \boldsymbol{X}, \boldsymbol{y})$ using a Gibbs Sampler and then approximates the

[10]Specifically, $q(\boldsymbol{\alpha}|\{\omega_i\}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\beta}, \sigma_\alpha^2)$ is normal and $q(\{\omega_i\}|\boldsymbol{\alpha}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\beta}, \sigma_\alpha^2)$ factors into independent Polya-Gammas. This shows how the Polya-Gamma approach makes a fully Bayesian sampler much more tractable insofar as all of the conditionals have tractable forms—as $q(\boldsymbol{\beta}|\boldsymbol{\alpha}, \{\omega_i\}, \boldsymbol{X}, \boldsymbol{y})$ is normally distributed.

expectations using those sampled latent variables.

This paper relies on a different strategy: variational EM. It approximates the intractable q distribution with a simpler one.[11] Variational methods make a trade-off where the approximating assumption dramatically improves computational speed at, hopefully, a minimal cost to the performance of the algorithm (see Grimmer 2011; Blei, Kucukelbir, and McAuliffe 2017 for a reviews; Imai, Lo, and Olmsted 2016 provides an application in the context of ideal point estimation). A variety of approximation assumptions exist: the most common assumption, that I use here, is a "mean-field" assumption. It assumes that q can be approximated by independent distributions: $q(\boldsymbol{\alpha}, \{\omega_i\}|-) \approx \tilde{q}_\alpha(\boldsymbol{\alpha})\tilde{q}_\omega(\{\omega_i\})$. This differs from existing variational approaches for hierarchical models, discussed shortly, that rely on stronger assumptions about the distributional form of the approximation distributions.

Appendix B provides full details and discussion of mean-field variational inference but the key point is that Polya-Gamma augmentation, unlike existing variational schemes, allows the mean-field assumption to be immediately tractable. The variational E-step thus consists of finding the approximating distributions that are closest to the (intractable) joint distribution. Equation 14 reports the variational distributions for the E-Step of the algorithm with a single random effect.[12]

$$\begin{aligned}
\tilde{q}(\alpha_j) &\sim N(\tilde{\mu}_{j,\alpha}, \tilde{\Lambda}_{j,\alpha}); \quad \tilde{\Lambda}_{j,\alpha} = \left(\sum_{i:j[i]=j} \omega_i + (\sigma_\alpha^2)^{-1} \right)^{-1} \\
\tilde{\mu}_{j,\alpha} &= \tilde{\Lambda}_{j,\alpha}^{-1} \left(\sum_{i:j[i]=j} I(y_i = 1) - 1/2 - E_{\tilde{q}(\omega_i)}[\omega_i](\boldsymbol{x}_i^T \boldsymbol{\beta}) \right)
\end{aligned} \tag{14a}$$

$$\tilde{q}(\omega_i) \sim PG\left(1, \tilde{c}_i\right); \quad \tilde{c}_i = \sqrt{\left(\boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{z}_i^T E_{\tilde{q}(\alpha_{j[i]})}[\alpha_{j[i]}] \right)^2 + Var_{\tilde{q}(\alpha_{j[i]})}(\alpha_{j[i]})} \tag{14b}$$

Thus, with these in hand, the variational EM algorithm can proceed by (i) updating the variational parameters given the current estimate of $\boldsymbol{\beta}$ and σ_α^2 and then (ii) optimizing the expectation

[11] Variational EM is a way of finding the "variational frequentist estimate" (Wang and Blei 2018). It differs from traditional "variational Bayes" by ways briefly discussed in Appendix B and in more detail in Wang and Blei (2018).

[12] Note that in this case, $\tilde{q}_\alpha(\boldsymbol{\alpha})$ factorizes into independent $\tilde{q}_\alpha(\alpha_j)$ distributions. Appendix B provides full details for an arbitrary number of random effects.

of the log of Equation 13, over the variational distributions, and iterating until convergence.[13]

A key benefit of this framework is that it also permits calculation of approximate standard errors on the parameters using a long-standing result from the EM literature. Louis (1982) notes that the information matrix for the non-augmented log-likelihood (i.e. the log of Equation 12) can be found using only the complete data log-likelihood (i.e. the log of Equation 13) and its score and Hessian. Equation 15 shows the result, where $\boldsymbol{\theta}$ denotes the concatenation of coefficients ($\boldsymbol{\beta}$) and the random effects parameters (e.g. σ_α^2).

$$- E\left[\frac{\partial \ln L_c}{\partial \boldsymbol{\theta}\boldsymbol{\theta}^T}\right] - E\left[\left(\frac{\partial \ln L_c}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \ln L_c}{\partial \boldsymbol{\theta}}\right)^T\right] + E\left[\left(\frac{\partial \ln L_c}{\partial \boldsymbol{\theta}}\right)\right] E\left[\left(\frac{\partial \ln L_c}{\partial \boldsymbol{\theta}}\right)\right]^T \tag{15}$$

As is standard, the information matrix can be inverted and evaluated at the estimated parameters to get the usual asymptotic variance-covariance matrix. Equation 15 cannot be evaluated directly, however, for reasons outlined above. I thus get "approximate" standard errors by evaluating those expectations with respect to the variational distributions found at the variational EM estimate.[14] At the moment, I only use this approximation for the logistic and negative binomial models.

2.1 Other Methods for Variational Inference with Hierarchical Models

It is worth stressing, however, that since all variational methods rely on some approximation, it is not necessarily the case that one set of approximations is better than others in particular settings. The key distinction between the variational EM scheme based on Polya-Gamma augmentation and others is that they seek to find a variational distribution for the random effects ($\boldsymbol{\alpha}$) and the parameters ($\boldsymbol{\beta}$) *alone*, insofar as they do not use data augmentation for the non-linear part of the generative model itself.[15] It is thus important to compare my framework against existing variational

[13]Specifically, the variational E-Step takes $\boldsymbol{\beta}^{(t-1)}$ and $[\sigma_\alpha^2]^{(t-1)}$, $\tilde{\boldsymbol{\mu}}_\alpha^{(t-1)}$, $\{\tilde{c}_i^{(t-1)}\}$ and updates the parameters in Equations 14a and b, sequentially. It then performs the M-Step as discussed in the main text.

[14](Ormerod and Wand 2012) suggest a related method for calculating standard errors given their variational approach: Specifically, they suggest inverting Fisher's information matrix for the main parameters and the variational parameters and then taking only the sub-matrix that corresponds to the main parameters. Another approach would be to evaluate Equation 15 using the Gibbs Sampler to get the exact distribution of the augmentation variables; this would, if run to convergence, exactly evaluate the information matrix at the variational EM point estimates.

[15]Goplerud et al. (2018) uses a related but importantly distinct algorithm for multilevel logistic regression with applications to MRP, although their major focus is on sparsity and not the estimation of random effects. Their approximation is algorithmically similar to the one employed here, but does not correspond to a well-defined variational procedure and thus lacks guarantees to increase a lower-bound on the objective function.

methods for hierarchical models.

I focus on two existing strategies: One is specifically designed for variational inference on generalized linear mixed models; the other is a generic method implemented in commonly used statistical software.[16] The first ("Gaussian Variational Approximation" [GVA] see Ormerod and Wand 2012 for an overview) uses a Gaussian approximation, i.e. they assume that the random effects have a normal distribution but rely on quadrature or other approximate methods to do inference in the logistic case and other general linear models, with the exception of hierarchical Poisson models (Hall, Ormerod, and Wand 2011; Ormerod and Wand 2012). Another popular strategy uses a different form of variational inference ("Automatic Differentiation Variational Inference" [ADVI]; Kucukelbir et al. 2017) and is implemented for any model in `STAN` (Carpenter et al. 2017). Roughly, it works by transforming all variables to have support on the real line and then uses stochastic gradient ascent to maximize the variational objective. This strategy has strong benefits; first, it is "automatic" insofar as one can feed it an arbitrary differentiable model and it will perform variational inference; the results in this paper only work for a class of model (random effects with a logistic link) that is amenable to Polya-Gamma augmentation. This framework, unlike, the variational EM scheme developed here or GVA targets the posterior distribution on the coefficients, instead of a point estimate.[17]

All of these variational inference strategies are designed to produce estimates that, hopefully, closely approximate the maximum likelihood (or posterior) estimates; whether the variational strategy will succeed in this objective in general is a very difficult and unresolved question. While some existing work has made progress on this front (see Wang and Blei 2018), it is typically only possible to prove the consistency of this method in special cases; focusing on non-linear hierarchical models, Hall and co-authors (Hall, Ormerod, and Wand 2011; Hall et al. 2011) show that for a Poisson model with one covariate and one random effect, the GVA estimator is consistent; Wang and Blei (2018) use this to show that fully variational procedure is also consistent for the true posterior. An important area of future research should examine whether the mean-field approximation

[16]Some other approaches are worth noting: First, Jaakkola and Jordan 1997 (see also Stewart 2014) rely on the shape of the logistic likelihood to further bound the objective function and thus can be estimated without needing numerical integration at the expense of targeting a (possibly) less good objective. Zhou et al. (2012) uses a fully variational strategy to estimate overdispersed counts, albeit with a different way of modelling dispersion, that involves Polya-Gamma variables but no random effects.

[17]Thus, ADVI is a scheme for "variational Bayes" whereas GVA and the scheme in this paper are targeting the "variational frequentist objective" (Wang and Blei 2018).

that relies on Polya-Gamma augmentation can be shown to have similar properties. I conjecture, however, that the better performance of the Polya-Gamma augmentation scheme turns on the fact that it makes a weaker assumption (mean-field independence) versus the stronger distributional assumptions of the other methods.

3 Examining the Performance on Simulated Data

Given the difficulty of general theoretical guarantees, however, I first examine the performance of this variational approximation against simulated data. I consider a number of possible scenarios to see where and when the performance of these algorithms degrade versus the "gold standard" method (e.g. numerical integration via `lme4`, Bates et al. 2015, or, where that does not exist, a fully Bayesian approach via `STAN`, Carpenter et al. 2017).

I create a dataset with one-thousand observations for each simulation.[18] I examine four scenarios corresponding to common uses of hierarchical models in political science. In each case, I used this same simulation environment for a logistic, negative binomial (where the dispersion parameter $r = 5$), and multinomial outcomes (with four choices). I examine four scenarios corresponding to four realistic cases. In each case, each random effect has ten levels.

1. One Random Intercept: $\alpha_j \sim N(0, 1)$ with $J = 10$ groups. All observations assigned at random to groups.

2. Two Random Intercepts: $\alpha_j \sim N(0, 1)$ and $\gamma_g \sim N(0, 5)$. $J = 10$ and $G = 10$ groups. All observations assigned at random to groups. This model thus uses cross-classified random effects.

3. Random Slope and Intercept: $\alpha_j \sim N(0, 1)$ and $\gamma_j \sim N(0, 1)$. The generative model is $\boldsymbol{x}i_i^T\boldsymbol{\beta} + \alpha_{j[i]} + \gamma_{j[i]}x_{i,1}$. $J = 10$. All observations are assigned at random to groups.

4. Violation of Random Effects Assumption: Generate α_j from equally spaced divisions of the interval $[-3, 3]$. Add $\epsilon_i \sim N(0, 1)$ to each $x_{i,1}$ and assign to groups based on the decile of the jittered value.

[18] I generate ten covariates x_{ij} such that the correlation between x_{ij} and x_{ik} is $0.5^{|j-k|}$. I generate the corresponding fixed effects ($\boldsymbol{\beta}$) independently from a standard normal for the logistic and multinomial cases. For the negative binomial, they are drawn from a $N(0, 1/25)$ distribution.

In all cases, I compare my variational approach against the 'gold standard' method—for the logistic and negative binomial case, I rely on the R implementation in `lme4` (Bates et al. 2015). For the multinomial case, I rely on one chain of Hamiltonian Monte Carlo in `STAN` (Carpenter et al. 2017; Bürkner 2017). I compare these results against three variational options: variational EM based on Polya-Gammma augmentation, ADVI, and GVA (where applicable).[19]

To begin, I examine computational time. Table 2 shows, averaging across one-hundred simulations, the run-time of each approach in minutes. We see that all variational methods are much faster than the 'gold standard' method of numerical integration or Hamiltonian Monte Carlo.

Table 2: Run Time of Estimation Methods

Outcome	Simulation	lmer	HMC	VI-ADVI	VI-GVA	VI-PG
Negative Binomial	1	3.267		$0.160^{\dagger}$		0.021^{*}
Negative Binomial	2	4.171		$0.214^{\dagger}$		0.030^{*}
Negative Binomial	3	5.149		$0.123^{\dagger}$		0.031^{*}
Negative Binomial	4	4.361		$0.163^{\dagger}$		0.078^{*}
Logistic	1	$0.117^{\dagger}$		0.741	5.946	0.061^{*}
Logistic	2	0.208		0.047^{*}		$0.117^{\dagger}$
Logistic	3	$0.137^{\dagger}$		0.763	2.183	0.067^{*}
Logistic	4	0.231		0.053^{*}		$0.087^{\dagger}$
Multinomial	1		5.209	$1.166^{\dagger}$		0.533^{*}
Multinomial	2		5.776	0.258^{*}		$1.025^{\dagger}$
Multinomial	3		5.143	1.213^{*}		$1.973^{\dagger}$
Multinomial	4		6.203	0.263^{*}		$0.957^{\dagger}$

Note: * indicates the best performing method; † indicates the second best performing method. The values are the average run time of each method in minutes. "VI-" indicates a variational method described in the main text. "VI-PG" stands for Polya-Gamma based VI. `lmer` is the baseline for logistic and negative binomial outcomes;`HMC` for multinomial.

Next, I examine performance of the variational approach in two ways: "relative" and "objective" performance. The first corresponds to the idea that, in some sense, a variational approach is good insofar as it closely approximates the (slower) gold standard method—even if the absolute performance of *both* methods are poor. Table 3 shows the results and the superior performance of inference based on Polya-Gamma random variables. It recovers nearly exactly the same estimates, while GVA performs poorly, and ADVI performs somewhere in between. Across all simulation environments, the variational EM algorithm derived in this paper is closest to the gold standard.

[19] I used the implementation of GVA in the code provided by Ormerod and Wand (2012). This works only in the case of one random effect, Simulations 1 and 3, for the logistic outcome but could likely be extended further. Given its inferior performance to ADVI, I focus on that as the best (and most widespread) variational alternative.

In the case of the negative binomial and logistic regressions, it is nearly identical—being generally around 5 to 10 times closer than ADVI. It is worth noting, however, that the correlation in all cases is nearly perfect suggesting that all variational methods broadly capture the aggregate relationship between the coefficients.

Table 3: Discrepancy versus Baseline

(a) Mean Absolute Error

Outcome	Simulation	VI-ADVI	VI-GVA	VI-PG
Negative Binomial	1	0.017†		0.001*
Negative Binomial	2	0.028†		0.005*
Negative Binomial	3	0.024†		0.003*
Negative Binomial	4	0.189†		0.002*
Logistic	1	0.056†	0.129	0.010*
Logistic	2	0.056†		0.010*
Logistic	3	0.056†	0.251	0.012*
Logistic	4	0.055†		0.021*
Multinomial	1	0.027†		0.018*
Multinomial	2	0.032†		0.025*
Multinomial	3	0.029†		0.018*
Multinomial	4	0.033†		0.023*

(b) Correlation

Outcome	Simulation	VI-ADVI	VI-GVA	VI-PG
Negative Binomial	1	0.991†		1.000*
Negative Binomial	2	0.932†		1.000*
Negative Binomial	3	0.957†		0.998*
Negative Binomial	4	0.903†		1.000*
Logistic	1	1.000†	0.999	1.000*
Logistic	2	1.000†		1.000*
Logistic	3	1.000†	0.943	1.000*
Logistic	4	0.999†		1.000*
Multinomial	1	0.993†		0.998*
Multinomial	2	0.990†		0.997*
Multinomial	3	0.991†		0.994*
Multinomial	4	0.964†		0.997*

Note: * indicates the best performing method; † indicates the second best performing method. The values in Panel (a) report the average mean absolute error for each method versus the baseline, i.e. the average of the absolute differences between the coefficients averaged across 100 simulations. The values in Panel (b) report the correlation between the methods and the baseline, averaged across 100 simulations. "VI-" indicates a variational method described in the main text. "VI-PG" stands for Polya-Gamma based VI. `lmer` is the baseline for logistic and negative binomial outcomes;`HMC` for multinomial.

It is worth seeing, additionally, how the models perform in an "objective" sense: How far are they from the simulated truth? Table 4 shows that on average the mean (absolute) difference between

the Polya-Gamma variational approach is very comparable (and sometimes better!) than the gold-standard method as well as beating the other variational approaches across most simulations. It is only beaten by another variational approach, in terms of mean absolute error, for some multinomial simulations.

Table 4: Discrepancy versus Truth

(a) Mean Absolute Error

Outcome	Simulation	lmer	HMC	VI-ADVI	VI-GVA	VI-PG
Negative Binomial	1	0.043†		0.048		0.043*
Negative Binomial	2	0.050*		0.062		0.050†
Negative Binomial	3	0.038*		0.049		0.039†
Negative Binomial	4	0.056*		0.224		0.056†
Logistic	1	0.148†		0.177	0.157	0.144*
Logistic	2	0.158†		0.186		0.156*
Logistic	3	0.147†		0.174	0.260	0.145*
Logistic	4	0.157†		0.182		0.150*
Multinomial	1		0.152	0.152†		0.143*
Multinomial	2		0.169	0.166†		0.155*
Multinomial	3		0.139†	0.141		0.131*
Multinomial	4		0.166	0.164†		0.154*

(b) Correlation

Outcome	Simulation	lmer	HMC	VI-ADVI	VI-GVA	VI-PG
Negative Binomial	1	0.933†		0.920		0.933*
Negative Binomial	2	0.876†		0.794		0.877*
Negative Binomial	3	0.962*		0.905		0.957†
Negative Binomial	4	0.878†		0.792		0.878*
Logistic	1	0.986*		0.986	0.986†	0.986
Logistic	2	0.982*		0.981		0.982†
Logistic	3	0.985*		0.985†	0.931	0.985
Logistic	4	0.982*		0.982		0.982†
Multinomial	1		0.793*	0.790		0.792†
Multinomial	2		0.756*	0.753		0.755†
Multinomial	3		0.827†	0.819		0.828*
Multinomial	4		0.769*	0.762		0.766†

Note: * indicates the best performing method; † indicates the second best performing method. The values in Panel (a) report the average mean absolute error for each method versus the truth, i.e. the average of the absolute differences between the coefficients averaged across 100 simulations. The values in Panel (b) report the correlation between the methods and the truth, averaged across 100 simulations. "VI-" indicates a variational method described in the main text. "VI-PG" stands for Polya-Gamma based VI.

Finally, I look at the coverage of these methods using the approximate standard errors: I report the proportion of times the 95% confidence or credible interval associated with the method contains

the true coefficient, averaged across all fixed effects ($\boldsymbol{\beta}$) and simulations. As noted before, insofar as the variational distributions closely approximate the true distribution of the augmentation variables, this method (Louis 1982) will recover the same standard errors as the "gold standard" methods.[20] My novel variational strategy has relatively good nominal coverage—it is slightly worse than the "gold standard" method but close to nominal.

Table 5: Coverage of Methods

Outcome	Simulation	lmer	HMC	VI-ADVI	VI-GVA	VI-PG
Negative Binomial	1	0.946†		0.824		0.950*
Negative Binomial	2	0.947*		0.825		0.940†
Negative Binomial	3	0.953*		0.820		0.910†
Negative Binomial	4	0.937*		0.787		0.889†
Logistic	1	0.934†		0.738	0.853	0.934*
Logistic	2	0.942*		0.738		0.939†
Logistic	3	0.938*		0.747	0.764	0.936†
Logistic	4	0.932*		0.750		0.927†
Multinomial	1		0.646*	0.426†		
Multinomial	2		0.670*	0.438†		
Multinomial	3		0.631*	0.424†		
Multinomial	4		0.665*	0.428†		

Note: * indicates the best performing method; † indicates the second best performing method. All numbers represent the frequentist coverage, averaged across simulations and variables: Does the 95% confidence interval created by the implied standard errors contain the truth? It reports the coverage for the fixed effects ($\boldsymbol{\beta}$). "VI-" indicates a variational method described in the main text. "VI-PG" stands for Polya-Gamma based VI. `lmer` is the baseline for logistic and negative binomial outcomes;`HMC` for multinomial.

Overall, the results are promising: In simulated cases, my novel variational algorithm very closely corresponds to the existing numerical integration approach—far more than existing variational methods—while being comparable in speed to existing approximate inference strategies.

4 Applying Variational EM on Existing Studies

In this section, I turn to recent published papers that rely on non-linear hierarchical models. Table 6 begins by the estimated coefficients between models used in the six papers and the variational EM approximation.[21] In the case of running multiple models for each paper, I report results aggregated

[20] The variational standard errors from the other approaches are calculated as follows. ADVI returns a variational posterior on all variables that is sampled and used to assess nominal coverage. GVA calculates approximate standard errors by inverting Fisher's Information for the main parameters and the variational parameters and then takes the sub-matrix corresponding to the coefficients (Ormerod and Wand 2012).

[21] See Appendix D for dis-aggregated results.

across all models. First, in terms of run time, the variational methods are markedly faster than existing implementations—taking minutes when the original models take hours.

The performance of the variational method for recovering the correct point estimates is also good: I evaluate this by standardizing all variables to have unit variance. This allows (non-intercept) variables to be interpreted in terms of standard deviations. I then calculate the mean absolute error (i.e. the average absolute discrepancy) between the variational method and the "gold standard". I also report the correlation between the (standardized) point estimates. It shows that (i) the estimates are nearly perfectly correlated and (ii) the mean absolute error is quite small—around one or two percent of a standard deviation. Appendix D shows the relationship visually.

Table 6: Performance of Methods on Actual Datasets

Paper	Method	N. Obs	N. Var	Run Time (Min)		MAE	Corr.
				Original	VI-EM		
Stoeckel (2013)	Multi	20782	81	130.51	3.23	0.015	0.989
Legewie and Schaeffer (2016)	NegBin	29631	17.6	113.80	1.94	0.025	0.994
Micozzi (2013)	NegBin	1429	14.5	5.21	0.17	0.002	0.999
Schumacher et al. (2015)	NegBin	1686	5	0.65	0.17	0.008	0.999
Giger and Klüver (2016)	Logit	20260	11	0.73	0.27	0.0002	0.999
Michelitch and Utych (2018)	Logit	465196	7	9.14	3.19	0.014	0.999

Note: Appendix D provides details on the models analyzed in each replication. "N. Obs" is the number of observations; "N. Var" is the number of variables; in the case of running multiple models for a single paper, "N.Var" is the average number of variables per model. All other results are reported pooling across multiple analyses. "Original" represents the run-time by the "gold standard" method for finding the maximum likelihood estimate. "VI-PG" is the run-time for the variational algorithm derived in this paper. "MAE" is the mean absolute error (discrepancy) between the variational and gold standard methods, where all variables are standardized to have a unit variance. "Corr" is the correlation between the standardized estimates

I also examined the properties of the approximate standard errors using the method described above. Sometimes the associated variance is invalidly estimated (e.g. negative) and thus no valid approximate standard error can be calculated. Table 7 shows the results. It reports, as before, the average absolute distance between the standard errors (standardized), focusing only on the cases where they can be estimated. It shows promising results; the discrepancy is rather small. Further, I examined the implications for hypothesis testing: I conducted a 95% two-tailed test for statistical significance for each coefficient. I then compared whether the results of such a test would differ between the exact and variational methods. The results are again reasonably good;

in most cases, the two tests coincide (e.g. both reject or both fail to reject) and, when they disagree, the approximate standard errors seem to be neither systematically too conservative or anti-conservative. Exploring why the standard errors cannot sometimes be calculated and whether this can be improved is an area of future research. The results here, however, suggest that, when they can be calculated, the approximate standard errors give a reasonable approximation to the correct standard errors.

Table 7: Performance of Approximate Standard Errors on Actual Datasets

Paper	Method	N. Obs	N.Var	MAE	% Invalid	% Same	% Con.	%Anti
Legewie and Schaeffer (2016)	NegBin	29631	17.6	0.01	7.8	92.3	7.6	0
Micozzi (2013)	NegBin	1429	14.5	0.02	34.5	76.3	17.4	2
Schumacher et al. (2015)	NegBin	1686	5	0.003	0	100	0	0
Giger and Klüver (2016)	Logit	20260	11	0.004	9.1	100	0	0
Michelitch and Utych (2018)	Logit	465196	7	0.004	28.6	80	0	14.3

Note: This figure compares the standard errors from the exact "gold standard" method against the approximate variational standard errors. "MAE" reports the mean absolute error (discrepancy) between the standard errors (standarized such that all variables have a unit variance). This is calculated only on valid standard errors. "% Invalid" reports the percent of standard errors that are invalid. "% Same" reports the percent of (valid) standard errors that have the same result from the exact and variational method on a two-sided 95% t-test. "% Con" reports the percentage of (valid) standard errors that are more conservative (i.e. fail to reject for variational when exact rejects) and "% Anti" reporst the percentage that are anti-conservative (i.e. varational rejects when the exact fails to reject).

From here, I examine two articles in more depth and show the ability to use the variational EM algorithm to efficiently extend existing results.

4.1 Attitudes towards the European Union

Stoeckel (2013) examines the support for the European Union (EU) amongst citizens across Europe. Building on existing research on public opinion (e.g. Zaller 1992), he divides citizen attitudes towards the EU as ones that are unambiguously positive or negative, ambivalent (i.e. holding both positive and negative attitudes), and indifferent (lacking both positive and negative attitudes). After including a variety of demographic controls, the article seeks to examine the role of cognitive and affective factors in predicting attitudes. On the cognitive side, he examines the effects of knowledge about the EU and media consumption. He hypothesizes and his results show that both of these factors increase ambivalence towards the EU by making citizens aware of both the positive

and negative things the EU does, while decreasing the prevalence of indifferent citizens.[22] On the affective side, he focuses on how attitudes towards the EU (e.g. trust and attachment in the institutions) affect ambivalence. He suggests that more positive affective ties work by decreasing *both* indifference and ambivalence and increasing positive views. To model these relationships, he uses a multinomial logistic regression. To address possible country-level confounding and to address correlation between respondents inside of a country, he includes random effects at the country level.

Given both the multinomial outcome and the reliance on random effects, the author turns to a common method in STATA (`gllamm`; Rabe-Hesketh, Skrondal, and Pickles 2004). As shown in Table 6, the variational algorithm is much faster—running in minutes versus two hours needed by the original software.[23] The vastly increased speed of the Polya-Gamma approach means that model exploration is now possible in ways that were prohibitively costly under existing implementations. I focus on one particular extension: As noted above, four variables of interest are knowledge of the European union and news media consumption (cognitive variables) and trust and attachment to EU institutions (affective variables). It is interesting to wonder whether the results that decreasing these variables corresponds to a drop in the probability of being indifferent on the European Union varies by country.A plausible way to model this is to include these variables as random slopes by country (Stegmueller 2013). This can be done easily using the Polya-Gamma framework above. This extension takes a very long time to run on `gllamm` whereas it does not markedly increase the speed for the Polya-Gamma estimation. If I include all four random slopes by country, the average time for the EM implementation increases by only about a minute (four minutes); by contrast, the STATA implementation takes nearly forty times as long (around seventy-five hours!). This illustrates how simple, theoretically motivated, extensions may become prohibitively expensive when using existing quadrature-based approaches.

To examine which sets of random effects improve the model's predictive power, I employ two strategies—one that explicitly incorporates the fast variational method. First, I rely on ten-fold cross-validation (Hastie, Tibshirani, and Friedman 2009); stratifying within country, I hold out one-tenth of the respondents and predict their responses using a model fit on the remaining 90% of the data. I average the log-likelihood for the predictions across folds to get a measure of the

[22] One additional factor, elite division, is not considered here because of its lack of variation within country.

[23] A full Bayesian Gibbs Sampler (i.e. a similar "exact" strategy) using Polya-Gamma augmentation takes around one hour to run to convergence. This is longer than the EM algorithm but still faster than author's original approach.

predictive power of each model. This requires running ten regressions per specification; with the EM algorithm, this takes about thirty minutes versus nearly a day for the traditional implementations. As I wish compare *sixteen* sets of random effects, this is prohibitively expensive to use existing software.

Second, to examine performance using an "exact" specification, I use the fully Bayesian estimation procedure (see Appendix B) and compare an information criterion that approximates cross-validation: the Widely Applicable Information Criterion (WAIC).[24] It can be interpreted like the AIC or BIC where smaller indicates better model fit; it looks at how well the model fits the data while penalizing more complex models (Gelman, Hwang, and Vehtari 2014; Vehtari, Gelman, and Gabry 2017).[25] Table 8 reports the results of the sixteen specifications.

When examining the cross-validated results, it is worth noting that most models are rather similar as *no* model does especially well at prediction. The WAIC shows more noticeable differences; it is worth noting, however, that the two approaches select the same top two models, and agree on the composition of the top-five models. Thus, this suggests that using the EM algorithm as a fast preliminary test, in this case, helps to narrow down the relevant universe of models to use for an exact procedure.

Substantively, it is interesting that neither the original specification (Model 1; no random slopes) or the most complex model (Model 16; four random slopes) is preferred. While the original

[24]Convergence is discussed in Appendix E. I ran three chains for each model with 2,500 burnin and 2,500 retained samples. Starting values were the EM point estimates plus random noise for the second and third chains. After doing this, the Gelman-Rubin statistic for all fixed and random effects is below 1.1 in all cases across almost all models. In two Models (9 and 14), only 98% of coefficients are below this value. Using a stricter test of a Gelman-Rubin statistic below 1.01, 66% of fixed effects and 85% of random effects pass this standard.

[25]Formally, the WAIC is build as follows, see Gelman, Hwang, and Vehtari (2014). We first calculate the expected log-pointwise predictive density (*elpd*) as a combination of two terms. First, we calculate the log-pointwise predictive density, *lpd*, using the posterior. However, this is known to be an upwardly biased quantity so we penalize it p_{WAIC} by the variance of the posterior.

The formulae for the estimators of these quantities are below, noting that N is the number of observations, S is the number of posterior draws and $\boldsymbol{\theta}^{(s)}$ as the vector of parameters associated with draw s. $Var(\{a^s\}_{s=1}^{S})$ indicates the sample variance of a set of observations $\boldsymbol{a}$ indexed by s.

$$\hat{lpd} = \sum_{i=1}^{N} \ln \left(\frac{1}{S} \sum_{s=1}^{S} p(y_i|\boldsymbol{\theta}^{(s)}) \right)$$

$$\hat{p}_{WAIC} = \sum_{i=1}^{N} Var \left(\{\ln p(y_i|\boldsymbol{\theta}^{(s)})\} \right)$$

$$\hat{elpd} = \hat{lpd} - \hat{p}_{WAIC}$$

Finally, to put the *elpd* on the same scale as the AIC, i.e. negative is better, we multiply it by negative two, i.e. $\hat{WAIC} = -2\hat{elpd}$.

Table 8: Adding Random Slopes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
CV (LL)	-0.9928	-0.9831	-0.9895	-0.9822	-0.9939	-0.9845	-0.9905	-0.9830
WAIC	41280.81	40880.83	41133.30	40842.59	41332.51	40951.82	41180.37	40872.75
Rank - CV	15	6	11	2	16	8	12	5
Rank - WAIC	14	6	11	2	16	8	12	4
RE: Trust		✓		✓		✓		✓
RE: Attachment			✓	✓			✓	✓
RE: Media					✓	✓	✓	✓
RE: Knowledge								
	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
CV (LL)	-0.9916	-0.9828	-0.9888	-0.9817	-0.9927	-0.9837	-0.9894	-0.9829
WAIC	41235.20	40875.46	41104.36	40811.36	41282.92	40903.25	41127.04	40850.14
Rank - CV	13	3	9	1	14	7	10	4
Rank - WAIC	13	5	9	1	15	7	10	3
RE: Trust		✓		✓		✓		✓
RE: Attachment			✓	✓			✓	✓
RE: Media					✓	✓	✓	✓
RE: Knowledge	✓	✓	✓	✓	✓	✓	✓	✓

Note: All models include a random intercept for country; the bottom four lines indicate whether there are random slopes for the specified variable by country. 'CV (LL)' indicates the average held-out log-likelihood from cross-validation, averaged over the ten folds. Larger values are better. 'WAIC' indicates the widely applicable information criterion, discussed in the main text. Smaller values are better. 'Rank' ranks the models where '1' is the best performing model.

specification is beaten handedly by most other ones, either selection procedure chooses only two characteristics to give country-specific variation (Model 4).[26] This also shines light on the author's original theory as including random slopes for the "affective" variables (trust and attachment) is warranted while the effect of "cognitive" factors (media and knowledge) are more likely stable across member states.

4.2 Ethnic Heterogeneity and Public Disorder

Focusing on an example in American politics, Legewie and Schaeffer (2016) examine the number of 311 calls about public disorder (e.g. noise complaints) in New York City. They argue for a reconceptualization of the drivers of these calls; rather than focusing on ethnic heterogeneity *per se*, the key explanatory variable is "contested boundaries". "Poorly defined boundaries that separate ethnic and racial groups" (Legewie and Schaeffer 2016, p. 126) should see higher conflict than either homogeneous areas *or*, more novel to their argument, areas with "sharp" boundaries

[26] A rough test for statistically detectable difference in the WAIC Vehtari, Gelman, and Gabry (2017) suggests that the best fitting model (Model 12) has a p-value of below 0.05 versus all other models except one (Model 4) where the p-value is 0.06.

between groups. They measure this by a variable referred to as "edge intensity" that captures how quickly the racial composition changes between census blocks. Their theory predicts, and importantly relies upon, the identification of an inverse "u-shaped" relationship; increasing edge intensity should, holding all other variables constant, increase calls to 311 up to an inflection point after which it should *decrease* 311 calls—as very high edge intensity indicates "sharp" demarcation between groups and thus, according to their theory, more limited conflict than in "contested" regions. Their main empirical analysis estimates negative binomial regressions with a random effect for the approximately 2,000 census tracts in their data. As shown above, the models fit using `lme4` (Bates et al. 2015) take a long time to run from 37 minutes to 174 for the most complex model. Replicating eight models found in the original paper, no single model takes longer than four minutes using the variational EM algorithm with a speed-up ratio ranging from 13 to 126 times as fast!

As before, this variational method allows extensions to further test the theoretical arguments. The key explanatory variable for Legewie and Schaeffer (2016) is "edge intensity" having a quadratic functional form. Their functional form, however, imposes the assumption that the effect of this variable is constant across all precincts. We can relax this by including random slopes for edge intensity and its square to allow each tract to have varying slopes. This model runs nearly twenty times as quickly using the variational method versus the existing exact approach.[27]

Adding in random slopes importantly qualifies their existing results as there is a wide degree of heterogeneity across tracts with many having effectively monotonic slopes (i.e. increasing edge intensity leads for almost all observed values to an increase in expected public disorder). I summarize the results by using the following statistic: I calculate the "point of inflection" (i.e. the point after which an increase in edge intensity leads to a *decrease* in expected disorder). If that point of inflection is very high, then the relationship is effectively monotonic over the range of plausible values in New York City. Note that the original analysis reports an inflection point of 0.36.[28] The results are summarized below:

[27] It takes around forty minutes to estimate the model using the variational approach, but 13.5 hours for the exact approach!

[28] Although this is not emphasized in the original paper, only 0.8% of blocks have a value higher than this. Thus, the decline they focus on as supporting their theory occurs between the 99.2 and 99.5th percentiles of the data. This suggests, however, that, at least for New York City, the vast majority of blocks exist see a monotonically increasing effect.

1. 13.5% of tracts have an inflection point *larger* than the one in the original paper. For almost all plausible edge intensities, there is a simple monotonic positive relationship in these tracts. Only extreme outliers in terms of edge intensity see the posited decrease.

2. Only 23.9% of tracts have an inflection point within the 5th and 95th percentiles of the observed edge intensities. Only 54.1% are within the 2.5th to 97.5th percentiles.

This qualifies the author's original results by suggesting that a "u-shaped" relationship only exists in a minority of census tracts. For most tracts, the relationship exists in a more straightforward way: Increasing edge intensity (contested boundaries) increases calls to 311. This bolsters the author's conjecture (Legewie and Schaeffer 2016, p. 141) to examine other cities with larger maximum edge intensities as this would provide clearer evidence of whether a non-monotonic relationship exists in those cities (e.g. Chicago).

5 Conclusions

In this paper, I have shown that Polya-Gamma data augmentation developed by Polson, Scott, and Windle (2013) is a powerful tool for political scientists. Relying on a data augmentation scheme similar to that for probit regression, it has become a powerful tool for estimating complex non-linear models. I applied those results to derive a novel set of algorithms for estimating non-linear hierarchical models. The algorithm runs quickly and has good performance as established on simulated and actual datasets. On simulated data, it closely recovers the point estimates of slower "gold standard" approaches, outperforming other variational methods based on different approximations. The approximate standard errors associated with this method have reasonable coverage.

I then examined the performance on a selection of recent papers. In general, variational EM always leads to gains in speed at limited costs in terms of accuracy. In the cases analyzed, the variational point estimates are, on average, no greater than 0.02 of a standardized effect from the exact methods and are often much closer. In all cases, they are highly correlated with the true point estimates. In terms of when the most improvement is expected, the analysis suggests that two conditions likely lead to very slow estimation of traditional methods: (i) multiple random effects or

(ii) uses a negative binomial or multinomial outcome. As datasets grow larger and more complex, these problems will become more severe and thus require fast, albeit approximate, methods for inference.

Substantively, I used two illustrations to show that this fast variational method allows researchers to better explore, in reasonable time, substantively important questions. When re-analyzing the results from Stoeckel (2013), including four random slopes leads the existing algorithm in `STATA` to take a prohibitively long length of time (over three days!). By contrast, the approximate method runs in the minutes and shows important country-level heterogeneity in the relationship between affective variables (trust and attachment) and EU attitudes. Re-examining a recent paper in sociology (Legewie and Schaeffer 2016) also provides insights into the published results. Their original analysis suggested a novel conclusion: "contested boundaries" (neighborhoods that correspond to fuzzy boundaries between ethnically homogeneous areas) have the highest number of calls to 311 about public disorder, but areas that are either very homogeneous or show very sharp boundaries between groups should see lower rates of these calls. This implied a u-shaped functional form in the estimated effect of the key variable of "edge intensity" measuring how sharply the ethnic composition of a neighborhood is varying. I showed that if we allow for tract-specific relationships via random effects, it seems clear that this u-shaped relationship only exists in a minority of tracts. Exploring this heterogeneity would have been prohibitively expensive using existing tools (around 17.5 hours to estimate a single model), but runs in under an hour using the variational approach.

Beyond the models outlined in this paper however, there will likely be other areas where Polya-Gamma augmentation will be useful for the social sciences. As it is conceptually similar to existing data augmentation schemes, it can be integrated into existing complex models relatively straightforwardly. Exploring those avenues is an exciting area for the methodological community to leverage recent statistical results to improve the ability for applied researchers to estimate complex models efficiently.

References

Albert, James H., and Siddhartha Chib. 1993. "Bayesian Analysis of Binary and Polychotomous Response Data". *Journal of the American Statistical Association* 88 (422): 669–679.

Bates, Douglas, et al. 2015. "Fitting Linear Mixed-Effects Models Using lme4". *Journal of Statistical Software* 67 (1): 1–48.

Bell, Andrew, and Kelvyn Jones. 2015. "Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data". *Political Science Research and Methods* 3 (1): 133–153.

Betancourt, Brenda, Abel Rodríguez, and Naomi Boyd. 2017. "Bayesian Fused Lasso Regression for Dynamic Binary Networks". *Journal of Computational and Graphical Statistics* 26 (4): 840–850.

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. 2017. "Variational inference: A review for statisticians". *Journal of the American Statistical Association* 112 (518): 859–877.

Blitzstein, Joseph K., and Jessica Hwang. 2014. *Introduction to Probability*. Chapman / Hall/CRC.

Brookes, Mike. 2011. *The Matrix Reference Manual*. `http://www.ee.imperial.ac.uk/hp/staff/dmb/matrix/intro.html`.

Bürkner, Paul-Christian. 2017. "brms: An R Package for Bayesian Multilevel Models Using Stan". *Journal of Statistical Software* 80 (1): 1–28.

Carpenter, Bob, et al. 2017. "Stan: A probabilistic programming language". *Journal of Statistical Software* 76 (1): 1–32.

Chen, Jianfei, et al. 2013. "Scalable Inference for Logistic-Normal Topic Models". In *Neural Information Processing Systems 2013*. `https://papers.nips.cc/paper/4981-scalable-inference-for-logistic-normal-topic-models.pdf`.

Clark, Tom S., and Drew A. Linzer. 2015. "Should I Use Fixed or Random Effects?" *Political Science Research and Methods* 3 (2): 399–408.

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society. Series B (Methodological)* 39 (1): 1–38.

Fenton, Lawrence. 1960. "The sum of log-normal probability distributions in scatter transmission systems". *IRE Transactions on Communications Systems* 8 (1): 57–67.

Gelman, Andrew. 2006. "Prior distributions for variance parameters in hierarchical models". *Bayesian Analysis* 1 (3): 515–533.

Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman, Andrew, Jessica Hwang, and Aki Vehtari. 2014. "Understanding predictive information criteria for Bayesian models". *Statistics and Computing* 24 (6): 997–1016.

Ghitza, Yair, and Andrew Gelman. 2013. "Deep Interactions with MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups". *American Journal of Political Science* 57 (3): 762–776.

Giger, Nathalie, and Heike Klüver. 2016. "Voting Against Your Constituents? How Lobbying Affects Representation". *American Journal of Political Science* 60 (1): 190–205.

Glynn, Chris, et al. 2018. "Bayesian Analysis of Dynamic Linear Topic Models". *Bayesian Analysis* Advanced Publication. doi:10.1214/18-BA1100.

Goplerud, Max. 2018. "A Multinomial Framework for Ideal Point Estimation". *Political Analysis* Advanced Access. https://doi.org/10.1017/pan.2018.31.

Goplerud, Max, et al. 2018. "Sparse Multilevel Regression (and Poststratification [sMRP])". *Working Paper*. https://scholar.harvard.edu/files/dtingley/files/sparsemultilevel.pdf.

Grimmer, Justin. 2011. "An Introduction to Bayesian Inference via Variational Approximations". *Political Analysis* 19 (1): 32–47.

Hall, Peter, John T. Ormerod, and Matt P. Wand. 2011. "Theory of Gaussian variational approximation for a Poisson mixed model". *Statistica Sinica* 21 (1): 369–389.

Hall, Peter, et al. 2011. "Asymptotic normality and valid inference for Gaussian variational approximation". *The Annals of Statistics* 39 (5): 2502–2532.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. 2nd. Springer-Verlang.

Imai, Kosuke, James Lo, and Jonathan Olmsted. 2016. "Fast Estimation of Ideal Points with Massive Data". *American Political Science Review* 110 (4): 631–656.

Jaakkola, Tommi S., and Michael I. Jordan. 1997. "A variational approach to Bayesian logistic regression models and their extensions". In *Sixth International Workshop on Artificial Intelligence and Statistics*.

Johndrow, James E., et al. 2018. "MCMC for Imbalanced Categorical Data". *Journal of the American Statistical Association* Advance Access:1–10. doi:10.1080/01621459.2018.1505626.

Kucukelbir, Alp, et al. 2017. "Automatic Differentiation Variational Inference". *Journal of Machine Learning Research* 18 (14): 1–45.

Laird, Nan M., and James H. Ware. 1982. "Random-Effects Models for Longitudinal Data". *Biometrics* 38 (4): 963–974.

Lax, Jeffrey R., and Justin H. Phillips. 2009. "How Should We Estimate Public Opinion in The States?" *American Journal of Political Science* 53 (1): 107–121.

Legewie, Joscha, and Merlin Schaeffer. 2016. "Contested Boundaries: Explaining Where Ethnoracial Diversity Provokes Neighborhood Conflict". *American Journal of Sociology* 122 (1): 125–161.

Linderman, Scott W., Matthew J. Johnson, and Ryan P. Adams. 2015. "Dependant Multinomial Models Made Easy". In *Neural Information Processing Systems 2015*. https://hips.seas.harvard.edu/files/linderman-dependent-nips-2015.pdf.

Liu, Chuanhai, Donald B. Rubin, and Ying Nian Wu. 1998. "Parameter expansion to accelerate EM: the PX-EM algorithm". *Biometrika* 85 (4): 755–770.

Louis, Thomas A. 1982. "Finding the Observed Information Matrix when Using the EM Algorithm". *Journal of the Royal Statistical Society. Series B (Methodological)* 44 (2): 226–233.

Makalic, Enes, and Daniel F. Schmidt. 2015. "A Simple Sampler for the Horseshoe Estimator". *IEEE Signal Processing Letters* 23 (1): 179–182.

Mardia, Kanti V., and Roger J. Marshall. 1984. "Maximum Likelihood Estimation of Models for Residual Covariance in Spatial Regression". *Biometrika* 71 (1): 135–146.

McLachlan, Geoffrey, and Thriyambakam Krishnan. 2008. *The EM Algorithm and Extensions*. 2nd. Wiley.

Meng, Xiao-Li, and Donald B. Rubin. 1993. "Maximum likelihood estimation via the ECM algorithm: A general framework". *Biometrika* 80 (2): 267–278.

Meng, Xiao-Li, and David Van Dyk. 1998. "Fast EM-type implementations for mixed effects models". *Journal of the Royal Statistical Society. Series B (Methodological)* 60 (3): 559–578.

Meng, Xiao-Li, and David Van Dyk. 1997. "The EM Algorithm–an Old Folk-song Sung to a Fast New Tune". *Journal of the Royal Statistical Society. Series B (Methodological)* 59 (3): 511–567.

Michelitch, Kristin, and Stephen Utych. 2018. "Electoral Cycle Fluctuations in Partisanship: Global Evidence from Eighty-Six Countries". *The Journal of Politics* 80 (2): 412–427.

Micozzi, Juan Pablo. 2013. "Does Electoral Accountability Make a Difference? Direct Elections, Career Ambition, and Legislative Performance in the Argentine Senate". *The Journal of Politics* 75 (1): 137–149.

Neal, Radford M. 2003. "Slice Sampling". *The Annals of Statistics* 31 (3): 705–767.

Neal, Radford M., and Geoffrey E. Hinton. 1998. "A view of the EM algorithm that justifies incremental, sparse, and other variants." In *Learning in Graphical Models*, ed. by Michael I. Jordan, 355–368. Springer.

Oakes, David. 1999. "Direct calculation of the information matrix via the EM". *Journal of the Royal Statistical Society. Series B (Methodological)* 61 (2): 479–482.

Ormerod, John T., and Michael P. Wand. 2012. "Gaussian Variational Approximate Inference for Generalized Linear Mixed Models". *Journal of Computational and Graphical Statistics* 21 (1): 2–17.

Park, David K., Andrew Gelman, and Joseph Bafumi. 2004. "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls". *Political Analysis* 12 (4): 375–385.

Petersen, Kaare Brandt, and Michael Syskind Pedersen. 2012. *The Matrix Cookbook*. http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=3274.

Pillow, Jonathan W., and James G. Scott. 2012. "Fully Bayesian inference for neural models with negative-binomial spiking". In *Neural Information Processing Systems 2012*.

Polson, Nicholas G., James G. Scott, and Jesse Windle. 2013. "Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables". *Journal of the American Statistical Association* 108 (504): 1339–1349.

Rabe-Hesketh, Sophia, and Anders Skrondal. 2008. *Multilevel and longitudinal modeling using STATA*. STATA Press.

Rabe-Hesketh, Sophia, Anders Skrondal, and Andrew Pickles. 2004. "Generalized Multilevel Structural Equation Modeling". *Psychometrika* 69 (2): 167–190.

Schumacher, Gijs, et al. 2015. "How Aspiration to Office Conditions the Impact of Government Participation on Party Platform Change". *American Journal of Political Science* 59 (4): 1040–1054.

Steenbergen, Marco R., and Bradford S. Jones. 2002. "Modeling Multilevel Data Structures". *American Journal of Political Science* 46 (1): 218–237.

Stegmueller, Daniel. 2013. "How Many Countries for Multilevel Modeling? A Comparison of Frequentist and Bayesian Approaches". *American Journal of Political Science* 57 (3): 748–761.

Stewart, Brandon. 2014. "Latent Factor Regressions for Social Sciences". *Working aper*. https://scholar.princeton.edu/sites/default/files/bstewart/files/tensorreg.pdf.

Stoeckel, Florian. 2013. "Ambivalent or Indifferent? Reconsidering the Structure of EU Public Opinion". *European Union Politics* 14 (1): 23–45.

Teh, Yee W., David Newman, and Max Welling. 2007. "A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation". In *Advances in Neural Information Processing Systems 2007*, 1353–1360.

Van Dyk, David A., and Taeyoung Park. 2008. "Partially collapsed Gibbs samplers: Theory and methods". *Journal of the American Statistical Association* 103 (482): 790–796.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2017. "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC". *Statistics and Computing* 27 (5): 1413–1432.

Wang, Yixin, and David M. Blei. 2018. *Journal of the American Statistical Association*: Advanced Access.

Wei, Greg C. G., and Martin A. Tanner. 1990. "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms". *Journal of the American Statistical Association* 85 (411): 699–704.

Zaller, John. 1992. *The Nature and Origins of Mass Opinion*. Cambridge University Press.

Zhou, Mingyuan, et al. 2012. "Lognormal and Gamma Mixed Negative Binomial Regression". In *International Conference on Machine Learning 2012*, 1343–1350.

epend

A Details on JSTOR Analysis

I searched the full text of JSTOR articles in the following political science journals, using the full coverage of JSTOR at the time of writing. The abbreviations used in Figure 1 are shown in parenthesis.

> American Journal of Political Science (AJPS), British Journal of Political Science (BJPS), Comparative Politics (CP), International Organization (IO), Journal of Peace Research (JPR), Journal of the American Statistical Association (JASA), Legislative Studies Quarterly (LSQ), Perspectives on Politics (Perspectives), Political Analysis (PA), Political Methodology (PA), Political Psychology (PolPsych), PS: Political Science and Politics (PS), Public Choice (Public Choice), The American Economic Review (AER), The American Political Science Review (APSR), The Journal of Conflict Resolution (JCR), The Journal of Politics (JOP), The Public Opinion Quarterly (POQ), World Politics (WP)

I counted the number of articles with any hit to the following regular expressions

1. "Skrondal and Rabe-Hesketh—Rabe-Hesketh and Skrondal" (for Rabe-Hesketh and Skrondal 2008; Rabe-Hesketh, Skrondal, and Pickles 2004)
2. "Gelman and Hill" (for Gelman and Hill 2006)
3. "Steenbergen and Jones" (for Steenbergen and Jones 2002)
4. "(hierarchical|multilevel) (model|logit|probit|poisson|negative binomial|multinomial|regression)"
5. "random (effect|slope|intercept)s? (model|logit|probit|poisson|negative binomial|multinomial|regression)?"

B Derivations of Models

This section derives the full algorithms for all models used in the main paper: For binomial (logistic), negative binomial, and multinomial regression. While the main analysis in the paper focuses on the fast variational EM algorithm, I first start by deriving the fully Bayesian Gibbs Sampler as, with that in hand, the variational algorithm can be read off relatively straightforwardly. The associated software can estimate models using either procedure.

B.1 Random Effects Notation

In this Appendix, I rely on a slightly different notation than in the main text to allow derivation of general results (i.e. multiple random effects and/or random slopes). Specifically, I consider a linear predictor of the following form:

$$\psi_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{z}_i^T \boldsymbol{\alpha}; \quad \boldsymbol{\alpha} \sim N(0, \boldsymbol{T}) \tag{16}$$

$\boldsymbol{\beta}$ denotes the coefficients on the fixed effects, i.e. no pooling across groups via a normal prior, whereas $\boldsymbol{\alpha}$ corresponds to the random effects. As noted in the main text, this differs from the common notation in Gelman and Hill (2006), e.g. $\alpha_{j[i]}$ to denote the random effect j associated with observation i.

Thus, I always use $\boldsymbol{\alpha}$ to denote all random effects stacked together into a single vector. In the case of a single random effect with J levels, it has a dimensionality of $J \times 1$. It is possible to define a (sparse) $\boldsymbol{z}_i$ such that the multiplication of $\boldsymbol{z}_i^T\boldsymbol{\alpha}$ pulls out the "correct" linear combination of random effects for observation i. In the case of a single random effect, $\boldsymbol{z}_i$ is a "one-hot" vector (i.e. with one element equal to "1" and the rest equal to "0").

The variance matrix for $\boldsymbol{\alpha}$ is defined as $\boldsymbol{T}$ and is block-diagonal (or banded block-diagonal) and *not* unstructured. In the single random effect case, $\boldsymbol{T} = \sigma_\alpha^2\boldsymbol{I}$ where σ_α^2 is the variance of the random effect, e.g. $\alpha_j \sim N(0, \tau_\alpha^2)$ and $\boldsymbol{I}$ has dimensionality $J \times J$. Note by standard rules of multivariate normals, $\boldsymbol{z}_i^T\boldsymbol{\alpha} \sim N(0, \sigma_\alpha^2)$.

This notation allows for a number of elegant extensions; for example, assume we had a random slope on some variable z_1. Again noting the random effect has J levels, $\boldsymbol{z}_i$ would be of length $2 \cdot J$. If the observation is in the first group, $\boldsymbol{z}_i^T = [1, z_{i,1}, 0, 0, \cdots, 0]$. Thus, it has two-nonzero elements; one to pull out the random intercept for that group and the other to pull out the random slope. In this case, $\boldsymbol{T}$ has dimensionality $2J \cdot 2J$ where it is block diagonal with 2-by-2 blocks corresponding to the variance-covariance matrix of the random slope and random intercept: $\boldsymbol{\Sigma}_{j,\alpha}$. When evaluating, $\boldsymbol{z}_i^T\boldsymbol{\alpha}$, this pulls out the following quantity: $\alpha^0_{j[i]} + \alpha^1_{j[i]}z_{i,1}$ — exactly the quantity corresponding to a random effect in the traditional notation. This could be further re-arranged to as $\check{\boldsymbol{z}}_i^T\boldsymbol{\alpha}_{j[i]}$ where $\check{\boldsymbol{z}}_i$ pulls out the non-zero elements—again mirroring the classic notation. In the case of multiple random effects (e.g. a random slope and intercept), $\boldsymbol{T}$ would be block diagonal with the variance-covariance matrix $\boldsymbol{\Sigma}_{j,\alpha}$ stacked together J times.

It is thus worth stressing that the prior (and inference) is *not* over $\boldsymbol{T}$ but rather over $\boldsymbol{\Sigma}_{j,\alpha}$. For single random effects, we put a single inverse Gamma (inverse Wishart, etc.) prior on $\boldsymbol{\Sigma}_{j,\alpha}$— regardless of J and the dimensionality of $\boldsymbol{T}$. The benefit of this procedure is that it allows us to incorporate multiple random effects that are cross-classified without changing our notation or sampling scheme. This also means that the approximate EM algorithm derived shortly will also generalize to cross-classified random effects. While $\boldsymbol{Z}$ may be high dimensional, it is very sparse (as is $\boldsymbol{T}$) and thus computational difficulties are mitigated.

B.2 Logistic Binomial

For a binomial regression, we define the data generating process as follows noting that the standard logistic case occurs when $N_i = 1$ for all observations.

$$Y_i \sim Binom(N_i, p_i); \quad p_i = \frac{\exp(\psi_i)}{1 + \exp(\psi_i)}; \quad \psi_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{\alpha}; \quad \boldsymbol{\alpha} \sim N(0, \boldsymbol{T}) \tag{17}$$

This is a simple logistic regression with random effects denoted by $\boldsymbol{\alpha}$. I assume we have conditionally conjugate priors on $\boldsymbol{\beta}$ (normal) and the elements $\boldsymbol{\Sigma}_\alpha$ of $\boldsymbol{T}$ (inverse Wishart). To do the Gibbs Sampler, we augment with the $\boldsymbol{\alpha}$ (the random effects) and then the Polya-Gamma random variables to get the following posterior where $p_0(\cdot)$ denotes the prior:

$$p\left(\{\omega_i\}, \boldsymbol{\beta}, \boldsymbol{T}, \boldsymbol{\alpha} | \{\boldsymbol{x}_i, y_i, N_i\}\right) \propto p_0(\boldsymbol{\beta}, \boldsymbol{T})|\boldsymbol{T}|^{-1/2} \exp\left(-1/2\boldsymbol{\alpha}^T\boldsymbol{T}^{-1}\boldsymbol{\alpha}\right) \prod_j \exp(s_i\psi_i - \omega_i\psi_i^2/2) f(\omega_i|N_i) \tag{18}$$

The Gibbs Sampler is defined below for the logistic case, relying on the results in the main text about Polya-Gamma random variables. I set simple conditionally conjugate priors but others (e.g. Gelman 2006) could be used with little extra difficulty.

Gibbs Sampler for Hierarchical Logistic Regression Preliminaries

- Define $s_i = y_i - N_i/2$; S_j is the number of rows of $\mathbf{\Sigma}_{j,\alpha}$
- Set priors such as $\boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{\Lambda}_\beta^0)$, $\mathbf{\Sigma}_{j,\alpha} \sim IW(\mathbf{\Psi}_j, \nu_j)$
- Set $\boldsymbol{\beta}^{(0)}$, $(\mathbf{\Sigma}_\alpha)^{(0)}$, $\boldsymbol{\alpha}^{(0)}$. This implicitly defines $\boldsymbol{T}^0$.

For t in $1, \cdots, T$:

1. Sample $\omega_i^{(t)} \quad \forall i$
$$\omega_i^{(t)}|- \sim PG\left(N_i, [\boldsymbol{\beta}^{(t-1)}]^T \boldsymbol{x}_i + \boldsymbol{z}_i^T \boldsymbol{\alpha}^{(t-1)}\right)$$
2. Rescale the data:
$$\tilde{s}_i^{(t)} = s_i/\sqrt{\omega_i^{(t)}}, \quad \tilde{\boldsymbol{x}}_i^{(t)} = \boldsymbol{x}_i \cdot \sqrt{\omega_i^{(t)}}, \quad \tilde{\boldsymbol{z}}_i^{(t)} = \boldsymbol{z}_i \cdot \sqrt{\omega_i^{(t)}}$$
3. Sample $\boldsymbol{\alpha}^{(t)} \sim N\left(\boldsymbol{\mu}_\alpha^{(t)}, \mathbf{\Lambda}_\alpha^{(t)}\right)$
$$\mathbf{\Lambda}_\alpha^{(t)} = \left(\left[\tilde{\boldsymbol{Z}}^{(t)}\right]^T \tilde{\boldsymbol{Z}}^{(t)} + \left[\boldsymbol{T}^{(t-1)}\right]^{-1}\right)^{-1}$$
$$\boldsymbol{\mu}_\alpha^{(t)} = \mathbf{\Lambda}_\alpha^{(t)} \left[\tilde{\boldsymbol{Z}}^{(t)}\right]^T \left(\tilde{\boldsymbol{s}}^{(t)} - \tilde{\boldsymbol{X}}^{(t)} \boldsymbol{\beta}^{(t-1)}\right)$$
4. Sample $\boldsymbol{\beta}^{(t)}$ $N\left(\boldsymbol{\mu}_\beta^{(t)}, \mathbf{\Lambda}_\beta^{(t)}\right)$
$$\mathbf{\Lambda}_\beta^{(t)} = \left(\left[\tilde{\boldsymbol{X}}^{(t)}\right]^T \left[\tilde{\boldsymbol{X}}^{(t)}\right] + \mathbf{\Lambda}_\beta^0\right)^{-1}$$
$$\boldsymbol{\mu}_{\boldsymbol{\beta}}^{(t)} = \mathbf{\Lambda}_\beta^{(t)} \left[\tilde{\boldsymbol{X}}^{(t)}\right]^T \left(\tilde{\boldsymbol{s}}^{(t)} - \tilde{\boldsymbol{Z}} \boldsymbol{\alpha}^{(t)}\right)$$
5. Update $\mathbf{\Sigma}_{\alpha,j}$, i.e. the variance for each random effect j. Re-arrange the posterior where G_j is the number of groups for random effect j:
$$p(\{\mathbf{\Sigma}_{\alpha,j}\}_{j=1}^J|-) \propto \prod_j \det\left(\mathbf{\Sigma}_{j,\alpha}\right)^{-G_j/2} \exp\left(-1/2 \sum_j \mathrm{tr}(\sum_{g=1}^{G_j} \boldsymbol{\alpha}_{j,g} \boldsymbol{\alpha}_{j,g}^T \mathbf{\Sigma}_{j,\alpha}^{-1}\right) \times$$
$$\prod_j |\mathbf{\Sigma}_{j,\alpha}|^{-[\nu_j + S_j + 1]/2} \exp\left(-1/2 \mathrm{tr}(\mathbf{\Psi}_j \mathbf{\Sigma}_{\boldsymbol{j,\alpha}}^{-1})\right)$$

   Thus, given the conjugate prior, we can sample as follows:
$$\mathbf{\Sigma}_{\alpha,j} \sim IW\left(\mathbf{\Psi}_j + \sum_{g=1}^{G_j} \boldsymbol{\alpha}_{j,g} \boldsymbol{\alpha}_{j,g}^T, \quad G_j + \nu_j\right)$$

B.2.1 Variational EM

The EM algorithm follows immediately from the Gibbs Sampler. In short, we plug the mean of $\omega_{ij}^{(t)}$ into the full conditionals (b)-(e) and then update the parameter to be its posterior *mode* rather than sampling directly. Thus, in the case without random effects, $\boldsymbol{\beta}^{(t)} = \boldsymbol{\mu}_{\beta}^{(t-1)}$ where again we have plugged in the conditional expectations of the Polya-Gammas. It is again worth stressing that this method is guaranteed to converge to the posterior mode (and the MLE in the case of a flat prior).

The random effects case requires more care. Notationally, I follow Neal and Hinton (1998) to cast the EM algorithm as a two-step optimization procedure (see also Blei, Kucukelbir, and McAuliffe 2017; Wang and Blei 2018). For some density $q(\{\omega_i\}_{i=1}^N, \boldsymbol{\alpha})$ and some parameter vector $\{\boldsymbol{\beta}, \{\boldsymbol{\Sigma}_{\alpha,j}\}\}$, we create the free-energy function F:

$$F(q(\{\omega_i\}_{i=1}^N, \boldsymbol{\alpha}), \{\boldsymbol{\beta}, \{\boldsymbol{\Sigma}_{\alpha,j}\}\}) = const. + E_q\left[\ln p(\boldsymbol{y}, \{\omega_i\}_{i=1}^N, \boldsymbol{\alpha} | \boldsymbol{X}, \boldsymbol{\beta}, \{\boldsymbol{\Sigma}_{\alpha,j}\})\right] - E_q\left[\ln q(\{\omega_i\}_{i=1}^N, \boldsymbol{\alpha})\right] \tag{19}$$

A key point in linking this to the variational literature is to note that we can re-arrange F to show the following:

$$F(q(\{\omega_i\}_{i=1}^N, \boldsymbol{\alpha}), \{\boldsymbol{\beta}, \{\boldsymbol{\Sigma}_{\alpha,j}\}\}) = const. + \ln p(\boldsymbol{y} | \boldsymbol{X}, \boldsymbol{\beta}, \{\boldsymbol{\Sigma}_{j,\alpha}\}) - KL\left(q || p(\{\omega_i\}_{i=1}^N, \boldsymbol{\alpha} | \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}, \{\boldsymbol{\Sigma}_{\alpha,j}\})\right) \tag{20}$$

The EM algorithm states that we can iteratively maximize F by first setting q equal to the full conditional of the missing data given the current parameter estimate and the data, i.e. $q = p(\{\omega_i\}_{i=1}^N, \boldsymbol{\alpha} | \boldsymbol{\beta}, \{\boldsymbol{\Sigma}_{j,\alpha}\}, \boldsymbol{y}, \boldsymbol{X})$ as this minimizes the KL divergence by setting it to zero. That is the E-Step. The M-step then maximizes the expectation of the log-conditional density with respect to $\boldsymbol{\beta}$ and $\{\boldsymbol{\Sigma}_{j,\alpha}\}$. This is the M-Step or the Conditional M-Step (Meng and Rubin 1993). Note further that the first term, $E_q\left[\ln p(\boldsymbol{y}, \{\omega_i\}_{i=1}^N, \boldsymbol{\alpha} | \boldsymbol{X}, \boldsymbol{\beta}, \{\boldsymbol{\Sigma}_{\alpha,j}\})\right]$ is known as the Q-function in the literature on the EM algorithm. Thus, the EM algorithm can be cast as a cyclical algorithm over the density q and the parameters.

As noted in the main text, the optimization with respect to q cannot be solved exactly as the E-Step is intractable. Thus, we rely on a variational approximation to F—Wang and Blei (2018) call this the "variational log-likelihood" and its optimum the "variational frequentist estimate".[29] This means that instead of trying to minimize the KL divergence over all distributions, we do this over a class of distributions that we assume q falls within.

I rely on the most common variational assumption—mean-field independence between $\boldsymbol{\Omega}$ and $\boldsymbol{\alpha}$. More simply, I maximize q over all distributions $\tilde{q}$ such that $\boldsymbol{\Omega}$ and $\boldsymbol{\alpha}$ are independent, i.e. $\tilde{q}(\boldsymbol{\Omega}, \boldsymbol{\alpha}) = \tilde{q}(\boldsymbol{\Omega})\tilde{q}(\boldsymbol{\alpha})$, and find the one that minimizes the KL divergence against the full conditional (that is intractable). As a variety of resources show (e.g. Grimmer 2011; Blei, Kucukelbir, and

[29] They note that this differs from the traditional 'evidence lower bound' (ELBO) in variational inference insofar as we are treating some of our parameters $(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\alpha,j})$ as an objective to maximize instead of placing a prior on it and finding its approximate posterior distribution.

McAuliffe 2017), the optimal variational distribution can be written as:

$$\tilde{q}(\boldsymbol{\Omega}) \propto \exp\left(E_{\tilde{q}(\boldsymbol{\alpha})}\left[\ln p(\boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\Omega}|\boldsymbol{\beta}, \{\boldsymbol{\Sigma}_{j,\alpha}\})\right]\right) = \exp\left(E_{\tilde{q}(\boldsymbol{\alpha})}\left[\ln p(\boldsymbol{\Omega}|\boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\boldsymbol{\Sigma}_{j,\alpha}\}) + \ln p(\boldsymbol{y}, \boldsymbol{\alpha}|\boldsymbol{\beta}, \{\boldsymbol{\Sigma}_{j,\alpha}\})\right]\right) \tag{21a}$$

$$\tilde{q}(\boldsymbol{\alpha}) \propto \exp\left(E_{\tilde{q}(\boldsymbol{\Omega})}\left[\ln p(\boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\Omega}|\boldsymbol{\beta}, \{\boldsymbol{\Sigma}_{j,\alpha}\})\right]\right) = \exp\left(E_{\tilde{q}(\boldsymbol{\Omega})}\left[\ln p(\boldsymbol{\alpha}|\boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\boldsymbol{\Sigma}_{j,\alpha}\}) + \ln p(\boldsymbol{y}, \boldsymbol{\Omega}|\boldsymbol{\beta}, \{\boldsymbol{\Sigma}_{j,\alpha}\})\right]\right) \tag{21b}$$

Note, importantly, this means that since both full conditionals of $\boldsymbol{\Omega}$ and $\boldsymbol{\alpha}$ are known, it means that the mean-field assumption implies that the variational distributions are in the same family as the full conditional. Thus, noting that the distribution for $\boldsymbol{\alpha}$ is normal and each ω_i is Polya-Gamma:

$$\tilde{q}(\boldsymbol{\Omega}, \boldsymbol{\alpha}) = \tilde{q}_{Normal}(\boldsymbol{\alpha}; \tilde{\boldsymbol{\mu}}_\alpha, \tilde{\boldsymbol{\Lambda}}_\alpha) \prod_{i=1}^{N} \tilde{q}_{PG}(\omega_i; \tilde{b}_i, \tilde{c}_i) \tag{22}$$

Thus, we have a set of variational parameters $\tilde{\boldsymbol{\mu}}_\alpha, \tilde{\boldsymbol{\Lambda}}_\alpha$ and $\{\tilde{b}_i, \tilde{c}_i\}_{i=1}^N$ that we optimize in the variational E-step and then, holding these parameters fixed, we optimize $\boldsymbol{\beta}$ and $\{\boldsymbol{\Sigma}_{j,\alpha}\}$ in the M-Step. Using this notation, we can write the variational objective for the logistic case, using the following helpful identity from Polson, Scott, and Windle (2013), where $f(\omega_i|b, c)$ denotes the Polya-Gamma density with parameters b and c as discussed in the main text:

$$f(\omega_i|b, c) = \exp(-c^2/2\omega_i) f(\omega_i|b, 0) \left[\cosh(c/2)\right]^b \tag{23}$$

This is important because $f(\omega_i|b, c)$ is itself intractable, and thus its entropy is also intractable; fortunately, to evaluate $\tilde{F}$ up to a constant that does not depend on any variational parameters, we can carefully cancel out terms. Considering the logistic case:

$$\begin{aligned}\tilde{F} = &-N\ln(2) + \sum_{i=1}^{N}(y_i - 1/2)[\boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\tilde{\mu}_\alpha] - E_{\tilde{q}(\omega_i|\tilde{b}_i,\tilde{c}_i)}[\omega_i]\left(\left[\boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\tilde{\mu}_\alpha\right]^2 + \boldsymbol{z}_i^T\tilde{\boldsymbol{\Lambda}}_\alpha\boldsymbol{z}_i\right)/2 + E_{\tilde{q}(\omega_i|\tilde{b}_i,\tilde{c}_i)}[\ln f(\omega_i|1,0)] \\ &1/2\ln(|2\pi\boldsymbol{T}|) - 1/2\mathrm{tr}\left(\boldsymbol{T}^{-1}\left[\tilde{\mu}_\alpha\tilde{\mu}_\alpha^T + \tilde{\boldsymbol{\Lambda}}_\alpha\right]\right) + 1/2\ln(|2\pi e\tilde{\boldsymbol{\Lambda}}_\alpha|) + \ln p_0(\boldsymbol{\beta}) + \ln p_0(\{\boldsymbol{\Sigma}_{j,\alpha}\}) \\ &\sum_{i=1}^{N}\tilde{c}_i^2/2E_{\tilde{q}(\omega_i|\tilde{b}_i,\tilde{c}_i)}[\omega_i] - E_{\tilde{q}(\omega_i|\tilde{b}_i,\tilde{c}_i)}[\ln f(\omega_i|1,0)] - \ln\cosh(\tilde{c}_i/2)\end{aligned} \tag{24}$$

Crucially, note that the intractable $E_{\tilde{q}}[\ln f(\omega_i|1, 0)]$ disappears as it appears in both the entropy term and the expectation of the complete data log-posterior. Further noting that $E_{\tilde{q}(\omega_i|\tilde{b}_i,\tilde{c}_i)}[\omega_i] = \frac{\tilde{b}_i}{2\tilde{c}_i}\tanh(\tilde{c}_i/2)$, we can simplify further to get:

$$\begin{aligned}\tilde{F} = &-N\ln(2)\sum_{i=1}^{N}(y_i - 1/2)[\boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\tilde{\mu}_\alpha] - \frac{\tilde{b}_i}{2\tilde{c}_i}\tanh(\tilde{c}_i/2)\left(\left[\boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\tilde{\mu}_\alpha\right]^2 + \boldsymbol{z}_i^T\tilde{\boldsymbol{\Lambda}}_\alpha\boldsymbol{z}_i\right)/2 + \frac{\tilde{b}_i\tilde{c}_i}{4}\tanh(\tilde{c}_i/2) - \ln\cosh(\tilde{c} \\ &1/2\ln(|2\pi\boldsymbol{T}|) - 1/2\mathrm{tr}\left(\boldsymbol{T}^{-1}\left[\tilde{\mu}_\alpha\tilde{\mu}_\alpha^T + \tilde{\boldsymbol{\Lambda}}_\alpha\right]\right) + 1/2\ln(|2\pi e\tilde{\boldsymbol{\Lambda}}_\alpha|) + \ln p_0(\boldsymbol{\beta}) + \ln p_0(\{\boldsymbol{\Sigma}_{j,\alpha}\})\end{aligned} \tag{25}$$

From this, we can derive the variational EM algorithm. As expected, this shows a remarkable resemblance to the Gibbs Sampler above. The variational distribution on the random effects $\boldsymbol{\alpha}$ is identical to the full conditional if we substitute ω_i with its (variational) expectation; the variational distribution on ω_i is similar.[30]

In the M-Step, therefore, we can simply perform a modified version of the linear mixed effects implementation of the EM algorithm (e.g. following Meng and Van Dyk 1998), noting that we are using a variational approximation. This intuitive makes sense; as stated above, if we knew the Polya-Gamma variables, the problem is trivial to solve via exact methods (e.g. EM) and thus we plug in a good guess as to their value from the variational approximation. The algorithm can thus be written as follows:

Variational Inference for Hierarchical Logistic Regression

Set $\boldsymbol{\beta}^{(0)}$, $\{\boldsymbol{\Sigma}_{j,\alpha}^{(0)}\}$, $\tilde{\boldsymbol{\mu}}_\alpha^{(0)}$, $\tilde{\boldsymbol{\Lambda}}_\alpha^{(0)}$ for $\boldsymbol{\alpha}$ and $\tilde{b}_i^{(0)}$ and $\tilde{c}_i^{(0)}$ for each i.

For each t in $1, \cdots, T$:

1. Variational E-Step (Polya-Gamma)

$$\tilde{b}_i^{(t)} = 1$$

$$\tilde{c}_i^{(t)} = \sqrt{E_q[\left(\boldsymbol{x}_i^T\boldsymbol{\beta}^{(t-1)} + z_i^T\boldsymbol{\alpha}\right)^2]} = \sqrt{\left(\boldsymbol{x}_i^T\boldsymbol{\beta}^{(t-1)} + \boldsymbol{z}_i^T\tilde{\boldsymbol{\mu}}_\alpha^{(t-1)}\right)^2 + \boldsymbol{z}_i^T\tilde{\boldsymbol{\Lambda}}_\alpha^{(t-1)}\boldsymbol{z}_i}$$

$$\tilde{\mu}_{\omega,i}^{(t)} = E_{q(\omega_i|\tilde{b}_i^{(t)},\tilde{c}_i^{(t)})}[\omega_i] = \frac{1}{2\tilde{c}_i^{(t)}}\tanh(\tilde{c}_i^{(t)}); \quad \tilde{\boldsymbol{\Omega}}^{(t)} = \text{diag}\left(\{\tilde{\mu}_{\omega,i}^{(t)}\}\right)$$

2. Variational E-Step (Random Effects). Update $\tilde{\boldsymbol{\mu}}_\alpha^{(t)}$ and $\tilde{\boldsymbol{\Lambda}}^{(t)}$.

$$\tilde{\boldsymbol{\Lambda}}_\alpha^{(t)} = \left(\boldsymbol{Z}^T\tilde{\boldsymbol{\Omega}}^{(t)}\boldsymbol{Z} + \left[\boldsymbol{T}^{(t-1)}\right]^{-1}\right)^{-1}$$

$$\tilde{\boldsymbol{\mu}}_\alpha^{(t)} = \tilde{\boldsymbol{\Lambda}}_\alpha^{(t)}\boldsymbol{Z}^T\left(\boldsymbol{s} - \tilde{\boldsymbol{\Omega}}^{(t)}\boldsymbol{X}\boldsymbol{\beta}^{(t-1)}\right)$$

3. Evaluate $\tilde{F}$. See main text for discussion.

4. M-Step for $\boldsymbol{\beta}$

$$\boldsymbol{\beta}^{(t)} = \left(\left[\tilde{\boldsymbol{X}}^{(t)}\right]^T\left[\tilde{\boldsymbol{X}}^{(t)}\right] + \boldsymbol{\Lambda}_\beta^0\right)^{-1}\left[\tilde{\boldsymbol{X}}^{(t)}\right]^T\left(\tilde{\boldsymbol{s}}^{(t)} - \tilde{\boldsymbol{Z}}\tilde{\boldsymbol{\mu}}_\alpha^{(t)}\right)$$

5. M-Step for

$$\boldsymbol{\Sigma}_{j,\alpha}^{(t)} = \frac{1}{G_j + \nu_j + S_j + 1}\left(\boldsymbol{\Psi}_j + \sum_{g=1}^{G_j}[\tilde{\boldsymbol{\mu}}_\alpha^{(t)}]_{j,g}\left[[\tilde{\boldsymbol{\mu}}_\alpha^{(t)}]_{j,g}\right]^T + [\tilde{\boldsymbol{\Lambda}}_\alpha^{(t)}]_{j,g}\right)$$

To evaluate the convergence of the algorithm, at the end of each iteration, the software uses two strategies. First, it looks for a sufficiently small decrease in the objective $\tilde{F}$; second, it looks for a sufficiently small change in $\boldsymbol{\beta}^{(t)}$ versus its prior value.

[30]Note that plugging in $E[\boldsymbol{\alpha}]$ recovers the approximate algorithm in Goplerud et al. 2018. This similarity to a variational algorithm likely explains its performance.

- Strategy 1: If $\tilde{F}^{(t)} - \tilde{F}^{(t-1)} < \epsilon$, i.e. the objective function is not increasing by an appreciable amount, terminate the algorithm. Note that, somewhat unusually, it is actually easier to evaluate $\tilde{F}$ after the E-Step but *before* the M-Step. In this case, it simplifies as follows:

$$\tilde{\tilde{F}}^{(t)} = -N\ln(2)\sum_{i=1}^{N}(y_i - 1/2)[\boldsymbol{x}_i^T\boldsymbol{\beta}^{(t-1)} + \boldsymbol{z}_i^T\tilde{\mu}_\alpha^{(t)}] - \ln\cosh(\tilde{c}_i^{(t)}/2)$$

$$1/2\ln(|2\pi\boldsymbol{T}^{(t)}|) - 1/2\text{tr}\left([\boldsymbol{T}^{(t)}]^{-1}\left[\tilde{\mu}_\alpha^{(t)}[\tilde{\mu}_\alpha^{(t)}]^T + \tilde{\boldsymbol{\Lambda}}_\alpha^{(t)}\right]\right) + 1/2\ln(|2\pi e\tilde{\boldsymbol{\Lambda}}_\alpha^{(t)}|)$$

  Since we are relying on a coordinate ascent algorithm, we know that both the E-Step and the M-Step increase $\tilde{F}$ and thus $\tilde{\tilde{F}}$ will also be monotonically increasing. Thus, we can track the progress of $\tilde{\tilde{F}}$ to assess convergence of the algorithm.

- Strategy 2: If $||\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}||_\infty < \epsilon$, stop the algorithm. That is, if the $\boldsymbol{\beta}$ are no longer moving by appreciable amounts.

B.3 Negative Binomial

This is a nearly exact replication of the logistic regression. As the densities in Table 1 show, the only real difference is the existence of the dispersion parameter r where larger values indicate *less* over-dispersion. Zhou et al. (2012) and Pillow and Scott (2012) have analyzed negative binomial models using Polya-Gamma augmentation, although they are not quite suited to our purposes.[31]

From before, the likelihood (regular and augmented) can be written as follows, defining $s_i = (y_i - r)/2$.

$$L(\boldsymbol{\beta}, r) = \int\prod_{i=1}^{N}\frac{\Gamma(y+r)}{\Gamma(r)\Gamma(y+1)}\frac{\exp(\psi_i - \ln r)^y}{[1+\exp(\psi_i - \ln r)]^{r+y}}f(\boldsymbol{\alpha}|\boldsymbol{0}, \boldsymbol{T})d\boldsymbol{\alpha} \tag{26a}$$

$$L_c(\boldsymbol{\beta}, r) = \prod_{i1=1}^{N}\frac{\Gamma(y+r)}{\Gamma(r)\Gamma(y+1)}2^{-(y_i+r)}\exp\left(s_i[\boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{\alpha} - \ln r] - \omega_i\left[[\boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{\alpha} - \ln r\right]^2/2\right)f(\boldsymbol{\alpha}|\boldsymbol{0}, \boldsymbol{T}) \tag{26b}$$

Using that result, the Gibbs Sampler is shown below

Gibbs Sampler for Hierarchical Negative Binomial Regression Preliminaries

- Define $s_i = (y_i - r)/2$.
- Set priors such as $\boldsymbol{\beta} \sim N(\boldsymbol{0}, \boldsymbol{\Lambda}_\beta)$, $r \sim Expo(r_0)$, $IW(\boldsymbol{\Psi}_j, \nu_j)$ for $\boldsymbol{\Sigma}_{j,\alpha}$
- Set $\boldsymbol{\beta}^{(0)}$, $(\boldsymbol{\Sigma}_{j,\alpha})^{(0)}$, $\boldsymbol{\alpha}^{(0)}$. $\boldsymbol{T}^{(0)}$ is defined implicitly.

For t in $1, \cdots, T$

[31] Pillow and Scott (2012) assumes that the dispersion parameter is fixed, though we wish to estimate this using the data in most social scientific applications. Zhou et al. (2012) rely on a different parameterization of the negative binomial that has the undesirable property that $E[Y_i] \neq \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})$. They do that to maintain tractable inference for r via a clever type of data augmentation.

1. Set $s_i^{(t)} = \left(y_i - r^{(t-1)}\right)/2$

2. Sample $\omega_i^{(t)}$.

$$\omega_i^{(t)}|- \sim PG\left(y_i + r^{(t-1)}, \boldsymbol{x}_i^T\boldsymbol{\beta}^{(t-1)} + \boldsymbol{z}_i^T\boldsymbol{\alpha}^{(t-1)} - \ln r^{(t-1)}\right)$$

3. Rescale the data

$$\tilde{s}_i = \frac{s_i^{(t)}}{\sqrt{\omega_i^{(t)}}} + \sqrt{\omega_i^{(t)}}\ln r^{(t-1)}; \quad \tilde{\boldsymbol{x}}_i = \sqrt{\omega_i^{(t)}}\boldsymbol{x}_i; \quad \tilde{\boldsymbol{z}}_i = \sqrt{\omega_i^{(t)}}\boldsymbol{z}_i$$

4. Sample $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{j,\alpha}$ as in in the logistic algorithm.

5. Sample r: Relying on partially collapsed Gibbs Samplers (Van Dyk and Park 2008) to sample r from its full conditional *integrating away* the weights. Its posterior density is

$$p(r|-) \propto p_0(r)\prod_i \frac{\Gamma(r+y_i)}{\Gamma(r)}(1-p_i)^{y_i}(p_i)^r; \quad 1 - p_i = \frac{\exp(\psi_i - \ln r)}{(1+\exp(\psi_i - \ln r))}$$

$$\psi_i^{(t)} = \boldsymbol{x}_i^T\boldsymbol{\beta}^{(t)} + \boldsymbol{z}_i^T\boldsymbol{\alpha}^{(t)}$$

The associated software uses slice sampling (Neal 2003), but other methods are acceptable.[a]

[a]Slice sampling is attractive because the full conditional is unimodal. The scheme implemented in the associated software is as follows: (a) find the posterior mode; (b) find the boundaries of the slice, and (c) following Neal's procedure for sampling from this random variable.

B.3.1 Variational EM

In the case without random effects, the EM algorithm proceeds very straightforwardly. To update the $\boldsymbol{\beta}$, we simply use the new expectation of $\omega_{ij}^{(t)}$ as implied by this Gibbs Sampler—that depends on $r^{(t-1)}$. There is one additional M-Step where we update r from the un-augmented posterior—i.e. find the posterior mode. This is valid by the AECM algorithm (Meng and Van Dyk 1997).

With random effects, again approximations are needed. We can use a similar strategy to derive the $\tilde{F}$ function again only assuming a mean-field approximation:

$$\begin{aligned}
\tilde{F} = &\sum_{i=1}^{N}(y_i - r)/2[\boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\tilde{\mu}_\alpha - \ln r] + \\
&- E_{\tilde{q}(\omega_i|\tilde{b}_i,\tilde{c}_i)}[\omega_i]\left(\left[\boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\tilde{\mu}_\alpha - \ln r\right]^2 + \boldsymbol{z}_i^T\tilde{\boldsymbol{\Lambda}}_\alpha\boldsymbol{z}_i\right)/2 + E_{\tilde{q}(\omega_i|\tilde{b}_i,\tilde{c}_i)}[\ln f(\omega_i|y_i + r, 0)] + \\
&1/2\ln(|2\pi\boldsymbol{T}|) - 1/2\mathrm{tr}\left(\boldsymbol{T}^{-1}\left[\tilde{\mu}_\alpha\tilde{\mu}_\alpha^T + \tilde{\boldsymbol{\Lambda}}_\alpha\right]\right) + 1/2\ln(|2\pi e\tilde{\boldsymbol{\Lambda}}_\alpha|) + \\
&\sum_{i=1}^{N}\tilde{c}_i^2/2E_{\tilde{q}(\omega_i|\tilde{b}_i,\tilde{c}_i)}[\omega_i] - E_{\tilde{q}(\omega_i|\tilde{b}_i,\tilde{c}_i)}[\ln f(\omega_i|\tilde{b}_i, 0)] - \tilde{b}_i\ln\cosh(\tilde{c}_i/2)
\end{aligned} \tag{27}$$

The key complication here is that the intractable terms, $E_{\tilde{q}(\omega_i)}[\ln f(\omega_i|b, 0)]$ do not cancel as

one involves $\ln f(\omega_i|y_i + r, 0)$ and the other involves $\ln f(\omega_i|\tilde{b}_i, 0)$. This will complicate inference for r and for actually evaluating the $\tilde{F}$ directly. Both of these points are addressed by noting the following point: $\tilde{F}$ can be re-written to eliminate the variational parameters. Since $\tilde{b}_i$ and $\tilde{c}_i$ are deterministic functions of r and the other parameters and data, they can be substituted away. Call this substituted function $\tilde{\tilde{F}}$. As in the logistic case, many terms—including the intractable ones—disappear.

$$\tilde{\tilde{F}} = \sum_{i=1}^{N}(y_i - r)/2[\boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\tilde{\mu}_\alpha - \ln r] - (y_i + r)\ln\cosh\left(1/2 \cdot \sqrt{\left(\boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\tilde{\boldsymbol{\mu}}_\alpha - \ln r\right)^2 + \boldsymbol{z}_i^T\tilde{\boldsymbol{\Lambda}}_\alpha\boldsymbol{z}_i}\right) \tag{28}$$
$$1/2\ln(|2\pi\boldsymbol{T}|) - 1/2\text{tr}\left(\boldsymbol{T}^{-1}\left[\tilde{\mu}_\alpha\tilde{\mu}_\alpha^T + \tilde{\boldsymbol{\Lambda}}_\alpha\right]\right) + 1/2\ln(|2\pi e\tilde{\boldsymbol{\Lambda}}_\alpha|)$$

This would lead to complicated inference for $\boldsymbol{\beta}$ and $\{\boldsymbol{\Sigma}_{j,\alpha}\}$, but it (i) can be evaluated to monitor convergence and (ii) provides a way to update r.

Variational Inference for Hierarchical Negative Binomial Regression

Preliminaries

- Initialize $\boldsymbol{\beta}^{(0)}$, $r^{(0)}$, and $\boldsymbol{\Sigma}^{(0)}$.
- Initialize the variational parameters: $\tilde{b}_i^{(0)}$, $\tilde{c}_i^{(0)}$ and $\tilde{\boldsymbol{\mu}}_\alpha^{(0)}$ and $\tilde{\boldsymbol{\Lambda}}_\alpha^{(0)}$.

For some number of iterations $t \in \{1, \cdots, T\}$.

1. Variational E-Step (Polya-Gamma). Update the parameters for $q(\omega_i)$, again by inspecting the full conditional on ω_i from the Gibbs Sampler. The key difference versus before is that $\tilde{b}_i^{(t)}$ depends on r.

$$\tilde{b}_i^{(t)} = y_i + r^{(t-1)}$$

$$\begin{aligned}\tilde{c}_i^{(t)} &= \sqrt{E_q[\left(\boldsymbol{x}_i^T\boldsymbol{\beta}^{(t-1)} + z_i^T\boldsymbol{\alpha} - \ln r^{(t-1)}\right)^2]} \\ &= \sqrt{\left(\boldsymbol{x}_i^T\boldsymbol{\beta}^{(t-1)} + \boldsymbol{z}_i^T\tilde{\boldsymbol{\mu}}_\alpha^{(t-1)} - \ln r^{(t-1)}\right)^2 + \boldsymbol{z}_i^T\tilde{\boldsymbol{\Lambda}}_\alpha^{(t-1)}\boldsymbol{z}_i}\end{aligned}$$

$$\tilde{\mu}_{\omega,i}^{(t)} = E_{q(\omega_i|\tilde{b}_i^{(t)},\tilde{c}_i^{(t)})}[\omega_i] = \frac{\tilde{b}_i^{(t)}}{2\tilde{c}_i^{(t)}}\tanh(\tilde{c}_i^{(t)}/2); \quad \tilde{\boldsymbol{\Omega}}^{(t)} = \text{diag}\left(\{\tilde{\mu}_{\omega,i}^{(t)}\}\right)$$

2. Variational E-Step (Random Effects). Update the parameters for $q(\boldsymbol{\alpha})$:

$$\tilde{\boldsymbol{\Lambda}}_\alpha^{(t)} = \left(\boldsymbol{Z}^T\tilde{\boldsymbol{\Omega}}^{(t)}\boldsymbol{Z} + \left[\boldsymbol{T}^{(t-1)}\right]^{-1}\right)^{-1}$$

$$\tilde{\boldsymbol{\mu}}_\alpha^{(t)} = \tilde{\boldsymbol{\Lambda}}_\alpha^{(t)}\boldsymbol{Z}^T\left(\boldsymbol{s} - \tilde{\boldsymbol{\Omega}}^{(t)}\boldsymbol{X}\boldsymbol{\beta}^{(t-1)} - \tilde{\boldsymbol{\Omega}}^{(t)}\mathbf{1}_{N\times 1}\ln r^{(t-1)}\right)$$

3. M-Step: $\boldsymbol{\beta}$. Scale the data by $E[\omega_i] = \tilde{\mu}_{\omega,i}^{(t)}$ as noted in the Gibbs Sampler. Update the

$\boldsymbol{\beta}$:

$$\boldsymbol{\beta}^{(t)} = \left(\left[\tilde{\boldsymbol{X}}^{(t)} \right]^T \left[\tilde{\boldsymbol{X}}^{(t)} \right] + \boldsymbol{\Lambda}_\beta^0 \right)^{-1} \left[\tilde{\boldsymbol{X}}^{(t)} \right]^T \left(\tilde{\boldsymbol{s}}^{(t)} - \tilde{\boldsymbol{Z}} \tilde{\boldsymbol{\mu}}_\alpha^{(t)} - [\tilde{\boldsymbol{\Omega}}^{(t)}]^{1/2} \mathbf{1}_{N \times 1} \ln r^{(t-1)} \right)$$

4. M-Step: Update $\{\boldsymbol{\Sigma}_{j,\alpha}\}$.

$$\boldsymbol{\Sigma}_{j,\alpha}^{(t)} = \frac{1}{G_j + \nu_j + S_j + 1} \left(\boldsymbol{\Psi}_j + \sum_{g=1}^{G_j} [\tilde{\boldsymbol{\mu}}_\alpha^{(t)}]_{j,g} \left[[\tilde{\boldsymbol{\mu}}_\alpha^{(t)}]_{j,g} \right]^T + [\tilde{\boldsymbol{\Lambda}}_\alpha^{(t)}]_{j,g} \right)$$

5. M-Step Update r: A naive attempt to optimize $\tilde{F}$ w.r.t. r runs into intractable trouble because of the terms $E_{\tilde{q}(\omega_i|\tilde{b}_i^{(t)}, \tilde{c}_i^{(t)})}[\ln f(\omega_i | y_i + r, 0)]$. As shown above, however, I optimize the (tractable) $\tilde{\tilde{F}}$:

$$r^{(t)} = \arg\max_r \tilde{\tilde{F}}$$

   This is a simple one-dimensional optimization problem.

6. Evaluate the convergence of the algorithm. Use the same strategies from the logistic case; either look at the stationarity of the $\boldsymbol{\beta}$ or the change in $\tilde{\tilde{F}}$.

B.4 Multinomial

Assume that each observation i engages in a choice between K choices. We can extend the logistic sampler to the K choice case with some additional work. For simplicity, assume the same covariates across levels though this can be generalized. The generative model is as follows:

$$p(y_i = k) = \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_k + \boldsymbol{z}_i^T \boldsymbol{\alpha}_k)}{\sum_l \exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta}_l + \boldsymbol{z}_i^T \boldsymbol{\alpha}_l\right)} \tag{29}$$

The only difference is now that there are different coefficients $\boldsymbol{\beta}_k$ for each level and random effects $\boldsymbol{\alpha}_k$. For identification, we assume that $\boldsymbol{\beta}_K = 0$ and $\boldsymbol{\alpha}_{j,K} = 0$ for all j.

For the random effect, we use a stacked notation: $\boldsymbol{\aleph}$ is a stacked vector of $\boldsymbol{\alpha}$ for $k \in \{1, \cdots, K-1\}$ with variance $\boldsymbol{\mathcal{T}}$.

$$\boldsymbol{\aleph} = \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \cdots \\ \boldsymbol{\alpha}_{K-1} \end{bmatrix} \sim MVN(0, \boldsymbol{\mathcal{T}}) \tag{30}$$

The structure of $\boldsymbol{\mathcal{T}}$ is banded block-diagonal. This can be derived by first thinking of a single level of a random effect: For some level g from random effect j, we have the following distribution:

$$\boldsymbol{\alpha}_{j,g} = \begin{bmatrix} \boldsymbol{\alpha}_{j,g,1} \\ \cdots \\ \cdots \\ \boldsymbol{\alpha}_{j,g,K-1} \end{bmatrix} \sim MVN(0, \boldsymbol{\mathcal{S}}_{j,\alpha}); \quad \boldsymbol{\mathcal{S}}_{j,\alpha} = \begin{pmatrix} \boldsymbol{\Sigma}_{j,1,\alpha} & \boldsymbol{\mathcal{C}}_{j,1-2,\alpha} & \cdots & \boldsymbol{\mathcal{C}}_{j,1-K-1,\alpha} \\ \boldsymbol{\mathcal{C}}_{j,1-2,\alpha} & \boldsymbol{\Sigma}_{j,2,\alpha} & \cdots & \boldsymbol{\mathcal{C}}_{j,2-K-1,\alpha} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \boldsymbol{\Sigma}_{j,K-1,\alpha} \end{pmatrix} \tag{31}$$

As before, $\boldsymbol{\Sigma}_{j,k,\alpha}$ represents the variance of a random effect for a particular level, i.e. random effect j and outcome choice k. $\boldsymbol{\mathcal{C}}_{j,k-k',\alpha}$ represents the *covariance* between levels of the random effect that should not, *a priori* be restricted to be zero.

Given the $\boldsymbol{\mathcal{S}}_{j,a}$, we can formulate two versions of $\boldsymbol{\mathcal{T}}$. One, perhaps the most intuitive one, stacks them in a block-diagonal fashion as in the logistic case. Call this $\boldsymbol{\mathcal{T}}'$. Each block of $G_j \cdot S_j \cdot K - 1$ rows/columns corresponds to the first random effect. This ordering, however, does not match what is used in the likelihood function as that picks out all elements of $\boldsymbol{\aleph}$ that correspond to the level k. Thus, we can define the primary $\boldsymbol{\mathcal{T}}$ as picking out the elements of each $\boldsymbol{\mathcal{S}}_{j,\alpha}$ that correspond to level k and then stacking those together for each G_j. For simplicity, consider the case of two random effects with one level each and $K = 3$:

$$\boldsymbol{\aleph}^T = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2] = [\alpha_{1,1}, \alpha_{2,1}, \alpha_{1,2}, \alpha_{2,2}] \quad \boldsymbol{\mathcal{T}} = \begin{pmatrix} \boldsymbol{\Sigma}_{1,1,\alpha} & 0 & \boldsymbol{\mathcal{C}}_{1,1-2,\alpha} & 0 \\ 0 & \boldsymbol{\Sigma}_{2,1,\alpha} & 0 & \boldsymbol{\mathcal{C}}_{2,1-2,\alpha} \\ \boldsymbol{\mathcal{C}}_{1,1-2,\alpha} & 0 & \boldsymbol{\Sigma}_{1,2,\alpha} & 0 \\ 0 & \boldsymbol{\mathcal{C}}_{2,1-2,\alpha} & 0 & \boldsymbol{\Sigma}_{2,2,\alpha} \end{pmatrix} \tag{32}$$

Given this notation, we can now turn to the model itself. Relying on the results in the main text, the full likelihood is:

$$\prod_i \prod_{k=1}^{K-1} \left[\frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_k + \boldsymbol{z}_i^T \boldsymbol{\alpha}_k)}{\sum_l \exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta}_l + \boldsymbol{z}_i^T \boldsymbol{\alpha}_l\right)} \right]^{I(y_i = k)}$$

This seems rather intractable and, indeed, it is not possible to sample all of the $\boldsymbol{\beta}_k$ in a single block in a simple fashion. Thus, we rely on insights in Polson, Scott, and Windle (2013) that we can focus on the conditional likelihood: Focusing on on *level* k and conditioning on all others, we can write the likelihood as follows:

$$p(\boldsymbol{\beta}_k, \boldsymbol{\alpha}_k | \boldsymbol{\beta}_{\neg k}, \boldsymbol{\alpha}_{\neg k}) \propto \prod_i \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_k + \boldsymbol{z}_i^T \boldsymbol{\alpha}_k - \ln C_{ik})^{I(y_i = k)}}{1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_k + \boldsymbol{z}_i^T \boldsymbol{\alpha}_k - \ln C_{ik})}; \quad C_{i,k} = \sum_{l \neq k} \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_l + \boldsymbol{z}_i^T \boldsymbol{\alpha}_k) \tag{33}$$

To be complete, however, we must also condition the random effect component, i.e. $p(\boldsymbol{\alpha}_k | \boldsymbol{\mathcal{T}}, \boldsymbol{\alpha}_{\neg k})$. This can be done most simply by focusing on $\boldsymbol{\mathcal{T}}'$ and then permuting this matrix. Focusing on a single $\boldsymbol{\alpha}_{j,g}$:

$$p(\boldsymbol{\alpha}_{j,g,k} | \boldsymbol{\alpha}_{j,g,\neg k}, \mathcal{S}_{j,\alpha}) \sim MVN\left(\boldsymbol{\mu}_{j,g|\neg k}, \boldsymbol{\Sigma}_{j,g|\neg k}\right) \tag{34a}$$

$$\boldsymbol{\mu}_{j,g|\neg k} = [\boldsymbol{\mathcal{S}}_{j,a}]_{k,\neg k} \left[[\boldsymbol{\mathcal{S}}_{j,a}]_{\neg k,\neg k}\right]^{-1} \boldsymbol{\alpha}_{j,g,\neg k} \tag{34b}$$

$$\boldsymbol{\Sigma}_{j,g|\neg k} = [\boldsymbol{\mathcal{S}}_{j,\alpha}]_{k,k} - [\boldsymbol{\mathcal{S}}_{j,a}]_{k,\neg k} \left[[\boldsymbol{\mathcal{S}}_{j,a}]_{\neg k,\neg k}\right]^{-1} [\boldsymbol{\mathcal{S}}_{j,a}]_{k,\neg k}^T \tag{34c}$$

These are then stacked together as before to create $\boldsymbol{\mathcal{S}}_{j,a|\neg k}$. These are then stacked together in a block-diagonal form to make $\boldsymbol{\mathcal{T}}'_{\neg k}$. Permuting the rows to match the order of $\boldsymbol{z}_i$ gives us $\mathcal{T}_{\neg k}$.

Given this book-keeping, I derive a Gibbs Sampler using exactly the algorithm noted above. The only difference is that our prior on the random effects is the conditional one noted above and there is an offset of $-\ln C_{ik}$ for each observation.

Gibbs Sampler for Multinomial Hierarchical Regression

Preliminaries

- Initialize some value of $\boldsymbol{\beta}^{(0)}$, $(\boldsymbol{\mathcal{S}}_{j,\alpha})^{(0)}$, $\boldsymbol{\alpha}^{(0)}$. We have therefore defined $\boldsymbol{\mathcal{T}}^{(0)}$ implicitly.
- The priors are as follows: $\boldsymbol{\beta}_k \sim N(\mathbf{0}, \boldsymbol{\Lambda}_\beta^0)$; for each random effect, $\boldsymbol{\mathcal{S}}_{j,\alpha} \sim IW(\boldsymbol{\Psi}_j, \nu_j)$; the dimensionality of $\boldsymbol{\Psi}_j$ is $S_j(K-1) \times S_j(K-1)$. Thus, for us to have a proper prior, $\nu_j > S_j(K-1)$.

For $t \in \{1, \cdots, T\}$

1. For $k \in \{1, \cdots, K-1\}$:
   (a) Define $s_{i,k} = I(y_i = k) - 1/2$
   (b) Sample $\omega_{i,k}^{(t)} \quad \forall i$

$$\omega_{i,k}^{(t)}|- \sim PG\left(1, \boldsymbol{x}_i^T \boldsymbol{\beta}_k^{(t-1)} + \boldsymbol{z}_i^T \boldsymbol{\alpha}_k^{(t-1)} - \ln C_{i,k}\right)$$

   (c) Turn the model into weighted least squares by rescaling the data as follows:

$$\tilde{s}_{i,k}^{(t)} = \frac{s_{i,k}}{\sqrt{\omega_{i,k}^{(t)}}} + \sqrt{\omega_{i,k}^{(t)}} \ln C_{i,k}$$

$$\tilde{\boldsymbol{x}}_{i,k}^{(t)} = \boldsymbol{x}_i \cdot \sqrt{\omega_{i,k}^{(t)}}$$

$$\tilde{\boldsymbol{z}}_{i,k}^{(t)} = \boldsymbol{z}_i \cdot \sqrt{\omega_{i,k}^{(t)}}$$

   (d) Calculate the (approximate) conditional prior:

$$\boldsymbol{\mu}_{j,g|\neg k} = [\boldsymbol{\mathcal{S}}_{j,a}]_{k,\neg k} \left[[\boldsymbol{\mathcal{S}}_{j,a}]_{\neg k,\neg k}\right]^{-1} \boldsymbol{\mu}_{j,g,\neg k}^{(t-1)}$$

$$\boldsymbol{\Sigma}_{j,g|\neg k} = [\boldsymbol{\mathcal{S}}_{j,\alpha}]_{k,k} - [\boldsymbol{\mathcal{S}}_{j,a}]_{k,\neg k} \left[[\boldsymbol{\mathcal{S}}_{j,a}]_{\neg k,\neg k}\right]^{-1} [\boldsymbol{\mathcal{S}}_{j,a}]_{k,\neg k}^T$$

   By permutation, generate $\boldsymbol{\mathcal{T}}_{\neg k}^{(t)}$. Calculate its inverse. Similarly, permute $\boldsymbol{\mu}_{j,g|\neg k}$ to get the correct alignment: $\tilde{\boldsymbol{\mu}}_{\neg k}$

   (e) Sample $\boldsymbol{\alpha}_k^{(t)}$:

$$\boldsymbol{\Lambda}_{\alpha,k}^{(t)} = \left(\left[\tilde{\boldsymbol{Z}}^{(t)}\right]^T \tilde{\boldsymbol{Z}}^{(t)} + [\boldsymbol{\mathcal{T}}_{\neg k}^{(t)}]^{-1}\right)^{-1}$$

$$\boldsymbol{\mu}_{\alpha,k}^{(t)} = \boldsymbol{\Lambda}_{\alpha,k}^{(t)} \left[[\boldsymbol{\mathcal{T}}_{\neg k}^{(t)}]^{-1} \tilde{\boldsymbol{\mu}}_{\neg k} + \left[\tilde{\boldsymbol{Z}}^{(t)}\right]^T \left(\tilde{\boldsymbol{s}}_k^{(t)} - \tilde{\boldsymbol{X}}^{(t)} \boldsymbol{\beta}_k^{(t-1)}\right)\right]$$

$$\boldsymbol{\alpha}_k^{(t)}|- \sim N\left(\boldsymbol{\mu}_{\alpha,k}^{(t)}, \boldsymbol{\Lambda}_{\alpha,k}^{(t)}\right)$$

   (f) Sample $\boldsymbol{\beta}_k^{(t)}$.

$$\boldsymbol{\Lambda}_{\beta,k}^{(t)} = \left(\left[\tilde{\boldsymbol{X}}^{(t)}\right]^T \left[\tilde{\boldsymbol{X}}^{(t)}\right] + \boldsymbol{\Lambda}_\beta^0\right)^{-1}$$

$$\boldsymbol{\mu}_{\beta,k}^{(t)} = \boldsymbol{\Lambda}_{\beta,k}^{(t)} \left[\tilde{\boldsymbol{X}}^{(t)}\right]^T \left(\tilde{\boldsymbol{s}}_k^{(t)} - \tilde{\boldsymbol{Z}}\boldsymbol{\alpha}_k^{(t)}\right)$$

$$\boldsymbol{\beta}_k^{(t)}|- \sim N\left(\boldsymbol{\mu}_{\beta,k}^{(t)}, \boldsymbol{\Lambda}_{\beta,k}^{(t)}\right)$$

2. Sample $\boldsymbol{\mathcal{S}}_{\alpha,j}$, i.e. the variance for each random effect j. We can rely on the block-diagonal representation of $\mathcal{T}$. By similar re-arrangement to above, we can write

$$\boldsymbol{\mathcal{S}}_{\alpha,j}^{(t)} \sim IW\left(\boldsymbol{\Psi}_j + \sum_{g=1}^{G_j} \boldsymbol{\alpha}_{j,g}^{(t)} \left[\boldsymbol{\alpha}_{j,g}^{(t)}\right]^T, G_j + \nu_j\right)$$

B.4.1 Variational Inference

As noted above, to estimate the model without random effects, we simply set all $\boldsymbol{z}_i = \boldsymbol{0}$, ignore all updates involving $\boldsymbol{\alpha}$ and $\boldsymbol{\mathcal{S}}_{j,\alpha}$, and replace $\omega_{i,k}^{(t)}$ with its expectation wherever it appears. This algorithm is valid via the AECM algorithm (Meng and Van Dyk 1997) where the algorithm can be thought of as an E-Step and an M-Step for each level k with having different Polya-Gamma variables. Without random effects, this procedure again has guaranteed convergence.

With random effects, point inference becomes challenging for the reasons outlined in the logistic case. Further, everything becomes even more challenging because of the need to condition on the random effects. Future work may try to use a different representation of the multinomial logistic regression (stick-breaking representation, see Linderman, Johnson, and Adams 2015; Goplerud 2018) that may be more tractable for EM inference. However, we can derive a variational approximation in two steps, although it requires additional approximations to the earlier approaches. This may explain its weaker performance.

First, the key factorization assumption is that the random effects are independent across levels *and* that the Polya-Gamms are independent across levels. This is a much stronger assumption than in prior models.

$$\tilde{q}(\aleph, \{\{\omega_{i,k}\}_{i=1}^N\}_{k=1}^{K-1}) = \prod_{k=1}^{K-1} \tilde{q}(\boldsymbol{\alpha}_k)\tilde{q}(\{\omega_{i,k}\}_{i=1}^N) \tag{35}$$

Next, to find the variational distribution of $\omega_{i,k}$ and $\boldsymbol{\alpha}_k$, we must know the mean and variance of $\ln C_{i,k} = \ln\left(\sum_{l\neq k} \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}_l + \boldsymbol{z}_i^T\boldsymbol{\alpha}_l)\right)$ over the variational distribution. Given the variational assumption of independence between the $\boldsymbol{\alpha}_k$, we can use the following insight: Define $L_{i,k} = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}_l + \boldsymbol{z}_i^T\boldsymbol{\alpha}_l)$. Given that, as we show momentarily, that $q(\boldsymbol{\alpha}_l)$ is multivariate normal, then $L_{i,k} \sim LogNormal\left(\boldsymbol{x}_i^T\boldsymbol{\beta}_l + \boldsymbol{z}_i^T E[\boldsymbol{\alpha}_l], \boldsymbol{z}_i^T Var(\boldsymbol{\alpha}_l)\boldsymbol{z}_i\right)$. Further, crucially, $L_{i,k}$ are independent for all k!

We can use the Fenton approximation for this (Fenton 1960): Assume that $L_{i,k}$ are independent

log-normals with parameters $m_{i,k}$ and $s^2_{i,k}$:

$$\tilde{L}_{i,k} \approx LogNormal\left(\tilde{m}_{i,k}, \tilde{s}^2_{i,k}\right) \tag{36a}$$

$$\tilde{m}_{i,k} = \ln\left(\sum_{l\neq k}\exp(m_{i,l} + s^2_{i,l}/2)\right) - \tilde{s}^2_i/2 \tag{36b}$$

$$\tilde{s}^2_{i,k} = \ln\left[\frac{\sum_{l\neq k}\exp(2\cdot m_{i,l} + s^2_{i,l})\left(\exp(s^2_{i,l}) - 1\right)}{\left(\sum_{l\neq k}\exp(m_{i,l} + s^2_{i,l}/2)\right)^2} + 1\right] \tag{36c}$$

Note, further, that since $\tilde{L}_i$ is log-normal, its logarithm is normal with the parameters noted above. Thus, by the approximation:

$$\ln C_{i,k} \approx N(\tilde{m}_{i,k}, \tilde{s}^2_{i,k}) \tag{37}$$

Further, note this agrees with a 'naive' approach: If $\boldsymbol{\alpha}_k$ has a density that was a point-mass at $\boldsymbol{x}_i^T\boldsymbol{\beta}_l + \boldsymbol{z}_i^T E[\boldsymbol{\alpha}_l]$, then all $s^2_{i,k} = 0$ and $\tilde{s}^2_{i,k} = 0$ and thus $\tilde{m}_{i,k} = \ln\left(\sum_{l\neq k}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}_l + \boldsymbol{z}_i^T E[\boldsymbol{\alpha}_l])\right)$.

To derive the variational algorithm, I proceed cyclically as in the Gibbs Sampler. I first re-arrange the variational objective to factor out all terms that do not involve some level k.

$$\tilde{F} = E_{\tilde{q}}[\ln p(\boldsymbol{y}, \boldsymbol{\aleph}, \{\boldsymbol{\Omega}\}|\boldsymbol{X}, \{\boldsymbol{\beta}\})] - E_{\tilde{q}}[\ln \tilde{q}(\boldsymbol{\aleph}, \{\{\omega_{i,k}\}\})] \tag{38a}$$

$$= E_{\tilde{q}}[\ln p(\boldsymbol{y}, \boldsymbol{\alpha}_k, \{\omega_{i,k}\}|\boldsymbol{X}, \{\boldsymbol{\beta}\})] + E_{\tilde{q}}[\ln p(\{\boldsymbol{\alpha}'_k\}_{k'\neq k}, \{\omega_{i,k'}\}_{k'\neq k})] - E_{\tilde{q}}[\ln \tilde{q}(\boldsymbol{\aleph}, \{\{\omega_{i,k}\}\})] \tag{38b}$$

Note, therefore, that this is tractable given the assumed independence of the Polya-Gamma and random effect variational distributions *across* levels k. The variational updates for each $\omega_{i,k}$ are as follows:[32]

$$\tilde{q}(\omega_{i,k}) \sim PG(1, \tilde{c}_{i,k}) \tag{39a}$$

$$\tilde{c}_{i,k} = \sqrt{\left(\boldsymbol{x}_i^T\boldsymbol{\beta}_k + \boldsymbol{z}_i^T E_{\tilde{q}(\alpha_k)}[\boldsymbol{\alpha}_k] - E_{\tilde{q}(\boldsymbol{\alpha}_{\neg k})}[\ln C_{i,k}]\right)^2 + \boldsymbol{z}_i^T Var_{\tilde{q}(\boldsymbol{\alpha}_k)}(\boldsymbol{\alpha}_k)\boldsymbol{z}_i + Var_{\tilde{q}(\boldsymbol{\alpha}_{\neg k})}(\ln C_{i,i})} \tag{39b}$$

Now, consider $\boldsymbol{\alpha}_k$. All terms that involve $\boldsymbol{\alpha}_k$ can be written as, noting the conditioning of the multivariate normal:

$$\sum_i (I(y_i = k) - 1/2)\boldsymbol{z}_i^T\boldsymbol{\alpha}_k - \omega_{i,k}\left(\boldsymbol{x}_i^T\boldsymbol{\beta}_k + \boldsymbol{z}_i^T\boldsymbol{\alpha}_k - \ln C_{i,k}\right)^2/2 + \ln f(\boldsymbol{\alpha}_k|\boldsymbol{\alpha}_{\neg k}, \boldsymbol{T}) \tag{40}$$

Recall that $f(\boldsymbol{\alpha}_k|\boldsymbol{\alpha}_{\neg k}, \boldsymbol{T})$ is multivariate normal. From the block diagonal structure of $\boldsymbol{T}$, we can conclude that for some $\boldsymbol{\alpha}_{j,g,k}$ (i.e. some random effect j with level g at choice k), the distribution is multivariate normal with mean $\boldsymbol{\mu}_{j,g|\neg k}$ and variance $\boldsymbol{\Sigma}_{j,g|\neg k}$, with $\neg k$ noting the conditioning on

[32] Note that the only terms that involve $\omega_{i,k}$ in the log (factorized) complete data likelihood are written below. The exponential of this expectation gives the variational distribution.

$$-\omega_{i,k}E_{q(\boldsymbol{\alpha}_k)}\left[\left(\boldsymbol{x}_i^T\boldsymbol{\beta}_k + \boldsymbol{z}_i^T\boldsymbol{\alpha}_k - \ln C_{i,k}\right)^2\right]/2 + \ln f(\omega_{i,k}|1, 0)$$

$\boldsymbol{\alpha}_{\neg k}$:

$$\boldsymbol{\mu}_{j,g|\neg k} = [\boldsymbol{\mathcal{S}}_{j,a}]_{k,\neg k} \left[[\boldsymbol{\mathcal{S}}_{j,a}]_{\neg k,\neg k}\right]^{-1} \boldsymbol{\alpha}_{j,g,\neg k} \tag{41a}$$

$$\boldsymbol{\Sigma}_{j,g|\neg k} = [\boldsymbol{\mathcal{S}}_{j,\alpha}]_{k,k} - [\boldsymbol{\mathcal{S}}_{j,a}]_{k,\neg k} \left[[\boldsymbol{\mathcal{S}}_{j,a}]_{\neg k,\neg k}\right]^{-1} [\boldsymbol{\mathcal{S}}_{j,a}]^T_{k,\neg k} \tag{41b}$$

Noting the independence of $\omega_{i,k}$ and $\boldsymbol{\alpha}_k \forall k$ under our variational mean-field assumption, it is clear that the correct form of the variational distribution for $\boldsymbol{\alpha}_k$ is normal—if we condition on $\boldsymbol{\alpha}_{\neg k}$. This is exactly as in the logistic case.[33] Taking the expectation and pattern matching gives the variational distribution outlined below.

Thus, for some k, given all other parameters, we can update the variational distributions. With those in hand, we then update $\boldsymbol{\beta}_k$. By cycling through all k levels, we can update all $\boldsymbol{\beta}_k$. This "carries through" the variational distributions from prior levels when updating the new variational distributions. $\boldsymbol{\mathcal{S}}_{j,\alpha}$ is updated at the end as shown in the full algorithm. While this order is non-standard for a variational algorithm, again casting the problem as a cyclical one provides some grounding to this approach.

Variational Inference for Multinomial Hierarchical Regression

Preliminaries

- Initialize some value of $\boldsymbol{\beta}^{(0)}$, $(\boldsymbol{\mathcal{S}}_{j,\alpha})^{(0)}$. We have therefore defined $\boldsymbol{\mathcal{T}}^{(0)}$ implicitly.
- Initialize the variational parameters for $\boldsymbol{\alpha}_k$, i.e. $\{\tilde{\boldsymbol{\mu}}^{(t)}_{\alpha,k}, \tilde{\boldsymbol{\Lambda}}^{(t)}_{\alpha,k}\}$.
- The priors are as follows: $\boldsymbol{\beta}_k \sim N(\mathbf{0}, \boldsymbol{\Lambda}^0_\beta)$; for each random effect, $\boldsymbol{\mathcal{S}}_{j,\alpha} \sim IW(\boldsymbol{\Psi}_j, \nu_j)$; the dimensionality of $\boldsymbol{\Psi}_j$ is $S_j(K-1) \times S_j(K-1)$. Thus, for us to have a proper prior, $\nu_j > S_j(K-1)$.

For some number of iterations $t \in \{1, \cdots, T\}$.

1. For each level $k \in \{1, \cdots, K\}$:
   (a) Define $s_{i,k} = I(y_i = k) - 1/2$. Stack these into a vector to form $\boldsymbol{s}_k$.
   (b) Variational E-Step: In this step, we use the variational distributions w.r.t. $q(\boldsymbol{\alpha}_{\neg k})$. Update the variational Polya-Gamma distribution, noting that $\tilde{m}_{i,k}$ and $\tilde{s}^2_{i,k}$ are defined via Fenton's approximation above—and are themselves functions of all of

[33] Expanding the relevant equation gives

$$\sum_i (I(y_i = k) - 1/2)\boldsymbol{z}_i^T \boldsymbol{\alpha}_k - E[\omega_{i,k}] \left(\boldsymbol{x}_i^T \boldsymbol{\beta}_k + \boldsymbol{z}_i^T \boldsymbol{\alpha}_k - \ln C_{i,k}\right)^2 /2 - 1/2 \left(\boldsymbol{\alpha}_k - \boldsymbol{\mu}_{\neg k}\right)^T \boldsymbol{\Sigma}^{-1}_{\neg k} \left(\boldsymbol{\alpha}_k - \boldsymbol{\mu}_{\neg k}\right)$$

Taking the expectation over all other $q(\boldsymbol{\alpha}_{\neg k})$ to get the more appropriate variational distribution: We do this and ignoring irrelevant terms:

$$\sum_i (I(y_i = k) - 1/2)\boldsymbol{z}_i^T \boldsymbol{\alpha}_k - E[\omega_{i,k}] \left[\left(\boldsymbol{x}_i^T \boldsymbol{\beta}_k + \boldsymbol{z}_i^T \boldsymbol{\alpha}_k - E[\ln C_{i,k}]\right)^2\right] /2$$
$$-1/2 \left(\boldsymbol{\alpha}_k - E[\boldsymbol{\mu}_{\neg k}]\right)^T \boldsymbol{\Sigma}^{-1}_{\neg k} \left(\boldsymbol{\alpha}_k - E[\boldsymbol{\mu}_{\neg k}]\right)$$

By pattern-matching, we note that $\boldsymbol{\alpha}_k$ is multivariate normal as found in the Gibbs Sampler except that we replace $\ln C_{i,k}$ with its expectation, $\omega_{i,k}$ with its expectation, and the conditional mean of the random effects with its expectation.

the variational parameters of $q(\boldsymbol{\alpha}_{\neg k})$.

$$\tilde{q}(\omega_{i,k}) \sim PG(1, \tilde{c}_{i,k})$$

$$\tilde{c}_{i,k} = \sqrt{\left(\boldsymbol{x}_i^T \boldsymbol{\beta}_k^{(t-1)} + \boldsymbol{z}_i^T \tilde{\boldsymbol{\mu}}_{\alpha,k}^{(t-1)} - \tilde{m}_{i,k}^{(t-1)}\right)^2 + \boldsymbol{z}_i^T \tilde{\boldsymbol{\Lambda}}_{\alpha,k}^{(t-1)} \boldsymbol{z}_i + [\tilde{s}_{i,k}^2]^{(t-1)}}$$

Note that as before, $\tilde{\boldsymbol{\Omega}}^{(t)}$ is a diagonal matrix with the expectation of each Polya-Gamma on the diagonal.

Update the variational normal density for $q(\boldsymbol{\alpha}_k)$. First, calculate the conditional prior:

$$\check{\boldsymbol{\mu}}_{j,g|\neg k}^{(t)} = [\boldsymbol{\mathcal{S}}_{j,a}^{(t-1)}]_{k,\neg k} \left[[\boldsymbol{\mathcal{S}}_{j,a}^{(t-1)}]_{\neg k,\neg k}\right]^{-1} \boldsymbol{\mu}_{\alpha,j,g,\neg k}^{(t-1)}$$

$$\check{\boldsymbol{\Sigma}}_{j,g|\neg k}^{(t)} = [\boldsymbol{\mathcal{S}}_{j,\alpha}^{(t-1)}]_{k,k} - [\boldsymbol{\mathcal{S}}_{j,a}^{(t-1)}]_{k,\neg k} \left[[\boldsymbol{\mathcal{S}}_{j,a}^{(t-1)}]_{\neg k,\neg k}\right]^{-1} [\boldsymbol{\mathcal{S}}_{j,a}^{(t-1)}]_{k,\neg k}^T$$

Stack these together to get $\check{\boldsymbol{\mu}}_{\neg k}^{(t)}$ and $\check{\boldsymbol{\mathcal{T}}}_{\neg k}^{(t)}$. By re-arrangement, we can get the variational density for $\boldsymbol{\alpha}_k$.

$$\tilde{q}(\boldsymbol{\alpha}_k) \sim N(\tilde{\boldsymbol{\mu}}_{\alpha,k}^{(t)}, \tilde{\boldsymbol{\Lambda}}_{\alpha,k}^{(t)})$$

$$\tilde{\boldsymbol{\Lambda}}_{\alpha,k}^{(t)} = \left(\boldsymbol{Z}^T \tilde{\boldsymbol{\Omega}}^{(t)} \boldsymbol{Z} + [\check{\boldsymbol{\mathcal{T}}}_{\neg k}^{(t)}]^{-1}\right)^{-1}$$

$$\tilde{\boldsymbol{\mu}}_{\alpha,k}^{(t)} = \tilde{\boldsymbol{\Lambda}}_{\alpha,k}^{(t)} \left[[\check{\boldsymbol{\mathcal{T}}}_{\neg k}^{(t)}]^{-1} \check{\boldsymbol{\mu}}_{\neg k}^{(t)} + \boldsymbol{Z}^T \left(\boldsymbol{s}_k - \tilde{\boldsymbol{\Omega}}^{(t)} \boldsymbol{X} \boldsymbol{\beta}_k^{(t-1)} - \tilde{\boldsymbol{\Omega}}^{(t)} \tilde{\boldsymbol{m}}_k^{(t-1)}\right)\right]$$

$$[\tilde{\boldsymbol{m}}^{(t-1)}]_i = \tilde{m}_{i,k}^{(t-1)}$$

(c) M-Step: Using these variational E-distributions, update $\boldsymbol{\beta}_k$:

$$\boldsymbol{\beta}^{(t)} = \left(\boldsymbol{X}^T \tilde{\boldsymbol{\Omega}}^{(t)} \boldsymbol{X} + \boldsymbol{\Lambda}_\beta^0\right)^{-1} \boldsymbol{X}^T \left(\boldsymbol{s}_k - \tilde{\boldsymbol{\Omega}}^{(t)} \boldsymbol{Z} \boldsymbol{\mu}_{\alpha,k}^{(t)} - \tilde{\boldsymbol{\Omega}}^{(t)} \tilde{\boldsymbol{m}}_k^{(t-1)}\right)$$

$$[\tilde{\boldsymbol{m}}^{(t-1)}]_i = \tilde{m}_{i,k}^{(t-1)}$$

2. The above procedure performs a variational update for all $\boldsymbol{\beta}_k$. We have also updated our variational approximations to $\boldsymbol{\alpha}_k$ for all k. The final update must be for $\boldsymbol{\mathcal{S}}_{j,\alpha}$. From the Gibbs Sampler, the actual quantity we would want to calculate is:

$$\boldsymbol{\mathcal{S}}_{j,\alpha}^{(t)} = \frac{1}{G_j + \nu_j + S_j(K-1) + 1} \left(\boldsymbol{\Psi}_j + \sum_{g=1}^{G_j} E[\boldsymbol{\alpha}_{j,g} \boldsymbol{\alpha}_{j,g}^T]\right)$$

Appealing to our earlier strategy, we can do this using the variational distributions on $q(\boldsymbol{\alpha}_k)$. Note that we have assumed that the random effect components across levels are independent in the variational distributions and thus we can break down the expectation into: (a) the outer product of the (stacked) variational means plus (b) a block-diagonal matrix where the variance of each variational distribution forms the blocks. Crucially, this does not mean that our estimated $\boldsymbol{\mathcal{S}}^(t)_{j,a}$ will be diagonal—this is good. Rather,

it means that our variational assumptions imply that we will likely under-estimate the correlation between the random effects across levels.

3. Evaluate the ELBO. This is more complicated given the alternating structure of the variational distributions on the Polya-Gammas. I rely on the following approximate metric to assess convergence (in addition to examining the change in the coefficients directly). Note that given the variational distribution on $\boldsymbol{\alpha}_k$, we can lower-bound the likelihood given the random effects—but not the Polya-Gammas—using a Taylor expansion. Specifically, for one observation i:

$$E_{\{\boldsymbol{\alpha}_k\}}\left[\sum_{k=1}^{K} I(y_i = k)\left[\boldsymbol{x}_i^T\boldsymbol{\beta}_k + \boldsymbol{z}_i^T\boldsymbol{\alpha}_k\right] - \ln\left(\sum_{l=1}^{K}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}_l + \boldsymbol{z}_i^T\boldsymbol{\alpha}_l)\right)\right]$$

The latter term can be bounded by a first-order Taylor expansion and noting that we are relying on the variational independence assumption between $\{\boldsymbol{\alpha}_k\}$:

$$-E_{\{\boldsymbol{\alpha}_k\}}\left[\ln\left(\sum_{l=1}^{K}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}_l + \boldsymbol{z}_i^T\boldsymbol{\alpha}_l)\right)\right] \geq -\left[\ln\left(\sum_{l=1}^{K} E_{\boldsymbol{\alpha}_k}\left[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}_l + \boldsymbol{z}_i^T\boldsymbol{\alpha}_l)\right]\right)\right]$$

The right hand side can be directly evaluated by using the fat that $\boldsymbol{\alpha}_l$ is assumed to be normal and thus the expectation is that of a log-normal random variable.This bound could be improved further by relying on a higher-order Taylor expansion; examining the quadratic case, e.g. Teh, Newman, and Welling (2007), for an application in a different variational context and picking a better expansion is an interesting area for future exploration.

B.5 Over-Parameterized Models

One way to speed convergence of the models is to over-parameterize the random effects; multiple ways of doing this exist. I follow a simple strategy suggested by Gelman and Hill (2006, ch. 19) of using a mean-over parameterization. Formally,

$$\boldsymbol{\alpha}_{j,g} \sim^{i.i.d.} N\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_{j,\alpha}\right)$$

Thus, I add $\boldsymbol{\mu}_j$ to the M-Step of the algorithm. The modification of the variational distribution is slight and the M-Step is simply the mean of the random effects, i.e. $[\boldsymbol{\mu}_j]^{(t)} = \frac{\sum_g \boldsymbol{\alpha}_{j,g}}{G_j}$. Whenever $\boldsymbol{\alpha}_{j,g}$ appear in the original algorithms, those should be adjusted by subtracting off the estimated mean. Otherwise, the algorithms are very similar and thus not derived in detail here.

As this method is un-identified, however, it is necessary to adjust the coefficients at the end. This is done by adding each element of $\boldsymbol{\mu}_j$ to the corresponding element of $\boldsymbol{\beta}$. This means that that estimated random effects will be mean-zero by construction, *a posteriori* by construction.

Both methods converge to the same coefficients on $\boldsymbol{\beta}$ and $\{\boldsymbol{\Sigma}_{j,\alpha}\}$; however, in practice, the over-parameterized model converges much more quickly as the mean of the random effects only slowly converges towards zero for the non-over-parameterized case.

C Variational Approximations of Standard Errors

As discussed in the main text, I rely on Louis (1982)'s method for calculating (approximate) standard errors.[34] This method requires three quantities of interest: (a) the score, (b) the Hessian, and (c) the outer product of the score from the complete (augmented) data log-likelihood. The expectation of these quantities is taken over the distribution found in the E-Step or, for the purposes of the approximate method, its variational approximating distribution.

- Logistic: Recall the complete data likelihood function if $\boldsymbol{\alpha}, \{\omega_i\}$ are assumed known:

$$\ln p(\boldsymbol{y}, \boldsymbol{\alpha}, \{\omega_i\} | \boldsymbol{\beta}, \{\boldsymbol{\Sigma}_{j,\alpha}\}) = \sum_i -\ln(2) + s_i(\boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{\alpha}) - \omega_i \left(\boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{\alpha}\right)^2/2 + \ln f(\omega_i|1,0) + \sum_j -G_j/2 \ln(|2\pi\boldsymbol{\Sigma}_{j,\alpha}|) - \sum_j \sum_g 1/2 \text{tr}\left(\boldsymbol{\Sigma}_{j,\alpha}^{-1}\left[\boldsymbol{\alpha}_{j,g}\boldsymbol{\alpha}_{j,g}^T\right]\right) \tag{42}$$

  The score with respect to the parameters ($\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_{j,a}$) can be expressed as follows. For notation, I use vech $(\boldsymbol{\Sigma}_{j,\alpha})$ to vectorize the lower half of the matrix; this therefore collects all the unique elements and contains $p(p+1)/2$ elements where $\boldsymbol{\Sigma}_{j,\alpha}$ is $p \times p$. $s(\text{vech}\,(\boldsymbol{\Sigma}_{j,\alpha})_i)$ denotes the score for the i-th element of vech $((\boldsymbol{\Sigma}_{j,\alpha}))$.

  Define $\boldsymbol{V}_i$ as a matrix that corresponds to all zeros except for the positions that the i-th element occupies in $\boldsymbol{\Sigma}_{j,\alpha}$ that gets a value of '1'. Alternatively put, it is the element-by-element derivative of $\boldsymbol{\Sigma}_{j,\alpha}$ w.r.t. the i-th element. Many results here use identities about matrix differentiation found in (Mardia and Marshall 1984; Petersen and Pedersen 2012; Brookes 2011).

$$s(\boldsymbol{\beta}) = \boldsymbol{X}^T\left(\boldsymbol{s} - \boldsymbol{\Omega}\left[\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha}\right]\right) \tag{43a}$$

$$s(\text{vech}\,(\boldsymbol{\Sigma}_{j,\alpha})_i) = \frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}_{j,\alpha}^{-1}\boldsymbol{V}_i\left[-G_j\mathbf{I} + \boldsymbol{\Sigma}_{j,\alpha}^{-1}\left[\sum_g \boldsymbol{\alpha}_{j,g}\boldsymbol{\alpha}_{j,g}^T\right]\right]\right) \tag{43b}$$

  The Hessian is block-diagonal with the diagonal blocks as follows. Each element of the Hessian of the score can be expressed as follows (Mardia and Marshall 1984):

$$H(\boldsymbol{\beta}) = -\boldsymbol{X}^T\boldsymbol{\Omega}\boldsymbol{X} \tag{44a}$$

$$H(\text{vech}\,(\boldsymbol{\Sigma}_{j,\alpha})_i, \text{vech}\,(\boldsymbol{\Sigma}_{j,\alpha})_k) = -\frac{1}{2}\text{tr}\left(-G_j \cdot \boldsymbol{\Sigma}_{j,\alpha}^{-1}\boldsymbol{V}_i\boldsymbol{\Sigma}_{j,\alpha}^{-1}\boldsymbol{V}_k + \boldsymbol{\Sigma}_{j,\alpha}^{-1}\left(\boldsymbol{V}_i\boldsymbol{\Sigma}_{j,\alpha}^{-1}\boldsymbol{V}_k + \boldsymbol{V}_k\boldsymbol{\Sigma}_{j,\alpha}^{-1}\boldsymbol{V}_i\right)\boldsymbol{\Sigma}_{j,\alpha}^{-1}\left[\sum_g \boldsymbol{\alpha}_{j,g}\boldsymbol{\alpha}_{j,}^T\right.\right. \tag{44b}$$

  Note that $\partial \boldsymbol{s}(\boldsymbol{\beta})/\partial \text{vech}\,(\boldsymbol{\Sigma}_{j,\alpha}) = \mathbf{0}$ and thus the off diagonal blocks of the Hessian are zero. Further, if there are multiple random effects (i.e. $\boldsymbol{\Sigma}_{j,\alpha}$ and $\boldsymbol{\Sigma}_{j',\alpha}$), these also have zero blocks in the Hessian.

[34] Other methods exist, e.g. Oakes (1999), that could also be explored for calculating approximate variational errors. These expectations could be calculated via a Monte Carlo approach either using the variational distribution or, more sensibly, using the Gibbs Sampler to sample the correct joint distribution of the latent variables given the observed data and other parameters.

- Negative Binomial: The complete data log-likelihood can be expressed as follows.

$$\ln p(\boldsymbol{y}, \boldsymbol{\alpha}|r, \boldsymbol{\beta}, \{\boldsymbol{\Sigma}_{j,\alpha}\}) = \sum_i \ln\Gamma(y_i + r) - \ln\Gamma(r) - \ln\Gamma(y_i + 1) \tag{45}$$

$$-\ln(2)(y_i + r) + (y_i - r)/2 \cdot (\boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{\alpha} - \ln r) - \omega_i \left(\boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{\alpha} - \ln r\right)^2/2 + \ln f(\omega_i|y_i + r, 0)$$

$$\sum_j -G_j/2\ln(|2\pi\boldsymbol{\Sigma}_{j,\alpha}|) - \sum_j\sum_g 1/2 \cdot \text{tr}\left(\boldsymbol{\Sigma}_{j,\alpha}^{-1}\left[\boldsymbol{\alpha}_{j,g}\boldsymbol{\alpha}_{j,g}^T\right]\right)$$

The score vector can be derived as before, noting that it now includes r as a parameter. For this section, I define $\boldsymbol{s}$ such that $s_i = (y_i - r)/2$. I further slightly abuse notation to use $\ln \boldsymbol{r}$ to be a vector of length $\boldsymbol{s}$ for each element consisting of $\ln r$.

$$s(\boldsymbol{\beta}) = \boldsymbol{X}^T\left(\boldsymbol{s} - \boldsymbol{\Omega}\left[\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha} - \ln \boldsymbol{r}\right]\right) \tag{46a}$$

$$s(\text{vech}\left(\boldsymbol{\Sigma}_{j,\alpha}\right)_i) = \frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}_{j,\alpha}^{-1}\boldsymbol{V}_i\left[-G_j\mathbf{I} + \boldsymbol{\Sigma}_{j,\alpha}^{-1}\left[\sum_g \boldsymbol{\alpha}_{j,g}\boldsymbol{\alpha}_{j,g}^T\right]\right]\right) \tag{46b}$$

$$s(r) = \sum_i \psi(y_i + r) - \psi(r) - \frac{(y_i - r)}{2r} - \frac{(\boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{\alpha} - \ln r)}{2} + \tag{46c}$$

$$\frac{\omega_i(\boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{\alpha} - \ln r)}{r} + \frac{\partial \ln f(\omega_i|y_i + r, 0)}{\partial r}$$

The Hessian is similar to the logistic case:

$$H(\boldsymbol{\beta}) = -\boldsymbol{X}^T\boldsymbol{\Omega}\boldsymbol{X} \tag{47a}$$

$$H(\text{vech}\left(\boldsymbol{\Sigma}_{j,\alpha}\right)_i, \text{vech}\left(\boldsymbol{\Sigma}_{j,\alpha}\right)_k) =$$
$$-\frac{1}{2}\text{tr}\left(-G_j \cdot \boldsymbol{\Sigma}_{j,\alpha}^{-1}\boldsymbol{V}_i\boldsymbol{\Sigma}_{j,\alpha}^{-1}\boldsymbol{V}_k + \boldsymbol{\Sigma}_{j,\alpha}^{-1}\left(\boldsymbol{V}_i\boldsymbol{\Sigma}_{j,\alpha}^{-1}\boldsymbol{V}_k + \boldsymbol{V}_k\boldsymbol{\Sigma}_{j,\alpha}^{-1}\boldsymbol{V}_i\right)\boldsymbol{\Sigma}_{j,\alpha}^{-1}\left[\sum_g \boldsymbol{\alpha}_{j,g}\boldsymbol{\alpha}_{j,g}^T\right]\right) \tag{47b}$$

$$H(r) = \sum_i \psi'(y_i + r) - \psi'(r) + \frac{y_i}{2r^2} + \frac{1}{2r} + \frac{\omega_i\left[1 + \boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{\alpha} - \ln r\right]}{r^2} + \frac{\partial^2 \ln f(\omega_i|y_i + r, 0)}{\partial r^2} \tag{47c}$$

$$\frac{\partial s(\boldsymbol{\beta})}{\partial r} = \sum_i \boldsymbol{x}_i\left[-1/2 + \omega_i/r\right] \tag{47d}$$

The key complication for the negative binomial model comes from the derivative of the Polya-Gamma density that, itself, depends on r. As noted in the derivation of the algorithm, this is intractable. A similar solution to above cannot be applied and thus I rely on a functional approximation. Specifically, note that the Polya-Gamma is an infinite convolution of Gamma random variables and thus has a "Gamma-like" shape. Further note that as b grows (i.e. $y_i + r$), it approaches a normal distribution (Glynn et al. 2018). Thus, for evaluating the terms that involve the log density of a Polya-Gamma and its derivative, I approximate it with a Gamma random variable such that the approximating Gamma has the same mean

and variance as the Polya-Gamma, i.e. a mean of $(y_i + r)/4$ and a variance of $(y_i + r)/24$.[35] Specifically, that means that it is approximated by $Gamma(3/2[y_i + r], 6)$. This density and its derivatives w.r.t. r are shown below

$$\ln f(\omega_i|y_i + r, 0) \approx 3/2[y_i + r]\ln(6) - \ln\Gamma(3/2[y_i + r]) + 6\omega_i + (3/2[y_i + r] - 1)\ln(\omega_i) \tag{48a}$$

$$\frac{\partial \ln f(\omega_i|y_i + r, 0)}{\partial r} \approx 3/2\left[\ln(6) - \psi(3/2[y_i + r]) + \ln(\omega_i)\right] \tag{48b}$$

$$\frac{\partial^2 \ln f(\omega_I|y_i + r, 0)}{\partial r^2} \approx -9/4\psi(3/2[y_i + r]) \tag{48c}$$

This can be used in the above analysis with the exception of $E[\ln(\omega_i)]$ that is, itself, intractable. In this case, I again approximate ω_i as a Gamma random variable noting that if $X \sim Gamma(\alpha_0, \beta_0)$, then $E[\ln(X)] = \psi(\alpha_0) - \ln(\beta_0)$.

- Over-Parameterization: Over-parameterizing the mean, that helps convergence in the algorithm, can also be used for calculating the standard errors of the over-parameterized model, although a reduced variance-covariance matrix is required for inference on the (identified) $\boldsymbol{\beta}$. This requires using the score and Hessian for $\boldsymbol{\mu}$. It can be defined as follows

$$s(\boldsymbol{\mu}_j) = \boldsymbol{\Sigma}_{j,\alpha}^{-1}\left(\sum_g \boldsymbol{\alpha}_{j,g} - G_j\boldsymbol{\mu}_j\right) \tag{49a}$$

$$H(\boldsymbol{\mu}_j) = -\boldsymbol{\Sigma}_{j,\alpha}^{-1} \cdot G_j \tag{49b}$$

$$\frac{\partial \boldsymbol{\mu}_j}{\partial\left[\text{vech}\left(\boldsymbol{\Sigma}_{j,\alpha}\right)\right]_i} = -\boldsymbol{\Sigma}_{j,\alpha}^{-1}\boldsymbol{V}_i \cdot s(\boldsymbol{\mu}_j) \tag{49c}$$

Given the results above, one can calculate all of the quantities of interest. Detailed analytical on the expectation of the outer product of the score are available upon request and heavily draw on identities from the Matrix Cookbook (Petersen and Pedersen 2012) and the Matrix Reference Manual (Brookes 2011).

[35] Using the definition of a Polya-Gamma as an infinite convolution, its variance can be derived as follows. For $c \to 0$, this converges to $b/24$.

$$Var(\omega) = \frac{b}{\cosh(c/2)^2} \cdot \frac{\sinh(c) - c}{4c^3}; \quad \omega_i \sim PG(b, c)$$

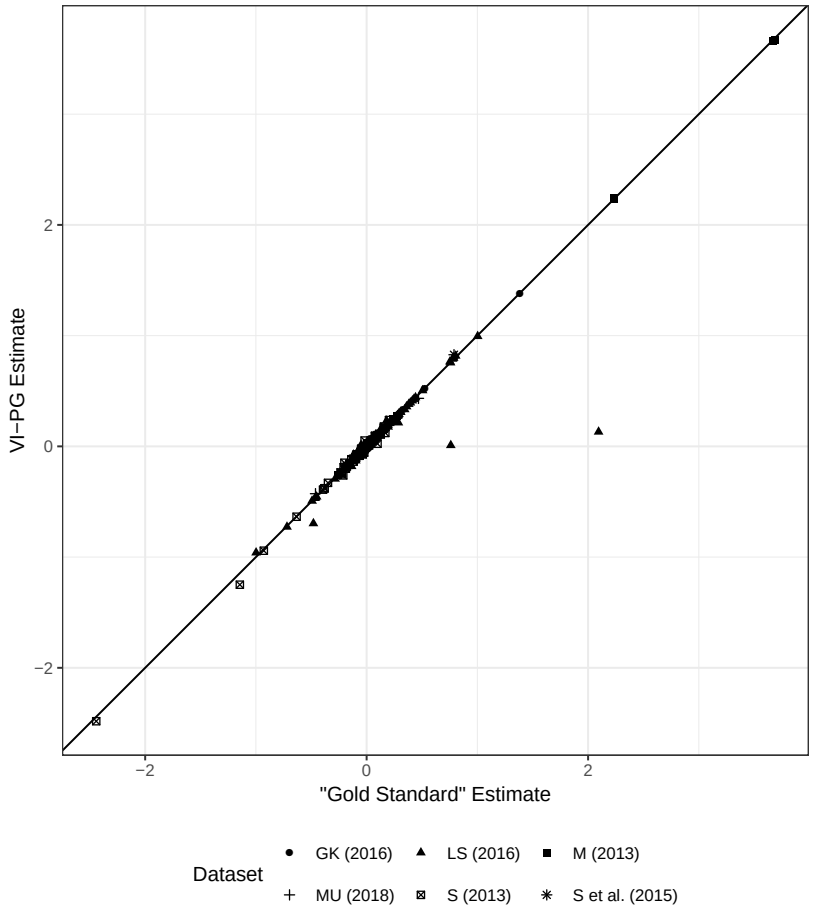
D Additional Information on Replicated Models

This section provides more detailed information on the performance of the variational algorithm on the six papers re-analyzed in Table 6. First, I report the exact models replicated.

- Stoeckel (2013): Table 3
- Legewie and Schaeffer (2016): Models 1-5 from Table 2 and Models 1,2, and 4 from Table 3
- Micozzi (2013): Models 3 and 4 from Table 1; Models 7 and 8 from Table 2
- Schumacher et al. (2015): Model 1 from Table 4
- Giger and Klüver (2016): Model 1 from Table 3
- Michelitch and Utych (2018): Model 1 from Table 1

Next, it shows the results of the analysis visually by plotting the estimated coefficients (standardized to have variance one) from the gold standard (`lmer`) and the variational methods. It plots the standardized coefficients against one another. With the exception of a very small number of coefficients, they are clearly tightly correlated and nearly exactly equal.

Figure 3: Visually Comparing Coefficients



E Additional Information on EU Public Opinion

This Appendix reports additional information on the models replicating (Stoeckel 2013).

E.1 Controls

The controls included in all models, following Stoeckel (2013) exactly, are briefly summarized here. First, there are variables for net fiscal transfer and membership length. Political orientations, occupation, age, gender, economic prospects, an index of posessions, the five major variables discussed in the main text (trust, attachment, elite division, median consumption, and political knowledge). Again following the original paper, these are all hierarchically centered, i.e. for all variables that vary within a country, it is standardized to be mean zero variance one *within a country*. Thus, the interpretation of a coefficient would be the change given a one standard deviation change of a respondent *relative to their country's baseline mean*.

E.2 Posterior Diagnostics

As noted in the main text, I conducted Gelman-Rubin and Geweke tests to examine whether the models have converged. Using the standard of a Gelman-Rubin statistic below 1.1, almost all models pass this test. All fixed and random effects in every model except for Model 9 and Model 14 pass this test. In Model 9 and Model 14, one coefficient (out of 81) fail to pass this test.

Using another test, I examined for each model, for each chain, does a variable have a z-score attached to the Geweke diagnostic that is below 1.96 in absolute value. The results here are more mixed, but 70% of model-chain-coefficient pairs are below this threshold alongside 60% of random effects.

F Additional Information on NYC 311 Calls

This Appendix reports additional information on the models in Legewie and Schaeffer 2016. The ten main models outlined are as follows. For simplicity, following the replication code from the authors, we can divide the variables into blocks. See Legewie and Schaeffer (2016, p.136) for a detailed exposition of the variables.

- Base Controls: Population in Census Block, Physical Area of Census Block, Calls to 311 on Other Topics, Borough Name
- Additional Controls: Concentrated Disadvantage, Crime Prone Population, Residential Instability, Immigrant Concentration, Number of Foreclosures, Public Housing Rate, Multigroup Segregation
- Diversity Controls: Ethnic Polarization, Ethnic Diversity
- Racial Proportion Controls: Proportion Black, Hispanic, and Asian (three variables)

The models I estimated to compare against their original results were as follows. All models include edge intensity, its square.

1. Base Controls with Random Effect for Census Tract
2. (1) + Additional Controls

3. (2) + Diversity Controls
4. (3) + Racial Proportion Controls
5. Adjusts the outcome and controls based on discussion on page 145 of the manuscript.
6. Model (1) with a calls about noise.
7. Model (1) focusing on calls at night. See their supporting information for more details.
8. Model (1) focusing on calls about other issues at night. See their supporting information for more details.

I modify Model 2 to add random effects by census tract as discussed in the main text.