Fast and Accurate Estimation of Non-Nested Binomial Hierarchical Models Using Variational Inference

Max Goplerud

University of Pittsburgh

2020 Annual Meeting of the Society for Political Methodology
(PolMeth XXXVII)

Political Scientists Study Complicated Data

Political Scientists Study Complicated Data

- Data in political science is messy

Political Scientists Study Complicated Data

- Data in political science is messy
  - Correlated across observations (voters within constituencies)

Political Scientists Study Complicated Data

- Data in political science is messy
  - Correlated across observations (voters within constituencies)
  - Nested observations (respondents in clusters in countries)

Political Scientists Study Complicated Data

- Data in political science is messy
  - Correlated across observations (voters within constituencies)
  - Nested observations (respondents in clusters in countries)
  - Effects vary across space and time (effect of income over time)

Political Scientists Study Complicated Data

- Data in political science is messy
  - Correlated across observations (voters within constituencies)
  - Nested observations (respondents in clusters in countries)
  - Effects vary across space and time (effect of income over time)
  - Non-linear outcomes (binary, count, multinomial)

Political Scientists Study Complicated Data

- Data in political science is messy
  - Correlated across observations (voters within constituencies)
  - Nested observations (respondents in clusters in countries)
  - Effects vary across space and time (effect of income over time)
  - Non-linear outcomes (binary, count, multinomial)

- Standard models ("i.i.d."; linear outcomes) are often unsuitable

Political Scientists Study Complicated Data

- Data in political science is messy
  - Correlated across observations (voters within constituencies)
  - Nested observations (respondents in clusters in countries)
  - Effects vary across space and time (effect of income over time)
  - Non-linear outcomes (binary, count, multinomial)
- Standard models ("i.i.d."; linear outcomes) are often unsuitable
- What to do?

Political Scientists Study Complicated Data

- Data in political science is messy
  - Correlated across observations (voters within constituencies)
  - Nested observations (respondents in clusters in countries)
  - Effects vary across space and time (effect of income over time)
  - Non-linear outcomes (binary, count, multinomial)
- Standard models ("i.i.d."; linear outcomes) are often unsuitable
- What to do? Hierarchical models, random effects, mixed effects, multilevel models, ...

Political Scientists Study Complicated Data

- Data in political science is messy
  - Correlated across observations (voters within constituencies)
  - Nested observations (respondents in clusters in countries)
  - Effects vary across space and time (effect of income over time)
  - Non-linear outcomes (binary, count, multinomial)
- Standard models ("i.i.d."; linear outcomes) are often unsuitable
- What to do? Hierarchical models, random effects, mixed effects, multilevel models, ...
- Popular in political science and use is going ↑↑

But...There's a Problem

But…There's a Problem

- Inference is *tough*:

But...There's a Problem

- Inference is *tough*:
  - Often requires evaluating many, intractable, integrals

But...There's a Problem

- Inference is *tough*:
  - Often requires evaluating many, intractable, integrals
  - Even worse when effects are "non-nested" (e.g. time + country)

But…There's a Problem

- Inference is *tough*:
  - Often requires evaluating many, intractable, integrals
  - Even worse when effects are "non-nested" (e.g. time + country)
- Estimation is thus usually rather *slow*

But...There's a Problem

- Inference is *tough*:
  - Often requires evaluating many, intractable, integrals
  - Even worse when effects are "non-nested" (e.g. time + country)
- Estimation is thus usually rather *slow*
  - Usually need to fit many models for hypothesis testing, robustness tests, model comparison, cross-validation...

But...There's a Problem

- Inference is *tough*:
  - Often requires evaluating many, intractable, integrals
  - Even worse when effects are "non-nested" (e.g. time + country)
- Estimation is thus usually rather *slow*
  - Usually need to fit many models for hypothesis testing, robustness tests, model comparison, cross-validation...
- For applied researchers, hierarchical models can be a pain to use.

But…There's a Problem

- Inference is *tough*:
  - Often requires evaluating many, intractable, integrals
  - Even worse when effects are "non-nested" (e.g. time + country)
- Estimation is thus usually rather *slow*
  - Usually need to fit many models for hypothesis testing, robustness tests, model comparison, cross-validation...
- For applied researchers, hierarchical models can be a pain to use.
- **Motivation:** Can we estimate these models *differently*, gain speed, and maintain accuracy?

Methods for Estimating Hierarchical Models

Methods for Estimating Hierarchical Models

	Bayesian	Laplace Approximation	Variational Bayes

Methods for Estimating Hierarchical Models

	Bayesian	Laplace Approximation	Variational Bayes
Software	`STAN`	`glmer`	...

Methods for Estimating Hierarchical Models

	Bayesian	Laplace Approximation	Variational Bayes
Software	`STAN`	`glmer`	...
Speed			
Accuracy			
Quantifying Uncertainty			

Methods for Estimating Hierarchical Models

	Bayesian	Laplace Approximation	Variational Bayes
Software	`STAN`	`glmer`	...
Speed	—		
Accuracy	++		
Quantifying Uncertainty	++		

Methods for Estimating Hierarchical Models

	Bayesian	Laplace Approximation	Variational Bayes
Software	`STAN`	`glmer`	...
Speed	—	?	
Accuracy	++	+	
Quantifying Uncertainty	++	?	

Methods for Estimating Hierarchical Models

	Bayesian	Laplace Approximation	Variational Bayes
Software	`STAN`	`glmer`	...
Speed	—	?	++
Accuracy	++	+	—
Quantifying Uncertainty	++	?	—

Methods for Estimating Hierarchical Models

	Bayesian	Laplace Approximation	Variational Bayes
Software	`STAN`	`glmer`	...
Speed	—	?	++
Accuracy	++	+	—
Quantifying Uncertainty	++	?	—

- **Goal for Today**: Keep speed but maintain quality

Methods for Estimating Hierarchical Models

	Bayesian	Laplace Approximation	Variational Bayes	MAVB
Software	`STAN`	`glmer`	...	`vglmer`
Speed	—	?	++	
Accuracy	++	+	—	
Quantifying Uncertainty	++	?	—	

- **Goal for Today**: Keep speed but maintain quality
- Marginally Augmented Variational Bayes

Methods for Estimating Hierarchical Models

	Bayesian	Laplace Approximation	Variational Bayes	MA**VB**
Software	`STAN`	`glmer`	...	`vglmer`
Speed	—	?	++	++
Accuracy	++	+	—	+
Quantifying Uncertainty	++	?	—	—

- **Goal for Today**: Keep speed but maintain quality
- Marginally Augmented **Variational Bayes**
  - Variational Bayes: New application of data augmentation to (non-linear) hierarchical models (Polson, Scott, and Windle 2013)

Methods for Estimating Hierarchical Models

	Bayesian	Laplace Approximation	Variational Bayes	**MA**VB
Software	`STAN`	`glmer`	...	`vglmer`
Speed	—	?	++	++
Accuracy	++	+	—	+
Quantifying Uncertainty	++	?	—	+

- **Goal for Today**: Keep speed but maintain quality
- **Marginally Augmented** Variational Bayes
  - Variational Bayes: New application of data augmentation to (non-linear) hierarchical models (Polson, Scott, and Windle 2013)
  - **Marginally Augmented:** Post-processing step to improve uncertainty

Methods for Estimating Hierarchical Models

	Bayesian	Laplace Approximation	Variational Bayes	MAVB
Software	`STAN`	`glmer`	...	`vglmer`
Speed	—	?	++	++
Accuracy	++	+	—	+
Quantifying Uncertainty	++	?	—	+

- **Goal for Today**: Keep speed but maintain quality
- Marginally Augmented Variational Bayes
  - Variational Bayes: New application of data augmentation to (non-linear) hierarchical models (Polson, Scott, and Windle 2013)
  - **Marginally Augmented:** Post-processing step to improve uncertainty
- Focus on logistic hierarchical models in paper
  - `R` package includes count and (soon!) multinomial and linear

Overview of Presentation

Overview of Presentation

- Motivating Example: Deep MRP (Ghitza and Gelman 2013)
- Outlining MAVB
- Advice for MRP Practitioners: How Deep is Deep Enough?

Motivating Example: Ghitza and Gelman (2013)

- Explain turnout differentials by state/age/ethnicity/income

Motivating Example: Ghitza and Gelman (2013)

- Explain turnout differentials by state/age/ethnicity/income
  - **But**: Only a few observations per cell $\rightarrow$ MRP!

Motivating Example: Ghitza and Gelman (2013)

- Explain turnout differentials by state/age/ethnicity/income
  - **But**: Only a few observations per cell $\rightarrow$ MRP!
  - Fit a **multilevel regression** on the survey and post-stratify

Motivating Example: Ghitza and Gelman (2013)

- Explain turnout differentials by state/age/ethnicity/income
  - **But**: Only a few observations per cell → MRP!
  - Fit a **multilevel regression** on the survey and post-stratify
  - Key contribution: Add "deep" interactions

Motivating Example: Ghitza and Gelman (2013)

- Explain turnout differentials by state/age/ethnicity/income
  - **But**: Only a few observations per cell $\rightarrow$ MRP!
  - Fit a **multilevel regression** on the survey and post-stratify
  - Key contribution: Add "deep" interactions
- Preferred model has 18 random effects and nearly 4,000 parameters!

Motivating Example: Ghitza and Gelman (2013)

- Explain turnout differentials by state/age/ethnicity/income
  - **But**: Only a few observations per cell → MRP!
  - Fit a **multilevel regression** on the survey and post-stratify
  - Key contribution: Add “deep” interactions
- Preferred model has 18 random effects and nearly 4,000 parameters!
  - Theory: Why use 18? Why not 4?

Motivating Example: Ghitza and Gelman (2013)

- Explain turnout differentials by state/age/ethnicity/income
  - **But**: Only a few observations per cell → MRP!
  - Fit a **multilevel regression** on the survey and post-stratify
  - Key contribution: Add "deep" interactions
- Preferred model has 18 random effects and nearly 4,000 parameters!
  - Theory: Why use 18? Why not 8 or 12?

Motivating Example: Ghitza and Gelman (2013)

- Explain turnout differentials by state/age/ethnicity/income
  - **But**: Only a few observations per cell → MRP!
  - Fit a **multilevel regression** on the survey and post-stratify
  - Key contribution: Add "deep" interactions
- Preferred model has 18 random effects and nearly 4,000 parameters!
  - Theory: Why use 18? Why not 8 or 12? Overfitting?

Motivating Example: Ghitza and Gelman (2013)

- Explain turnout differentials by state/age/ethnicity/income
  - **But**: Only a few observations per cell → MRP!
  - Fit a **multilevel regression** on the survey and post-stratify
  - Key contribution: Add "deep" interactions
- Preferred model has 18 random effects and nearly 4,000 parameters!
  - Theory: Why use 18? Why not 8 or 12? Overfitting?
  - Computation: Expensive to fit the "deep" model (prohibitive for CV)

Motivating Example: Ghitza and Gelman (2013)

- Explain turnout differentials by state/age/ethnicity/income
  - **But**: Only a few observations per cell → MRP!
  - Fit a **multilevel regression** on the survey and post-stratify
  - Key contribution: Add "deep" interactions
- Preferred model has 18 random effects and nearly 4,000 parameters!
  - Theory: Why use 18? Why not 8 or 12? Overfitting?
  - Computation: Expensive to fit the "deep" model (prohibitive for CV)
- Consider a spectrum of nine models:

Motivating Example: Ghitza and Gelman (2013)

- Explain turnout differentials by state/age/ethnicity/income
  - **But**: Only a few observations per cell → MRP!
  - Fit a **multilevel regression** on the survey and post-stratify
  - Key contribution: Add "deep" interactions
- Preferred model has 18 random effects and nearly 4,000 parameters!
  - Theory: Why use 18? Why not 8 or 12? Overfitting?
  - Computation: Expensive to fit the "deep" model (prohibitive for CV)
- Consider a spectrum of nine models:
  - Simple: `... + (1 | state) + (1|eth) + (1|age) + (1|inc)`
  - Deep:
```
(1 | inc) + (1 + z.inc | eth) + (1 + z.inc | stt) + (1 + z.inc | age) +
+ (1 | eth.inc) + (1 | eth.age) + (1 | inc.age) + (1 | stt.eth) + (1 | stt.inc) + (1 | stt.age) +
(1 + z.inc | reg) + (1 | reg.eth) + (1 | reg.inc) + (1 | reg.age) + (1 | eth.inc.age) +
(1 | stt.eth.inc) + (1 | stt.eth.age) + (1 | stt.inc.age)
```

Motivating Example: Ghitza and Gelman (2013)

- Explain turnout differentials by state/age/ethnicity/income
  - **But**: Only a few observations per cell → MRP!
  - Fit a **multilevel regression** on the survey and post-stratify
  - Key contribution: Add "deep" interactions
- Preferred model has 18 random effects and nearly 4,000 parameters!
  - Theory: Why use 18? Why not 8 or 12? Overfitting?
  - Computation: Expensive to fit the "deep" model (prohibitive for CV)
- Consider a spectrum of nine models:
  - Simple: `... + (1 | state) + (1|eth) + (1|age) + (1|inc)`
  - Deep: `(1 | inc) + (1 + z.inc | eth) + (1 + z.inc | stt) + (1 + z.inc | age) +`
    `+ (1 | eth.inc) + (1 | eth.age) + (1 | inc.age) + (1 | stt.eth) + (1 | stt.inc) + (1 | stt.age) +`
    `(1 + z.inc | reg) + (1 | reg.eth) + (1 | reg.inc) + (1 | reg.age) + (1 | eth.inc.age) +`
    `(1 | stt.eth.inc) + (1 | stt.eth.age) + (1 | stt.inc.age)`
  - Intermediate: `(1 + z.inc | stt) + (1 + z.inc | eth) + (1 | inc)`
    `(1 + z.inc | age) + (1 | eth.inc) + (1 | eth.age) + (1 | inc.age)`
    `(1 | stt.eth) + (1 | stt.inc) + (1 | stt.age)`

Outlining MAVB

- VB - Variational Bayes
- MA - Marginal Augmentation

Variational Bayes (VB): Approximating the Posterior

Variational Bayes (VB): Approximating the Posterior

- Model: Logistic (Binomial) Random Effects
  - J random effects (e.g. age, county, gender) each with d_j variables
  - p "fixed effects"

$$y_i \sim \text{Binom}(n_i, p_i) \quad p_i = \frac{\exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta} + \sum_{j=1}^J \boldsymbol{z}_{i,j}^T \boldsymbol{\alpha}_{j,g[i]}\right)}{1 + \exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta} + \sum_{j=1}^J \boldsymbol{z}_{i,j}^T \boldsymbol{\alpha}_{j,g[i]}\right)} \qquad \begin{array}{c} \boldsymbol{\alpha}_{j,g} \sim^{i.i.d.} N(\boldsymbol{0}, \boldsymbol{\Sigma}_j) \\ \boldsymbol{\Sigma}_j \sim \text{IW}(\nu_j, \boldsymbol{\Phi}_j) \end{array}$$

Variational Bayes (VB): Approximating the Posterior

- Model: Logistic (Binomial) Random Effects
  - J random effects (e.g. age, county, gender) each with d_j variables
  - p "fixed effects"

$$y_i \sim \text{Binom}(n_i, p_i) \quad p_i = \frac{\exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta} + \sum_{j=1}^{J} \boldsymbol{z}_{i,j}^T \boldsymbol{\alpha}_{j,g[i]}\right)}{1 + \exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta} + \sum_{j=1}^{J} \boldsymbol{z}_{i,j}^T \boldsymbol{\alpha}_{j,g[i]}\right)} \qquad \begin{array}{c} \boldsymbol{\alpha}_{j,g} \sim^{i.i.d.} N(\mathbf{0}, \boldsymbol{\Sigma}_j) \\ \boldsymbol{\Sigma}_j \sim \text{IW}(\nu_j, \boldsymbol{\Phi}_j) \end{array}$$

- Goal: Approximate posterior of $\boldsymbol{\theta} = \boldsymbol{\beta}, \{\boldsymbol{\alpha}_j\}, \{\boldsymbol{\Sigma}_j\}$

Variational Bayes (VB): Approximating the Posterior

- Model: Logistic (Binomial) Random Effects
  - J random effects (e.g. age, county, gender) each with d_j variables
  - p "fixed effects"

$$y_i \sim \mathrm{Binom}(n_i, p_i) \quad p_i = \frac{\exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta} + \sum_{j=1}^J \boldsymbol{z}_{i,j}^T \boldsymbol{\alpha}_{j,g[i]}\right)}{1 + \exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta} + \sum_{j=1}^J \boldsymbol{z}_{i,j}^T \boldsymbol{\alpha}_{j,g[i]}\right)} \qquad \begin{array}{c} \boldsymbol{\alpha}_{j,g} \sim^{i.i.d.} N(\mathbf{0}, \boldsymbol{\Sigma}_j) \\ \boldsymbol{\Sigma}_j \sim \mathrm{IW}(\nu_j, \boldsymbol{\Phi}_j) \end{array}$$

- Goal: Approximate posterior of $\boldsymbol{\theta} = \boldsymbol{\beta}, \{\boldsymbol{\alpha}_j\}, \{\boldsymbol{\Sigma}_j\}$
- Mean-Field VB: Assume independence, $q(\boldsymbol{\beta})q(\{\boldsymbol{\alpha}_j\})q(\{\boldsymbol{\Sigma}_j\})$, and find best approximation to true posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$

Variational Bayes (VB): Approximating the Posterior

- Model: Logistic (Binomial) Random Effects
  - J random effects (e.g. age, county, gender) each with d_j variables
  - p "fixed effects"

$$y_i \sim \text{Binom}(n_i, p_i) \quad p_i = \frac{\exp\left(\boldsymbol{x}_i^T \beta + \sum_{j=1}^J \boldsymbol{z}_{i,j}^T \boldsymbol{\alpha}_{j,g[i]}\right)}{1 + \exp\left(\boldsymbol{x}_i^T \beta + \sum_{j=1}^J \boldsymbol{z}_{i,j}^T \boldsymbol{\alpha}_{j,g[i]}\right)} \qquad \begin{array}{c} \boldsymbol{\alpha}_{j,g} \sim^{i.i.d.} N(\mathbf{0}, \boldsymbol{\Sigma}_j) \\ \boldsymbol{\Sigma}_j \sim \text{IW}(\nu_j, \boldsymbol{\Phi}_j) \end{array}$$

- Goal: Approximate posterior of $\boldsymbol{\theta} = \beta, \{\boldsymbol{\alpha}_j\}, \{\boldsymbol{\Sigma}_j\}$
- Mean-Field VB: Assume independence, $q(\beta)q(\{\boldsymbol{\alpha}_j\})q(\{\boldsymbol{\Sigma}_j\})$, and find best approximation to true posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$
  - As posed, no specialized algorithm for arbitrary J (see $J = 2$ in Jeon, Rijmen, and Rabe-Hesketh 2017)
  - Requires evaluating many integrals

Variational Bayes (VB): Approximating the Posterior

- Model: Logistic (Binomial) Random Effects
  - J random effects (e.g. age, county, gender) each with d_j variables
  - p "fixed effects"

$$y_i \sim \mathrm{Binom}(n_i, p_i) \quad p_i = \frac{\exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta} + \sum_{j=1}^J \boldsymbol{z}_{i,j}^T \boldsymbol{\alpha}_{j,g[i]}\right)}{1 + \exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta} + \sum_{j=1}^J \boldsymbol{z}_{i,j}^T \boldsymbol{\alpha}_{j,g[i]}\right)} \qquad \begin{array}{c} \boldsymbol{\alpha}_{j,g} \sim^{i.i.d.} N(\mathbf{0}, \boldsymbol{\Sigma}_j) \\ \boldsymbol{\Sigma}_j \sim \mathrm{IW}(\nu_j, \boldsymbol{\Phi}_j) \end{array}$$

- Goal: Approximate posterior of $\boldsymbol{\theta} = \boldsymbol{\beta}, \{\boldsymbol{\alpha}_j\}, \{\boldsymbol{\Sigma}_j\}$
- Mean-Field VB: Assume independence, $q(\boldsymbol{\beta})q(\{\boldsymbol{\alpha}_j\})q(\{\boldsymbol{\Sigma}_j\})$, and find best approximation to true posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$
  - As posed, no specialized algorithm for arbitrary J (see $J = 2$ in Jeon, Rijmen, and Rabe-Hesketh 2017)
  - Requires evaluating many integrals
- Solution: Augment posterior using Polya-Gammas (Polson, Scott, and Windle 2013)

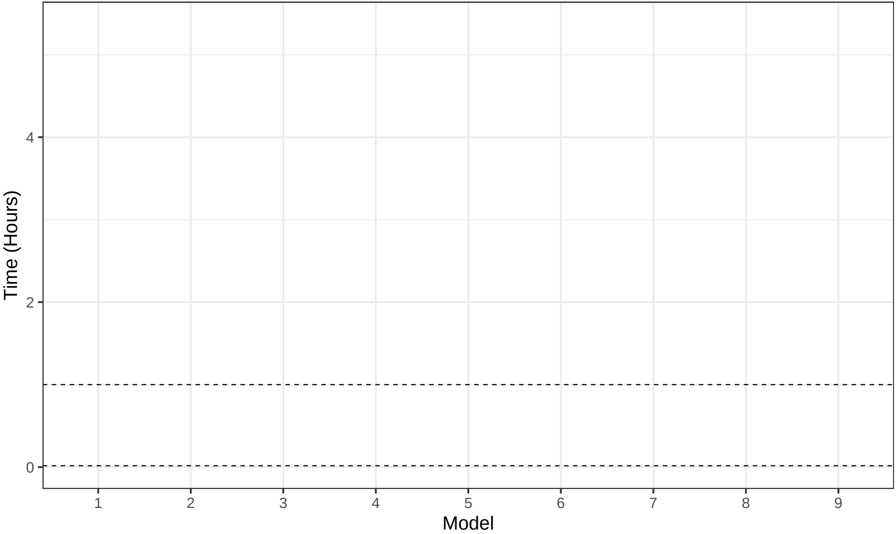
Variational Bayes (VB): Approximating the Posterior

- Model: Logistic (Binomial) Random Effects
  - J random effects (e.g. age, county, gender) each with d_j variables
  - p "fixed effects"

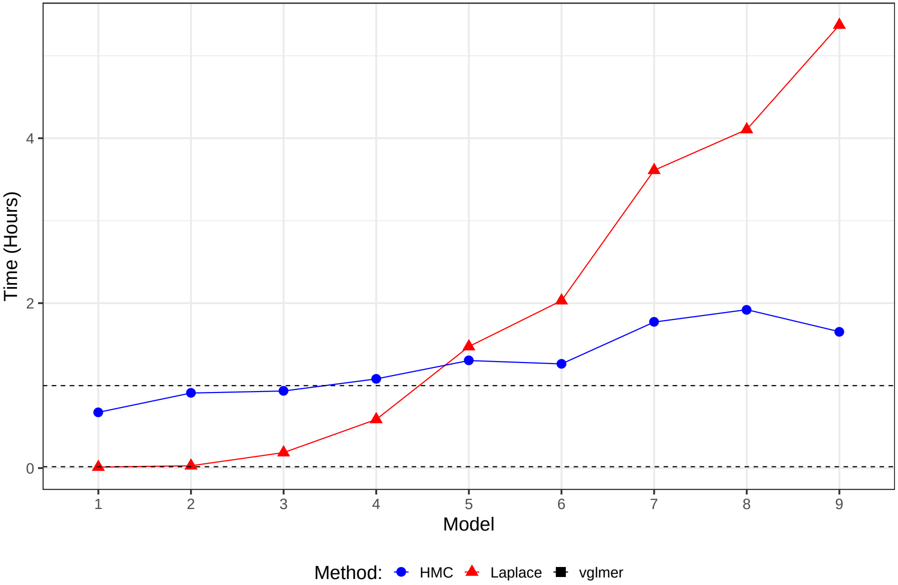
$$y_i \sim \mathrm{Binom}(n_i, p_i) \quad p_i = \frac{\exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta} + \sum_{j=1}^J \boldsymbol{z}_{i,j}^T \boldsymbol{\alpha}_{j,g[i]}\right)}{1 + \exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta} + \sum_{j=1}^J \boldsymbol{z}_{i,j}^T \boldsymbol{\alpha}_{j,g[i]}\right)} \qquad \begin{array}{c} \boldsymbol{\alpha}_{j,g} \sim^{i.i.d.} N(\boldsymbol{0}, \boldsymbol{\Sigma}_j) \\ \boldsymbol{\Sigma}_j \sim \mathrm{IW}(\nu_j, \boldsymbol{\Phi}_j) \end{array}$$

- Goal: Approximate posterior of $\boldsymbol{\theta} = \boldsymbol{\beta}, \{\boldsymbol{\alpha}_j\}, \{\boldsymbol{\Sigma}_j\}$
- Mean-Field VB: Assume independence, $q(\boldsymbol{\beta})q(\{\boldsymbol{\alpha}_j\})q(\{\boldsymbol{\Sigma}_j\})$, and find best approximation to true posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$
  - As posed, no specialized algorithm for arbitrary J (see $J = 2$ in Jeon, Rijmen, and Rabe-Hesketh 2017)
  - Requires evaluating many integrals
- Solution: Augment posterior using Polya-Gammas (Polson, Scott, and Windle 2013)
  - Tractable mean-field for $p(\boldsymbol{\theta}, \{\omega_i\}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z})$
  - Easily scalable to arbitrary J, no integration required, simple updates
  - Different "strengths" of assumption to trade-off speed & accuracy

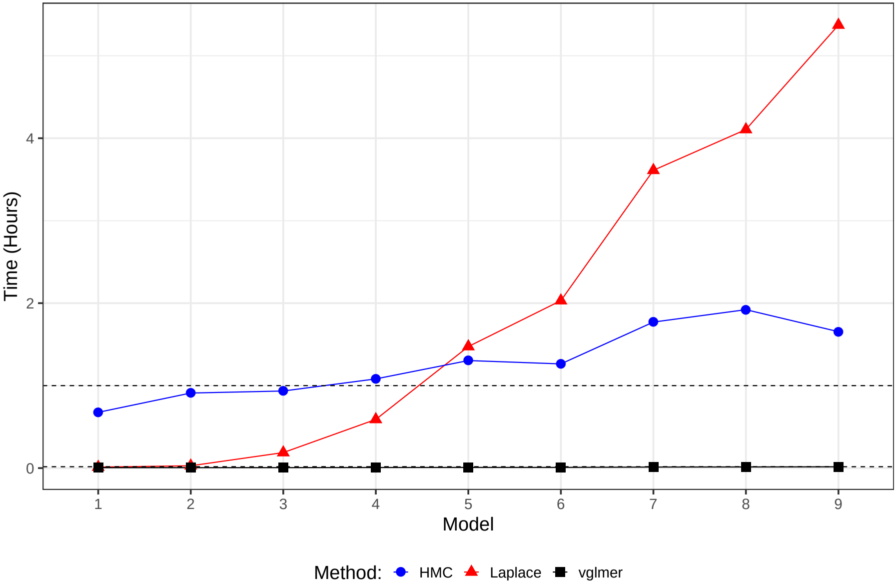
Application: Dramatic Gains in Speed



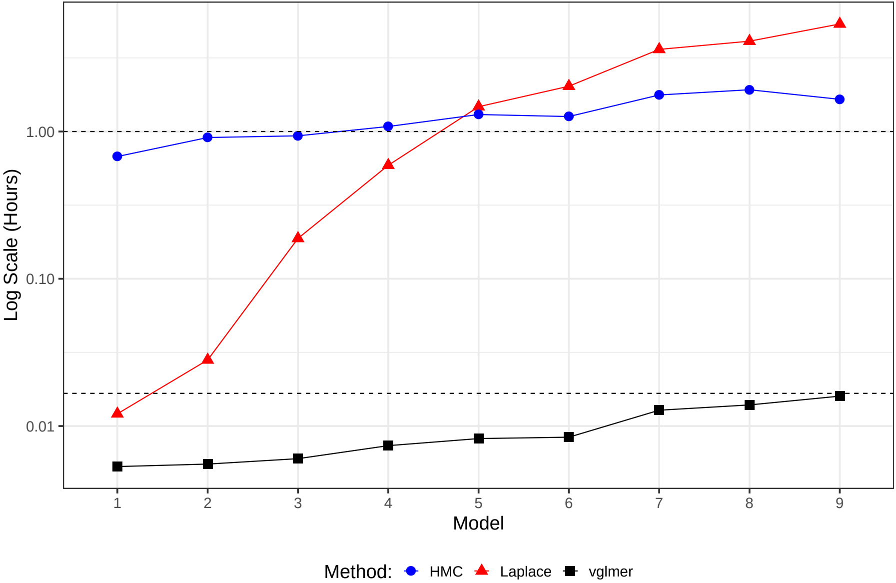
Application: Dramatic Gains in Speed



Application: Dramatic Gains in Speed



Application: Dramatic Gains in Speed



Mixed Results on Performance

- Dramatic success with speed ✓

Mixed Results on Performance

- Dramatic success with speed ✓
- Point estimates are good ✓

Mixed Results on Performance

- Dramatic success with speed ✓
- Point estimates are good ✓
  - Parameter blocks correlate highly with `glmer` (0.976) and `STAN` (0.977)

Mixed Results on Performance

- Dramatic success with speed ✓
- Point estimates are good ✓
  - Parameter blocks correlate highly with `glmer` (0.976) and `STAN` (0.977)
- Issues with variance estimates for both `glmer` and VB

Mixed Results on Performance

- Dramatic success with speed ✓
- Point estimates are good ✓
  - Parameter blocks correlate highly with `glmer` (0.976) and `STAN` (0.977)
- Issues with variance estimates for both `glmer` and VB
  - `glmer`: Some REs collapse to zero (no prior! Chung et al. 2015)
  - `vglmer`: Noticeably too small variance (well-known, general problem)

Mixed Results on Performance

- Dramatic success with speed ✓
- Point estimates are good ✓
  - Parameter blocks correlate highly with `glmer` (0.976) and `STAN` (0.977)
- Issues with variance estimates for both `glmer` and VB
  - `glmer`: Some REs collapse to zero (no prior! Chung et al. 2015)
  - `vglmer`: Noticeably too small variance (well-known, general problem)
  - Median parameter block has

Mixed Results on Performance

- Dramatic success with speed ✓
- Point estimates are good ✓
  - Parameter blocks correlate highly with `glmer` (0.976) and `STAN` (0.977)
- Issues with variance estimates for both `glmer` and VB
  - `glmer`: Some REs collapse to zero (no prior! Chung et al. 2015)
  - `vglmer`: Noticeably too small variance (well-known, general problem)
  - Median parameter block has
    - `vglmer`: 17% smaller standard deviation than HMC

Mixed Results on Performance

- Dramatic success with speed ✓
- Point estimates are good ✓
  - Parameter blocks correlate highly with `glmer` (0.976) and `STAN` (0.977)
- Issues with variance estimates for both `glmer` and VB
  - `glmer`: Some REs collapse to zero (no prior! Chung et al. 2015)
  - `vglmer`: Noticeably too small variance (well-known, general problem)
  - Median parameter block has
    - `vglmer`: 17% smaller standard deviation than HMC
    - `glmer`: 36% smaller standard deviation than HMC

Mixed Results on Performance

- Dramatic success with speed ✓
- Point estimates are good ✓
  - Parameter blocks correlate highly with `glmer` (0.976) and `STAN` (0.977)
- Issues with variance estimates for both `glmer` and VB
  - `glmer`: Some REs collapse to zero (no prior! Chung et al. 2015)
  - `vglmer`: Noticeably too small variance (well-known, general problem)
  - Median parameter block has
    - `vglmer`: 17% smaller standard deviation than HMC
    - `glmer`: 36% smaller standard deviation than HMC
- Simulations show a similar story:
  - All recover point estimates well
  - `glmer` has poor coverage for REs
  - `vglmer` undercovers somewhat
  - Alternative variational methods (`ADVI`) do very poorly

Marginal Augmentation to the Rescue!

Marginal Augmentation to the Rescue!

- **Second Goal of Paper:** Cheap way to improve initial approximation (although it still is an approximation!)

Marginal Augmentation to the Rescue!

- **Second Goal of Paper:** Cheap way to improve initial approximation (although it still is an approximation!)
- Procedure:

Marginal Augmentation to the Rescue!

- **Second Goal of Paper:** Cheap way to improve initial approximation (although it still is an approximation!)
- Procedure:
  - Find approximation using VB and draw m samples

MA Marginal Augmentation to the Rescue!

- **Second Goal of Paper:** Cheap way to improve initial approximation (although it still is an approximation!)
- Procedure:
  - Find approximation using VB and draw m samples
  - Run m chains of MCMC for one step using some transition kernel k (e.g. marginal augmentation [MA], Gibbs, HMC, etc.)

Marginal Augmentation to the Rescue!

- **Second Goal of Paper:** Cheap way to improve initial approximation (although it still is an approximation!)
- Procedure:
  - Find approximation using VB and draw m samples
  - Run m chains of MCMC for one step using some transition kernel k (e.g. marginal augmentation [MA], Gibbs, HMC, etc.)
  - Use new samples as approximation!

Marginal Augmentation to the Rescue!

- **Second Goal of Paper:** Cheap way to improve initial approximation (although it still is an approximation!)
- Procedure:
  - Find approximation using VB and draw m samples
  - Run m chains of MCMC for one step using some transition kernel k (e.g. marginal augmentation [MA], Gibbs, HMC, etc.)
  - Use new samples as approximation!
- Use MA because (i) simple & (ii) known to work well for MCMC on hierarchical models (Van Dyk and Meng 2001)

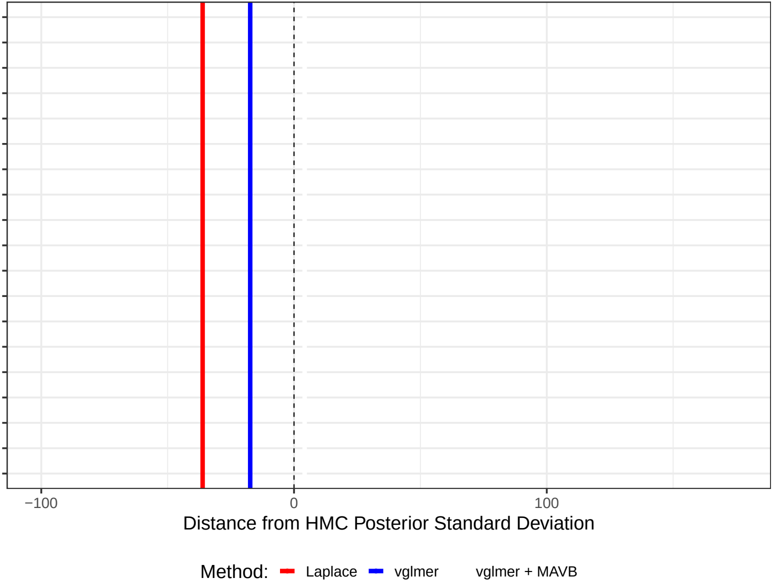
Marginal Augmentation to the Rescue!

- **Second Goal of Paper:** Cheap way to improve initial approximation (although it still is an approximation!)
- Procedure:
  - Find approximation using VB and draw m samples
  - Run m chains of MCMC for one step using some transition kernel k (e.g. marginal augmentation [MA], Gibbs, HMC, etc.)
  - Use new samples as approximation!
- Use MA because (i) simple & (ii) known to work well for MCMC on hierarchical models (Van Dyk and Meng 2001)
- Provides a guaranteed improvement (e.g. Ruiz and Titsias 2019)

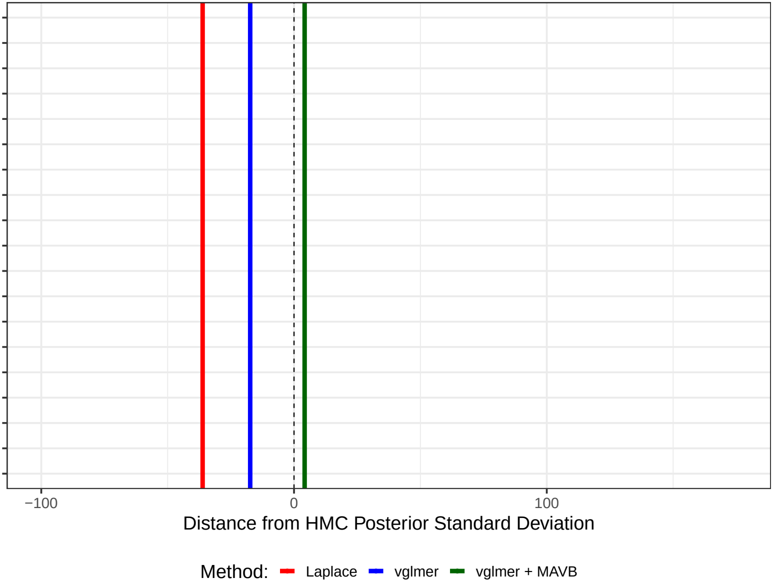
Marginal Augmentation to the Rescue!

- **Second Goal of Paper:** Cheap way to improve initial approximation (although it still is an approximation!)
- Procedure:
  - Find approximation using VB and draw m samples
  - Run m chains of MCMC for one step using some transition kernel k (e.g. marginal augmentation [MA], Gibbs, HMC, etc.)
  - Use new samples as approximation!
- Use MA because (i) simple & (ii) known to work well for MCMC on hierarchical models (Van Dyk and Meng 2001)
- Provides a guaranteed improvement (e.g. Ruiz and Titsias 2019)
- **Intuition**: Running one step of MCMC makes approximation better $\rightarrow$ induces dependencies between parameters

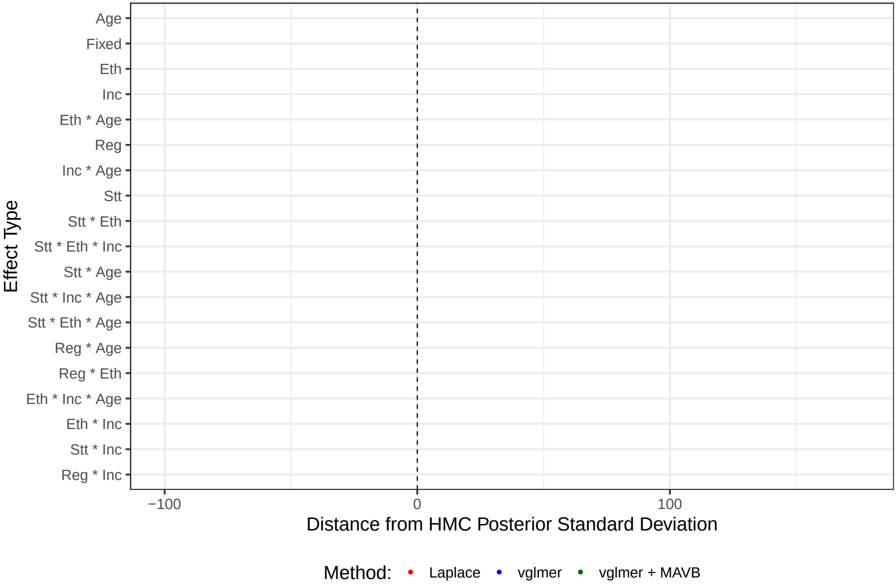
Improving Estimates with MAVB



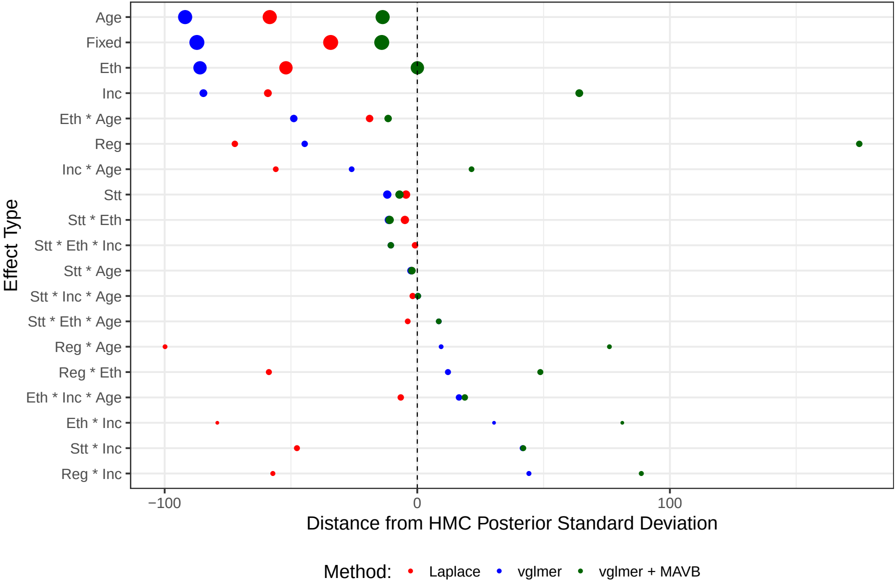
Improving Estimates with MAVB



Improving Estimates with MAVB



Improving Estimates with MAVB



Advice for MRP Practitioners

Advice for MRP Practitioners

- Ghitza and Gelman use $J = 18$; what about other choices?

Advice for MRP Practitioners

- Ghitza and Gelman use $J = 18$; what about other choices?
- Use 10-fold cross-validation to compare 9 models

Advice for MRP Practitioners

- Ghitza and Gelman use $J = 18$; what about other choices?
- Use 10-fold cross-validation to compare 9 models
  - Prohibitive for `STAN` or `glmer`

Advice for MRP Practitioners

- Ghitza and Gelman use $J = 18$; what about other choices?
- Use 10-fold cross-validation to compare 9 models
  - Prohibitive for `STAN` or `glmer`
  - `vglmer` $\rightarrow$ 20 minutes for all 9 models!

Advice for MRP Practitioners

- Ghitza and Gelman use $J = 18$; what about other choices?
- Use 10-fold cross-validation to compare 9 models
  - Prohibitive for `STAN` or `glmer`
  - `vglmer` $\rightarrow$ 20 minutes for all 9 models!
- Summary:
  - Adding demographic x state two-way interactions $\rightarrow$ big lift
  - Intermediate complexity ($J = 10$) performs better than $J = 18$

Advice for MRP Practitioners

- Ghitza and Gelman use $J = 18$; what about other choices?
- Use 10-fold cross-validation to compare 9 models
  - Prohibitive for `STAN` or `glmer`
  - `vglmer` $\rightarrow$ 20 minutes for all 9 models!
- Summary:
  - Adding demographic x state two-way interactions $\rightarrow$ big lift
  - Intermediate complexity ($J = 10$) performs better than $J = 18$
- Improve models by some interactions, but don't go too deep!

Conclusions

Conclusions

- Hierarchical models are popular in political science

Conclusions

- Hierarchical models are popular in political science
- Estimation for non-linear outcomes is time-consuming—limiting model exploration & checking

Conclusions

- Hierarchical models are popular in political science
- Estimation for non-linear outcomes is time-consuming—limiting model exploration & checking
- Developed a new approximate algorithm (MAVB)
  - Can be used for binomial, (count, and multinomial outcomes)
  - Can include any number or type of (normal) random effects

Conclusions

- Hierarchical models are popular in political science
- Estimation for non-linear outcomes is time-consuming—limiting model exploration & checking
- Developed a new approximate algorithm (MAVB)
  - Can be used for binomial, (count, and multinomial outcomes)
  - Can include any number or type of (normal) random effects
- Considerable speed gains with limited cost in terms of accuracy

Conclusions

- Hierarchical models are popular in political science
- Estimation for non-linear outcomes is time-consuming—limiting model exploration & checking
- Developed a new approximate algorithm (MAVB)
    - Can be used for binomial, (count, and multinomial outcomes)
    - Can include any number or type of (normal) random effects
- Considerable speed gains with limited cost in terms of accuracy
- Can improve poor uncertainty estimates by simple "post-processing"

Conclusions

- Hierarchical models are popular in political science
- Estimation for non-linear outcomes is time-consuming—limiting model exploration & checking
- Developed a new approximate algorithm (MAVB)
  - Can be used for binomial, (count, and multinomial outcomes)
  - Can include any number or type of (normal) random effects
- Considerable speed gains with limited cost in terms of accuracy
- Can improve poor uncertainty estimates by simple "post-processing"
- Competitive with `glmer` in performance & much faster!

Conclusions

- Hierarchical models are popular in political science
- Estimation for non-linear outcomes is time-consuming—limiting model exploration & checking
- Developed a new approximate algorithm (MAVB)
  - Can be used for binomial, (count, and multinomial outcomes)
  - Can include any number or type of (normal) random effects
- Considerable speed gains with limited cost in terms of accuracy
- Can improve poor uncertainty estimates by simple "post-processing"
- Competitive with `glmer` in performance & much faster!
- On-Going Work: Looking for more papers & models to examine!

Conclusions

- Hierarchical models are popular in political science
- Estimation for non-linear outcomes is time-consuming—limiting model exploration & checking
- Developed a new approximate algorithm (MAVB)
  - Can be used for binomial, (count, and multinomial outcomes)
  - Can include any number or type of (normal) random effects
- Considerable speed gains with limited cost in terms of accuracy
- Can improve poor uncertainty estimates by simple "post-processing"
- Competitive with `glmer` in performance & much faster!
- On-Going Work: Looking for more papers & models to examine!

github.com/mgoplerud/vglmer | j.mp/goplerud_MAVB
mgoplerud.com | mgoplerud@pitt.edu

References I

Chung, Yeojin, Andrew Gelman, Sophia Rabe-Hesketh, Jingchen Liu, and Vincent Dorie. 2015. “Weakly Informative Prior for Point Estimation of Covariance Matrices in Hierarchical Models.” *Journal of Educational and Behavioral Statistics* 40 (2): 136–157.

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.

Jeon, Minjeong, Frank Rijmen, and Sophia Rabe-Hesketh. 2017. “A Variational Maximization–Maximization Algorithm for Generalized Linear Mixed Models with Crossed Random Effects.” *Psychometrika* 82 (3): 693–716.

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.

References II

Ruiz, Francisco J.R., and Michalis K. Titsias. 2019. "A Contrastive Divergence for Combining Variational Inference and MCMC." In *International Conference on Machine Learning.* http://proceedings.mlr.press/v97/ruiz19a/ruiz19a.pdf.

Van Dyk, David A., and Xiao-Li Meng. 2001. "The Art of Data Augmentation." *Journal of Computational and Graphical Statistics* 10 (1): 1–50.