Dissertation Defense: Numerical Methods for Approximating High-Dimensional Posterior Distributions

Jameson Quinn

12/9/19

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{\int_{\theta' \in \Theta} P(x|\theta')P(\theta')d\theta'}$$

Structure of thesis

- Chapter 1: online data assimilation in spatiotemporal systems
- Chapter 2: new method for variational inference on latent variable models
  - Contributions: Laplace guide families; analytic amortization
- Chapter 3: application to ecological inference (EI)
  - Contributions: extensible model for EI; full algorithm and implementation of Laplace VI for this model

Collaborator on Ch. 2-3: Mira Bernstein

- On most things, equal collaborator and coauthor
- All the major motivating ideas, and $>95\%$ of the coding, is mine
- We checked that this is OK

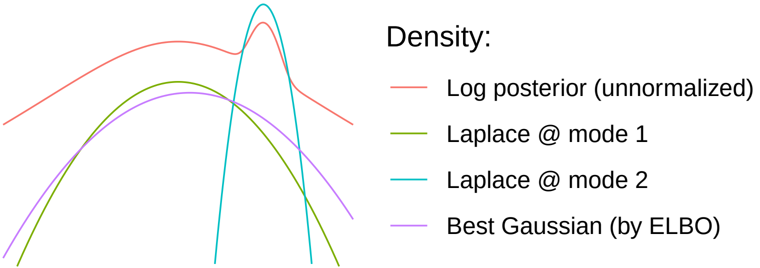
Variational Inference

Approximate w/ guide distribution $q_\phi(\boldsymbol{\theta})$; choose ϕ to mimize KL:

$$\hat{\phi} = \operatorname{argmin}_\phi \left[D_{\mathrm{KL}} \left(\, q_\phi(\boldsymbol{\theta}) \, || \, p(\boldsymbol{\theta}|\boldsymbol{x}) \, \right) \right].$$

Equivalent to maximizing ELBO:

$$\mathrm{ELBO}(\phi) := E_{q_\phi} \left[\log p(\boldsymbol{x}, \boldsymbol{\theta}) - \log q_\phi(\boldsymbol{\theta}) \right]$$



Computational tool: Pyro

Released in 2017 and still under very active development, pyro is a cutting-edge python package for black-box VI.

- Stochastic optimization (hill-climbing)
- Automatic differentiation via PyTorch ML

Choosing a guide family

This talk will focus on Gaussian guide families.

The first obvious possibility for the guide family of a d-parameter model is just the unrestricted set of d-dimensional Gaussians.

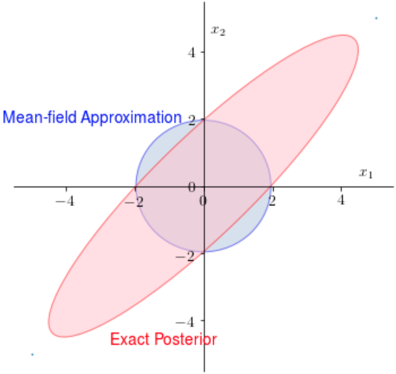
- Mean: d guide parameters (1 per model parameter)
- Covariances: $\mathcal{O}(d^2)$ guide parameters

Meanfield guide family

A common assumption is posterior independence of parameters, referred to as "meanfield" guides. Thus guide parameters:

- Mean: d guide parameters (1 per model parameter)
- Variances: d guide parameters (1 per model parameter; diagonal covariance matrix)

Problem:



Who will guide us?

Among Gaussian guide families:

- Set of all normals, with unrestricted covariance, is too big
- Meanfield subfamily doesn't actually contain any good approximations
- We want subfamily that contains at least some good approximations without being too big

Introducing: Laplace family

Let's guarantee that the family contains the Laplace approximation around any posterior mode. This allows us to parametrize only the mean, and then derive the precision matrix by the taking the observed information of the posterior:

$$\mathcal{I}_p\left(\boldsymbol{\theta}^*\right) := -H\left[\log p(\boldsymbol{\theta})\right]\Big|_{\boldsymbol{\theta}^*}$$

Thus, the guide parameters for a model $p(\theta)$ would be θ^*, defining the point at which to take a Laplace approximation.

- Means (θ^*): d guide parameters (1 per model parameter)
- Covariance: 0 guide parameters! Just compute $\mathcal{I}_p(\theta^*)$.

Boosting function

$\mathcal{I}_p$ not guaranteed to be positive definite. So define "boosting" function $f(\mathcal{I}_p)$ s.t.:

- Guaranteed p.d.
- Smooth almost everywhere.
- $f(\mathcal{I}_p) \approx \mathcal{I}_p$ if $\mathcal{I}_p$ already p.d.

A similar problem arises in optimization (quasi-Newton methods); solved via modified Cholesky algorithms (Surveyed in Fang, 2008; we use GMW81 by Gill, Murray, & Wright)

Boosting family

Boosting functions are designed to ensure that their output is positive definite "enough". But how much is "enough"?

Instead of deciding arbitrarily, we can parametrize f to create a boosting family $f_{\boldsymbol{\psi}}$, for $\psi_i \in \mathbb{R}_+^D$, s.t. as $\boldsymbol{\psi} \to \vec{\mathbf{0}}$, $f(\mathcal{I}_p) \to \mathcal{I}_p$ if $\mathcal{I}_p$ already p.d.

Formal definition of Laplace family

Let $p(\boldsymbol{\theta})$ be a (twice-differentiable) probability density over $\mathbb{R}^d$.

Let $\Theta \subseteq \mathbb{R}^d$, $\Psi \subseteq \mathbb{R}^d_+$, and let f_Ψ be a boosting family.

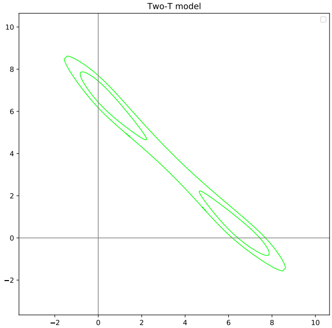
Laplace guide: $q_{\boldsymbol{\theta}^* \in \Theta, \boldsymbol{\psi} \in \Psi}(\boldsymbol{\theta})$, a d-dimensional Gaussian with mean $\boldsymbol{\theta}^*$ and precision precision matrix $f_{\boldsymbol{\psi}}\left(\mathcal{I}_p(\boldsymbol{\theta}^*)\right)$.

Laplace guide family $\mathcal{L}_{\Theta \times \Psi}(p, f_\Psi)$: $\{q_{\boldsymbol{\theta}^*, \boldsymbol{\psi}} : \boldsymbol{\theta}^* \in \Theta,\ \boldsymbol{\psi} \in \Psi\}$

Thus, $2d$ guide parameters.

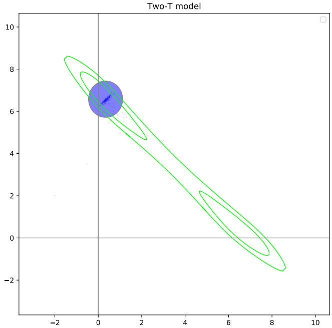
Toy model

$$p(T_1, T_2 | x = 7)$$
$$x = T_1 + T_2 + \epsilon$$
$$T_i \sim StudentT_\nu(0,1); i \in \{1,2\}$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$



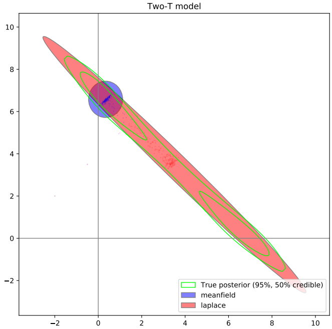
Toy model

$$p(T_1, T_2 | x = 7)$$
$$x = T_1 + T_2 + \epsilon$$
$$T_i \sim StudentT_\nu(0, 1); i \in \{1, 2\}$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$



Toy model

$$p(T_1, T_2 | x = 7)$$
$$x = T_1 + T_2 + \epsilon$$
$$T_i \sim StudentT_\nu(0, 1); i \in \{1, 2\}$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$



Latent variable models (or: why hi-D?)

A latent variable model has 3 core elements:

- Global parameters: $\boldsymbol{\gamma} \in \Gamma \cong \mathbb{R}^g$,
- Latent parameter vectors: $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_N \in \Lambda \cong \mathbb{R}^l$
- Observation vectors: $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$

$$p\left(\gamma, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_N, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\right) = p(\gamma) \prod_{i=1}^{N} p(\boldsymbol{\lambda}_i|\gamma)\, p(\boldsymbol{x}_i|\boldsymbol{\lambda}_i, \gamma)$$

Laplace guide parameters: $\gamma^*, \boldsymbol{\lambda}_1^* \ldots \boldsymbol{\lambda}_N^*, \boldsymbol{\psi}$

Latent variable models: Block Arrowhead Hessians

$$
\mathcal{I}_p(\boldsymbol{\theta}^*) = \begin{array}{cc}
 & \begin{array}{cccccc} \boldsymbol{\gamma} & \boldsymbol{\lambda}_1 & \boldsymbol{\lambda}_2 & \dots & \boldsymbol{\lambda}_N \end{array} \\
\begin{array}{c} \boldsymbol{\gamma} \\ \boldsymbol{\lambda}_1 \\ \boldsymbol{\lambda}_1 \\ \vdots \\ \boldsymbol{\lambda}_N \end{array} &
\begin{pmatrix}
G & C_1 & C_2 & \dots & C_N \\
C_1^T & U_1 & 0 & \dots & 0 \\
C_2^T & 0 & U_2 & \dots & 0 \\
\vdots & 0 & 0 & \ddots & 0 \\
C_N^T & 0 & 0 & \dots & U_N
\end{pmatrix}
\end{array}
$$

- Easy to boost.
- Easy to sample from. Note that marginal covariance for $\boldsymbol{\gamma}$ is $[\mathcal{I}_p(\boldsymbol{\theta}^*)^{-1}]_{\Gamma,\Gamma} = (G - \sum_i C_i U_i^{-1} C_i^T)^{-1}$

SVI (Stochastic Variational Inference)

At each optimization step, let $\mathcal{S}$ be a random sample of units with $p(i \in \mathcal{S}) = \pi_i$.

Replace

$$\log p(\boldsymbol{\theta}, \boldsymbol{x}) := \log p(\boldsymbol{\gamma}) + \sum_{i=1}^{N} \left[\log p(\boldsymbol{\lambda}_i | \boldsymbol{\gamma}) \; + \log p(\boldsymbol{x}_i | \boldsymbol{\lambda}_i, \boldsymbol{\gamma}) \right]$$

with the unbiased estimator

$$\log p_{\mathcal{S}}(\boldsymbol{\theta}_{\mathcal{S}}, \boldsymbol{x}_{\mathcal{S}}) := \log p(\boldsymbol{\gamma}) + \frac{1}{\pi_i} \sum_{i \in \mathcal{S}} \left[\log p(\boldsymbol{\lambda}_i | \boldsymbol{\gamma}) \; + \log p(\boldsymbol{x}_i | \boldsymbol{\lambda}_i, \boldsymbol{\gamma}) \right]$$

Then compute Laplace guide for the latter expression, and find the ELBO gradient.

SVI (Stochastic Variational Inference): unbiased?

With Laplace guide, this makes:

- Log density: unbiased
- Guide covariance of globals for given $\boldsymbol{\theta}^*$: Up to boosting, unbiased for both conditional precision and "marginal precision" (inverse of marginal covariance).
- ELBO and ELBO gradient: Not unbiased (unlike meanfield)

Amortization

Suppose we can find an analytic function for the (approximate) conditional MAP: $f(\boldsymbol{\gamma}, \boldsymbol{x}_i) \approx \mathrm{argmax}_{\boldsymbol{\lambda}_i} p(\boldsymbol{\lambda}_i | \gamma, \boldsymbol{x}_i)$

Then we can reduce the number of guide parameters by restricting to a lower-dimensional subfamily of $\mathcal{L}$:

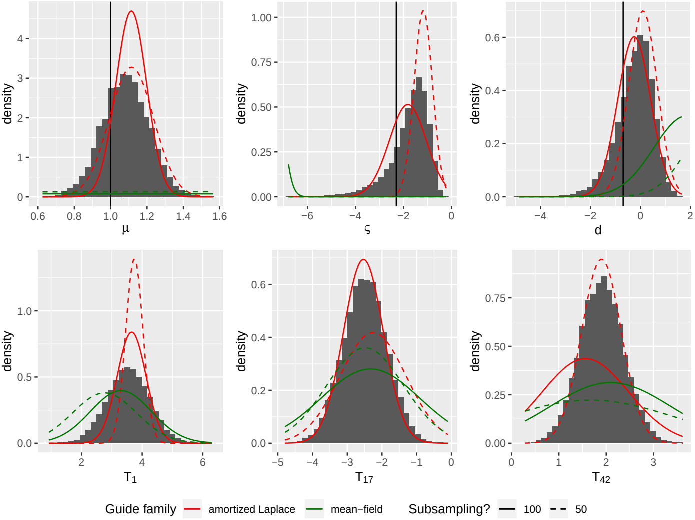
- Restrict Θ: set $\boldsymbol{\lambda}_i^*$ to $f(\boldsymbol{\gamma}^*, \boldsymbol{x}_i)$
- Restrict Ψ: reuse the same boosting parameters for each unit

Laplace guide parameters: $\boldsymbol{\gamma}^*, \boldsymbol{\psi}_\gamma, \boldsymbol{\psi}_\lambda$

Multisite model

$$T_i/\sigma \sim StudentT_\nu; \quad i \in \{1, ..., N\}$$
$$\tau_i = \mu + T_i$$
$$x_i \sim \mathcal{N}(\tau_i, s_i^2)$$
$$d := \log(\nu - \nu_{\min}) \sim \mathcal{N}(1, 1.5^2)$$
$$\varsigma := \log(\sigma - \sigma_{\min}) \sim \mathcal{N}(0, 2^2)$$
$$\mu \sim \mathcal{N}(0, 20)$$

Results


density
μ
ς
d
T_1
T_{17}
T_{42}
Guide family
amortized Laplace
mean−field
Subsampling?
100
50

Ch. 3: Ecological inference (EI)

EI: inferring individual behavior from aggregated data.

Motivating example: voting behavior by racial or other groups

Candidate

Race	X	Y	Z	n_u
White	?	?	?	400
Black	?	?	?	200
Hispanic	?	?	?	100
Other	?	?	?	100
v_u	400	200	200	800

precinct

Ecological inference (EI)

EI: inferring individual behavior from aggregated data.

Motivating example: voting behavior by racial or other groups

		Candidate			
		X	Y	Z	n_u
Race	*White*	400	0	0	400
	Black	0	200	0	200
	Hispanic	0	0	100	100
	Other	0	0	100	100
	v_u	400	200	200	800

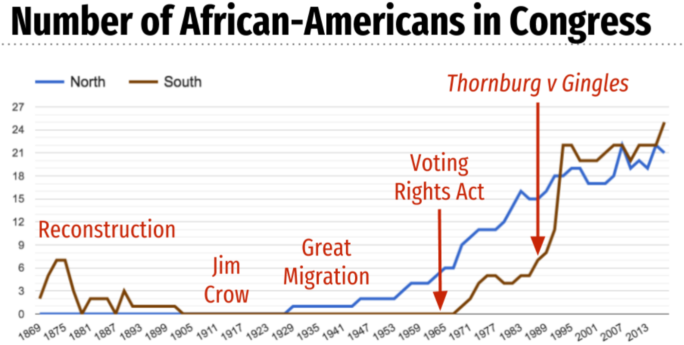
precinct

		Candidate			
		X	Y	Z	n_u
Race	*White*	200	100	100	400
	Black	100	50	50	200
	Hispanic	50	25	25	100
	Other	50	25	25	100
	v_u	400	200	200	800

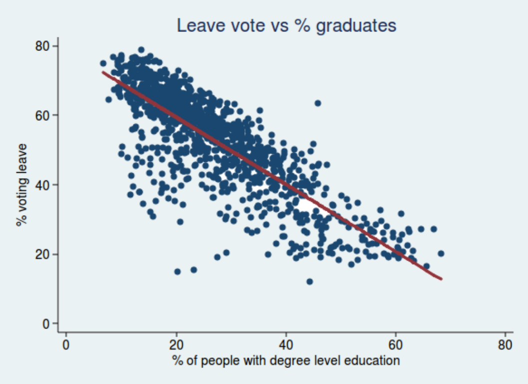
precinct

Thornburg v. Gingles, 1986

When you can show racially polarized voting, a minority community is entitled to a majority-minority district. Result:

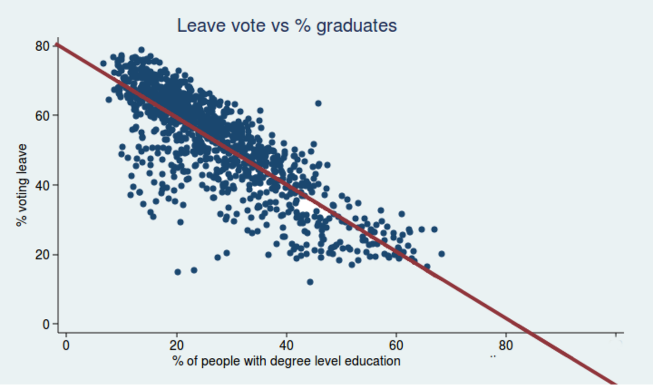


First attempt: Ecological regression (ER)



Brexit voting data. (Example by Adam Jacobs.)

First attempt: Ecological regression (ERrrrr. . .)

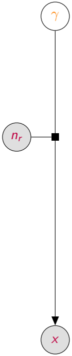


Brexit support: -16% of those with a degree???

Comparison of models

ER
(1953)

Global voting propensities
(for each race)



Precinct vote totals
by candidate (observed)

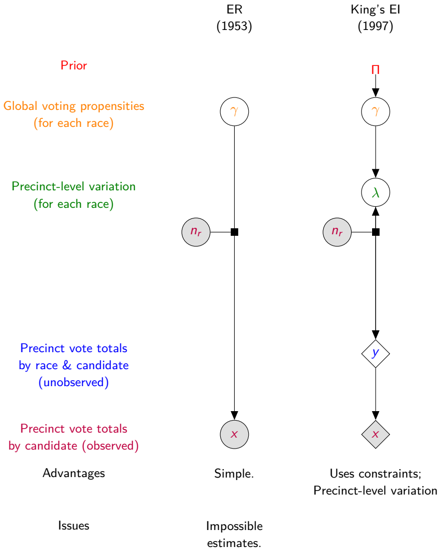
Advantages

Simple.

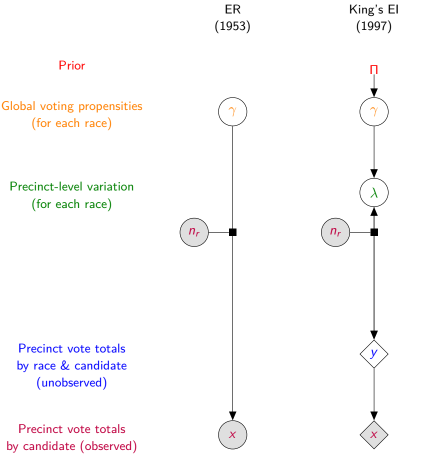
Issues

Impossible
estimates.

Comparison of models



Comparison of models

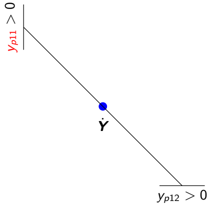


	ER (1953)	King's EI (1997)
Advantages	Simple.	Uses constraints; Precinct-level variation
Issues	Impossible estimates.	Only 2×2; No voter-level randomness.

Polytope

Candidate

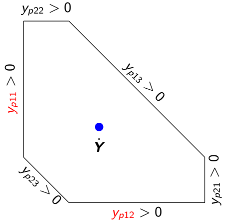
Race		X	Y	$\boldsymbol{n}_u$
	White	50	50	100
	Black	70	70	140
	$\boldsymbol{v}_u$	120	120	240



$\mathcal{Y}_u$

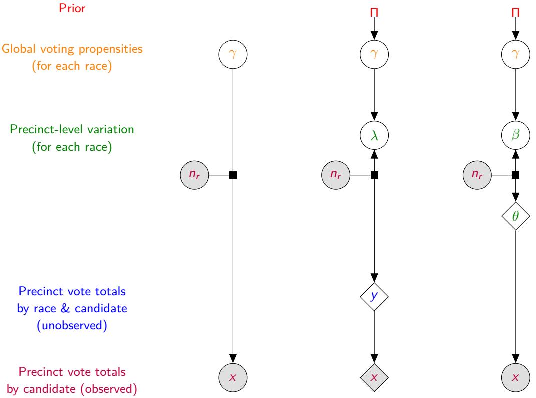
Candidate

Race		X	Y	Z	$\boldsymbol{n}_u$
	White	50	50	50	150
	Black	70	70	70	210
	$\boldsymbol{v}_u$	120	120	120	360



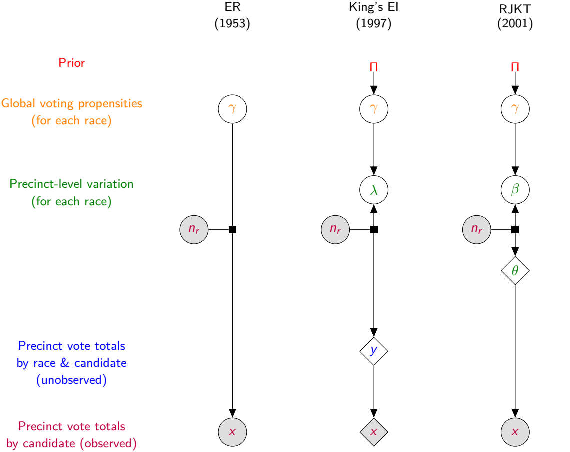
$\mathcal{Y}_u$

Comparison of models



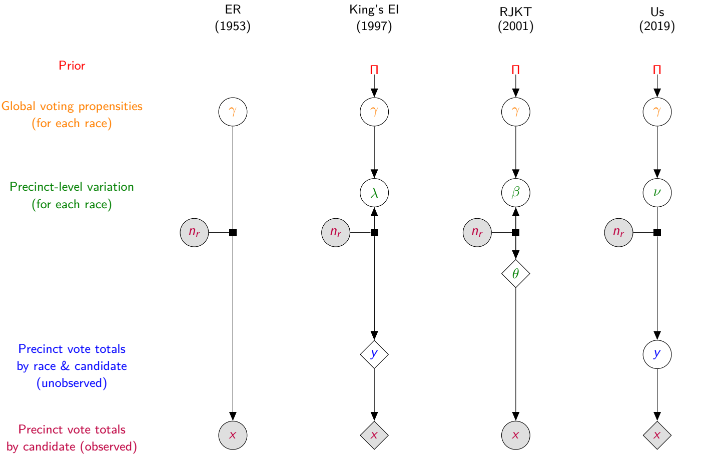
	ER (1953)	King's EI (1997)	RJKT (2001)
Advantages	Simple.	Uses constraints; Precinct-level variation	$R \times C$; includes voter multinomial.
Issues	Impossible estimates.	Only 2×2; No voter-level randomness.	

Comparison of models



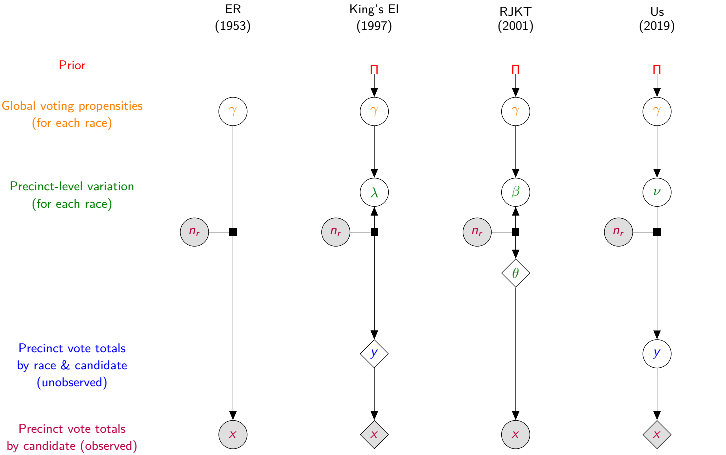
	ER (1953)	King's EI (1997)	RJKT (2001)
Advantages	Simple.	Uses constraints; Precinct-level variation	$R \times C$; includes voter multinomial.
Issues	Impossible estimates.	Only 2×2; No voter-level randomness.	Cheats a bit on constraints: no y.

Comparison of models



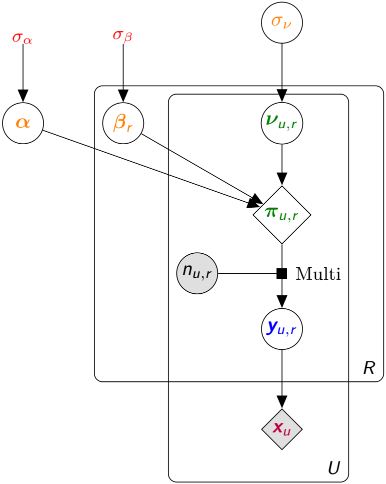
	ER (1953)	King's EI (1997)	RJKT (2001)	Us (2019)
Advantages	Simple.	Uses constraints; Precinct-level variation	$R \times C$; includes voter multinomial.	Respects constraints; $R \times C$; "multinomial".
Issues	Impossible estimates.	Only 2×2; No voter-level randomness.	Cheats a bit on constraints: no y.	

Comparison of models

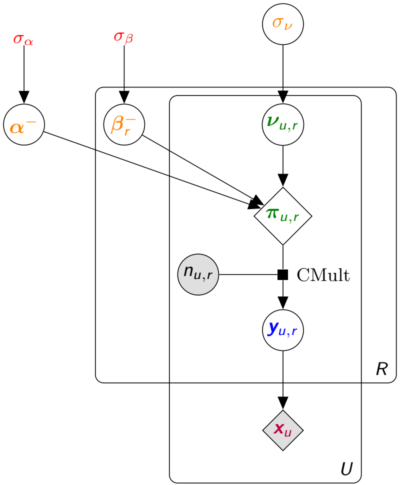


	ER (1953)	King's EI (1997)	RJKT (2001)	Us (2019)
Advantages	Simple.	Uses constraints; Precinct-level variation	$R \times C$; includes voter multinomial.	Respects constraints; $R \times C$; "multinomial". Extensible!
Issues	Impossible estimates.	Only 2×2; No voter-level randomness.	Cheats a bit on constraints: no y.	

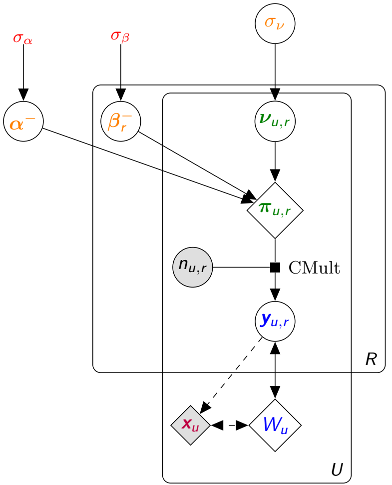
Our model


σ_α
σ_β
σ_ν
α
β_r
$\nu_{u,r}$
$\pi_{u,r}$
$n_{u,r}$
Multi
$y_{u,r}$
R
x_u
U

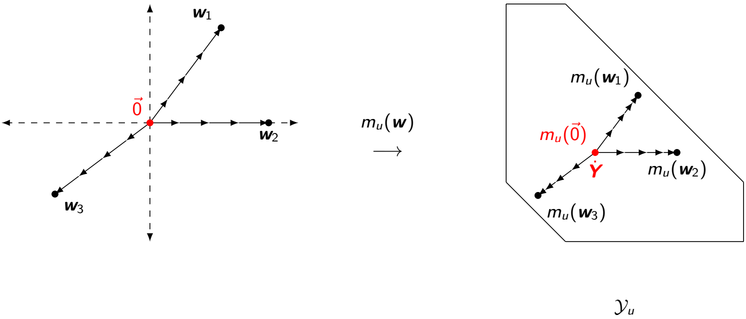
Modified model



Modified model



Polytopize



- Start at $m_u(\mathbf{0}) = \dot{Y}$: independence point
- Project to correct subspace
- Apply mapping which preserves angle but shrinks length
- Smooth almost everywhere

Simulation study

- 3×3:
  - Races: White/Black/Other
  - Candidates: Republican/Democrat/Other (where "other" includes both 3rd party and not voting)
- Realistic parameter settings for NC Pres 2016
  - Use racial demographics by precinct in NC (available because of VRA)
  - Use exit polls to calculate realistic α and β for 2016 presidential election
- Simulate datasets from our model using arbitrary choices of σ_ν: 0.02, 0.3

Simulation study

- Run LVI using pyro, RJKT using eiPack in R.
- Let $Q_{rc} = \frac{\sum_u y_{urc}}{\sum_u n_{ur}}$, the percentage of race r that voted for candidate c
- Report mean and standard deviation of Q_{rc}, based on 1000 samples from the estimated posterior

Results for $\sigma_\nu = 0.02$

		Other/none		Clinton (D)		Trump (R)	
	$\mathcal{A}$	$\overline{Q}$	s_Q	$\overline{Q}$	s_Q	$\overline{Q}$	s_Q
White	Truth	32.4%		22.6%		45.0%	
	RJKT	32.3%	0.059%	22.5%	0.069%	45.0%	0.057%
	LVI	32.3%	0.028%	22.7%	0.028%	45.0%	0.030%
Black	Truth	38.1%		56.6%		5.28%	
	RJKT	38.4%	0.32%	56.2%	0.19%	4.87%	0.19%
	LVI	37.9%	0.067%	56.3%	0.058%	5.79%	0.046%
Other	Truth	43.4%		32.7%		24.0%	
	RJKT	41.5%	0.65%	32.9%	0.41%	24.2%	0.72%
	LVI	44.4%	0.25%	32.4%	0.21%	23.3%	0.25%

Results for $\sigma_\nu = 0.3$

		Other/none		Clinton (D)		Trump (R)	
	$\mathcal{A}$	$\overline{Q}$	s_Q	$\overline{Q}$	s_Q	$\overline{Q}$	s_Q
White	Truth	32.3%		22.7%		45.0%	
	RJKT	32.3%	0.080%	22.8%	0.11%	44.8%	0.14%
	LVI	33.2%	0.067%	23.5%	0.049%	43.0%	0.070%
Black	Truth	38.9%		55.6%		5.48%	
	RJKT	40.4%	0.39%	53.7%	0.38%	5.33%	0.20%
	LVI	36.3%	0.17%	54.6%	0.15%	9.21%	0.16%
Other	Truth	43.0%		32.7%		24.3%	
	RJKT	38.3%	1.1%	35.4%	0.52%	24.9%	1.1%
	LVI	45.0%	0.46%	33.6%	0.40%	21.7%	0.47%

Discussion/future work (Ch. 3)

- Including racial makeup of precinct as a covariate
- Hierarchical model (counties)
- Multiple elections
- Actual NC data
- Compare hierachical model without EI, Standard RJKT, and our model
- Cross-validation

Discussion/future work (Ch. 2)

- More on subsampling:
  - General theory of how to assign weights to minimize variance of estimator in subsampling
  - Some theory to help choose sample size for SVI
- Investigate replacing normal with T in guide

Thanks

Thank you!

Directory of extra slides

- Prior work
- EI amortization
- *Thornburg v. Gingles*
- Model (for discussing extensions)

Non-meanfield prior work

- Copula VI (Han et al 2015): create arbitrary transformations, to allow "quasi-correlation matrices" for non-Gaussian families. The values for such matrices are unrestricted, though.
- Time-series (Zhang et al 2017): model-specific tricks.
- Hierarchical VI (Ranganath et al 2015): put a prior on the guide then marginalize it out. Relies on conjugacy.
- Variational Boosting (Miller et al 2016): Correlation structure: low-rank plus diagonal.
- Normalizing flows (Rezende et al 2015): Kinda like adding a step of MCMC after sampling from guide.
- "Laplace Variational Inference" (Wang & Blei, 2012): Use Laplace approximation to approve update step in a conjugate meanfield VI.
- Imaging (Zhang et al 2017): Use Laplace approximation around posterior mode for certain fixed-size subsets of parameters. No boosting.

How our EI amortization works

Which variables are we amortizing: Y, nu, sigma_nu

Steps:

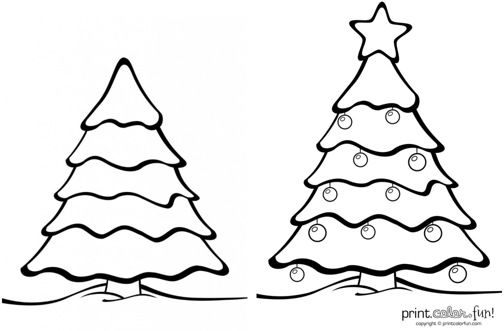
- Find approximate mode of $p(\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{\beta}; \boldsymbol{\nu} = 0)$ constrained to lie on polytope (this is linear algebra plus Stirling's approximation)
- Not yet done: One-dimensional Newton's method to find approximate mode of W. (Not the same thing, because there's Jacobian, mode of W is further away from boundary)
- Find approximate mode of $p(\boldsymbol{\nu}_u, \sigma_\nu|\gamma, \boldsymbol{W}_u)$ (ad hoc algorithm, but see next step)
- Newton's method (for free!!)

Thornburg v Gingles, 1986

A majority-minority district must be created if:

1. A minority group is "sufficiently numerous and compact to form a majority in a single-member district"; and
2. The minority group is **"politically cohesive"**; and
3. The "majority **votes sufficiently as a bloc** to enable it . . . usually to defeat the minority's preferred candidate."

Flexible model



$$\vec{y}_{p,r} = y_{p,r,c}|_{c=1}^{C} \sim \text{CMult}\left(n_{p,r}, \frac{exp(\alpha_c + \beta_{r,c} + \nu_{r,c,p})|_{c=1}^{C}}{\sum_{c=1}^{C} exp(\alpha_c + \beta_{r,c} + \nu_{r,c,p})}\right)$$

$$\alpha_c \sim \mathcal{N}(0, 2)$$

$$\beta_{r,c} \sim \mathcal{N}(0, 2)$$