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

Jameson Quinn

12/9/19

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

Always good to start a stats talk with Bayes' Thm.
You've all seen this before.
As you know, the trickiest part is the denominator.
If integral is over hi-D space, numerical methods for estimating will have unacceptably high variance.

So you need tricks.

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

Ch. 2: The VI framework is to assume the posterior is well-approximated... construct a new guide family that's able to...
Ch 3: More than a simple application. The model is more realistic and more extensible than the most common method, and applying VI here requires several tricks.

You will notice a few changes to what I sent you, particularly in the chapter 3 results;

I will point them out as we go along

Collaborator: 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

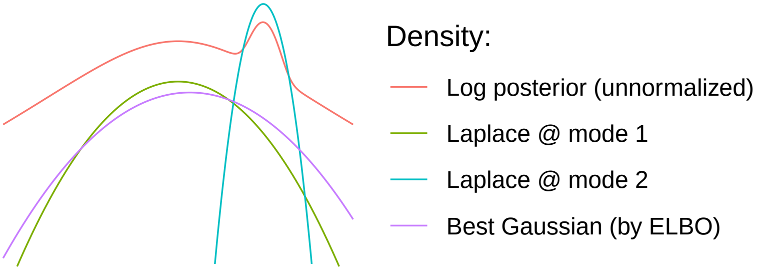
Variational Inference

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

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

Equivalent to maximizing ELBO:

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



minimizing KL-divergence, or without normalization, the ELBO)
Note model parameters here and guide parameters there
Optimizing, but over distributions, so entropy;

hill-climing means automatic differentiation, so...

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

Explain automatic differentiation

All the significant software engineering I had to do

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 Gaussians.

Guide parameters:

- Means: 1 per model parameter
- Covariances: ...

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 Gaussians.

Guide parameters:

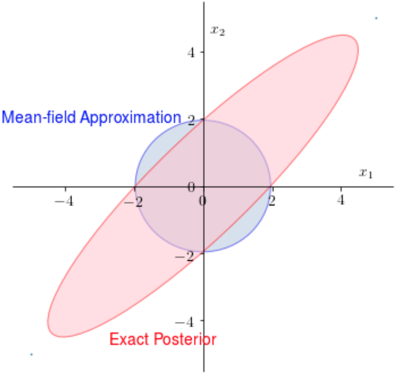
- Means: 1 per model parameter
- Covariances: $\mathcal{O}(d^2)$

Meanfield guide family

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

can't capture posterior correlations

systematically underestimates posterior marginals



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.

Don't ignore correlation. Don't optimize over it. Just get it by calculus.

Boosting

$\mathcal{I}_p$ 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)

Furthermore, we can parametrize f to create a boosting family f_ψ, 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.

Boosting family is better than just boosting function.
D-dimensional so we can boost dif params dif.

Version of thesis sent previously has quasi-boosting which we're no longer using

Formal definition of Laplace family

Let $p(\boldsymbol{\theta})$ be a 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: $q_{\boldsymbol{\theta}^*,\boldsymbol{\psi}}(\boldsymbol{\theta})$, a d-dimensional Gaussian with mean $\boldsymbol{\theta}^*$ and precision precision matrix $f_{\boldsymbol{\psi}}\left(\mathcal{I}_p(\boldsymbol{\theta}^*)\right)$.

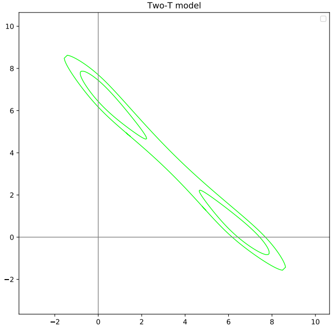
Laplace guide family $\mathcal{L}_{\Theta\times\Psi}(p, f_\Psi)$: $\{q_{\boldsymbol{\theta}^*,\boldsymbol{\psi}} : \boldsymbol{\theta}^* \in \Theta,\ \boldsymbol{\psi} \in \Psi\}$

Boosting family is better than just boosting function.
D-dimensional so we can boost dif params dif.
Version of thesis sent previously has quasi-boosting which we're no longer using

Citation for method that we're actually using

Toy model

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



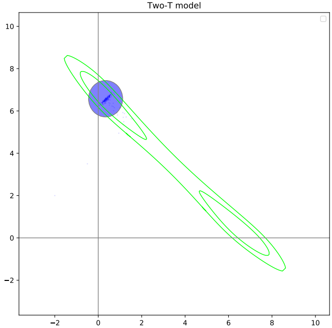
Simple model with bimodal posterior
Shows several things: Importance of covariance; case where laplace of MAP isn't

Toy model

$$x = T_1 + T_2 + \epsilon$$

$$T_i \sim StudentT_\nu(0, 1); i \in \{1, 2\}$$

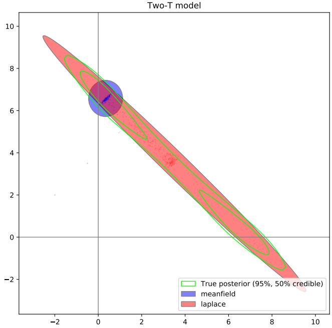
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$



Simple model with bimodal posterior
Shows several things: Importance of covariance; case where laplace of MAP isn't

Toy model

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



Simple model with bimodal posterior
Shows several things: Importance of covariance; case where laplace of MAP isn't

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

A latent variable model has 3 core elements:

- Global parameters: $\gamma \in \Gamma \cong \mathbb{R}^g$,
- 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: $\boldsymbol{\gamma}^*, \boldsymbol{\lambda}^*, \boldsymbol{\psi}_\gamma, \boldsymbol{\psi}_\lambda$

This is why we need a high-dimensional solution.

Latent variable models: Hessians

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

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

easy to sample from and easy to boost.

SVI (Stochastic Variational Inference)

Two methods useful w/ LVM. Conceptually independent, but combine. 1st, standard:

Replace

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

with

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

With Laplace guide, this makes:

- Posterior log density: unbiased (as in meanfield)
- Hessian: Up to boosting, unbiased for both conditional precision and "marginal precision" (inverse of marginal covariance).
- ELBO and ELBO gradient: Not unbiased (unlike meanfield)

If cheap way to predict bigger terms, weights; we haven't implemented

ELBO & gradient not unbiased as in meanfield, but seem to work well

Amortization

Lower-dimensional subfamily of a Laplace family: reduce the number of guide parameters by setting $\boldsymbol{\lambda}^*$ to $f(\boldsymbol{\gamma}^*)$

The aim is to find an analytic a priori function that sets the latents to (approximately) their conditional MAP values.

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

Note that it's MAP not MLE.
MAP is not perfect but it's close.

computationally cheap refinement is available

Multisite model

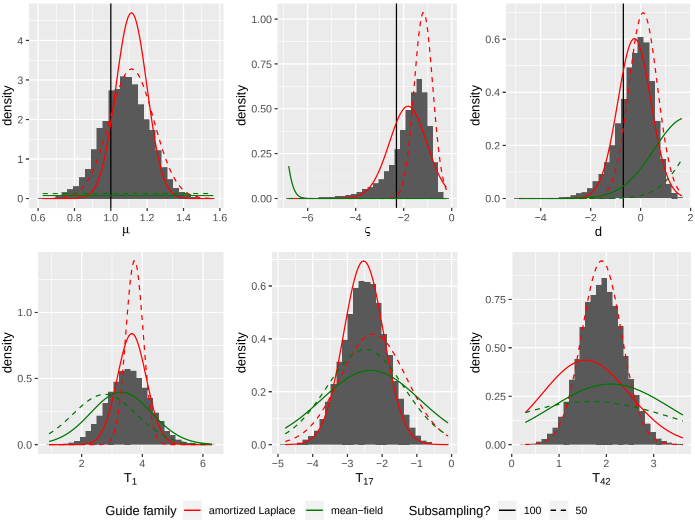
$$d := \log(\nu - \nu_{\text{min}}) \sim \mathcal{N}(1, 1.5^2)$$

$$\varsigma := \log(\sigma - \sigma_{\text{min}}) \sim \mathcal{N}(0, 2^2)$$

$$\mu \sim \mathcal{N}(0, 20)$$

$$\nu_{\text{min}} = 2.5,\ \sigma_{\text{min}} = \max(s_i) * 1.9$$

Results



Changes: df; boosting

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

Mainly comes up in voting rights cases, so I'll talk about it in this setting.

Z could represent not voting.

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	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

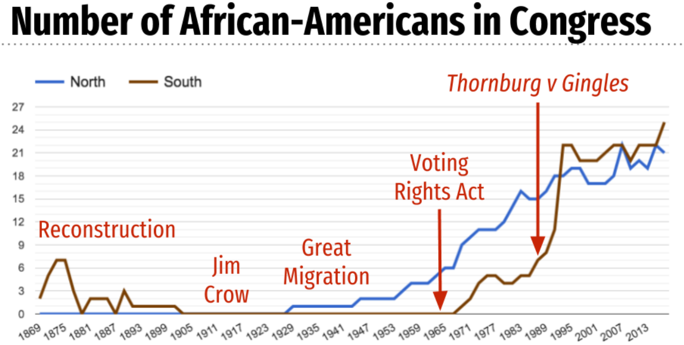
Candidate

Race		X	Y	Z	n_u
	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

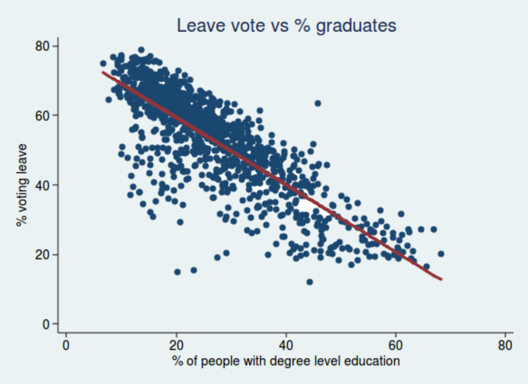
precinct

Thornburg v. Gingles, 1986

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



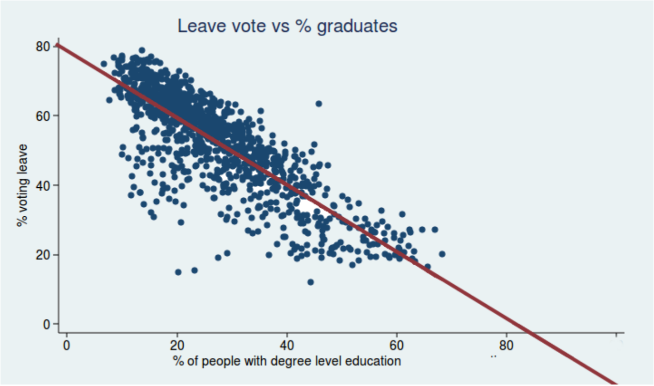
First attempt: Ecological regression (ER)



Brexit voting data. (Example by Adam Jacobs.)

Want to know how Brexit support differed by education. So...

First attempt: Ecological regression (ERrrrr. . .)



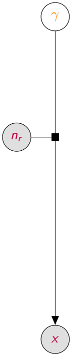
Brexit support: -16% of those with a degree???

Strong model assumption, which is incorrect: no precinct-level variation

Comparison of models

ER
(1953)

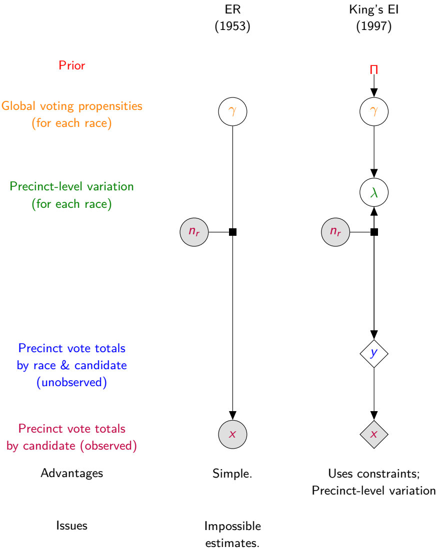
Global voting propensities
(for each race)



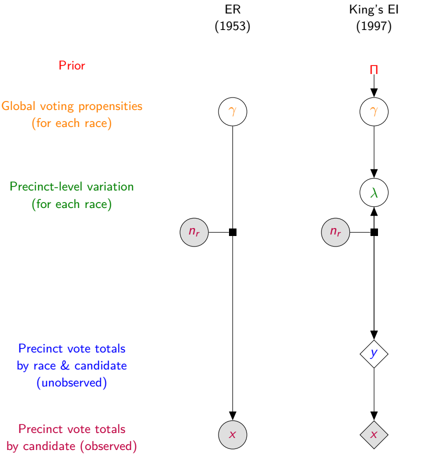
Precinct vote totals
by candidate (observed)

	ER (1953)
Advantages	Simple.
Issues	Impossible estimates.

Comparison of models



Comparison of models



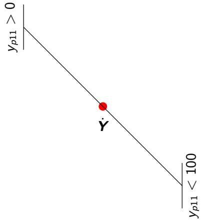
	ER (1953)	King's EI (1997)
Advantages	Simple.	Uses constraints; Precinct-level variation
Issues	Impossible estimates.	Only 2×2; No voter-level randomness.

Polytopize

Candidate

Race

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

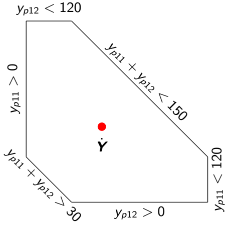


$\mathcal{Y}_u$

Candidate

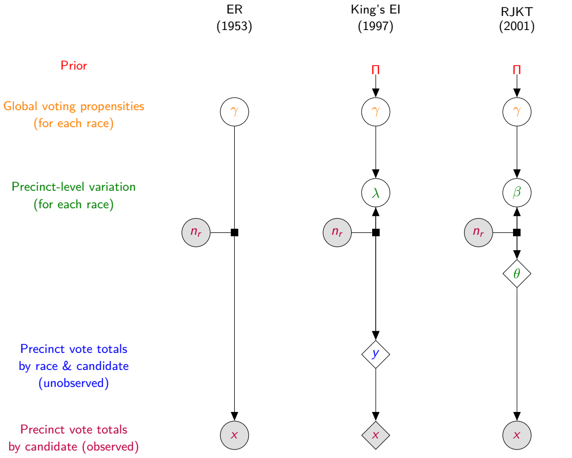
Race

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



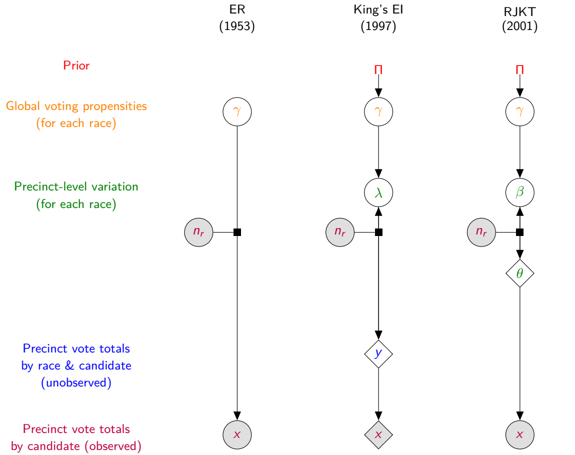
$\mathcal{Y}_u$

Comparison of models



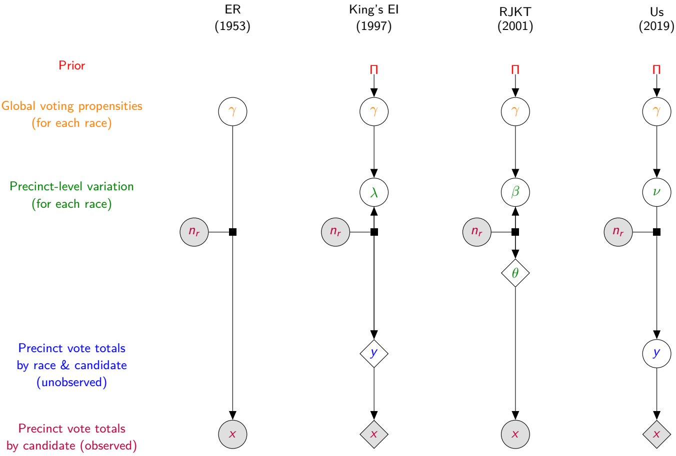
	ER (1953)	King's EI (1997)	RJKT (2001)
Advantages	Simple.	Uses constraints; Precinct-level variation	$R \times C$; includes voter multinomial.
Issues	Impossible estimates.	Only 2×2; No voter-level randomness.	

Comparison of models



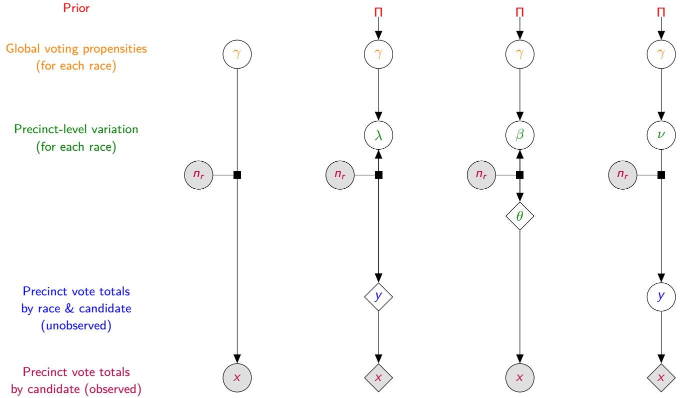
Advantages	Simple.	Uses constraints; Precinct-level variation	$R \times C$; includes voter multinomial.
Issues	Impossible estimates.	Only 2×2; No voter-level randomness.	Cheats a bit on constraints: no y.

Comparison of models



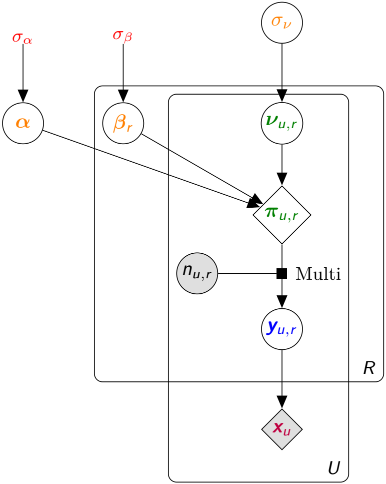
Advantages	Simple.	Uses constraints; Precinct-level variation	$R \times C$; includes voter multinomial.	Respects constraints; $R \times C$; "multinomial".
Issues	Impossible estimates.	Only 2×2; No voter-level randomness.	Cheats a bit on constraints: no y.	

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.	

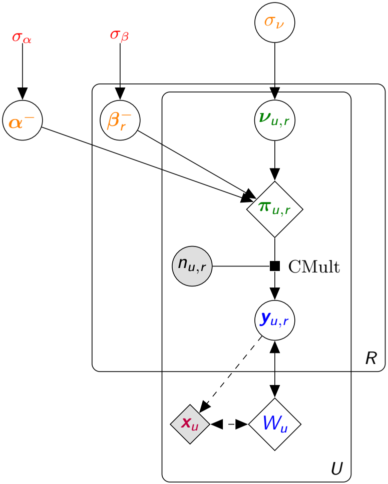
Our model



Talk briefly about how you could add other Christmas tree ornaments

Talk about why this is hard to make a guide for

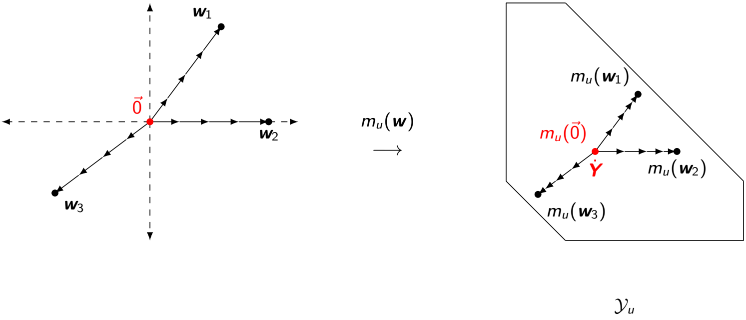
Modified model



Cmult; Polytopize: ae smooth bijective map from R^n to polytope
Pseudovoters because of boundary issues that arise from Cmult and polytopize

All of these make the model itself slightly less-realistic, but make VI work.

Polytopize



Let $m_u(\boldsymbol{w}) := g(a(\boldsymbol{w}))$, where a is affine projection, g is retraction

$b(M)$: the intersection of ray $\overrightarrow{\dot{Y}M}$ with boundary of $\mathcal{Y}_u$

$$g(M) := \begin{cases} \dot{Y} & \text{if } M = \dot{Y}, \\ \dot{Y} + \exp\left(-\frac{|b(M) - \dot{Y}|}{|M - \dot{Y}|}\right) \cdot (b(M) - \dot{Y}) & \text{otherwise.} \end{cases}$$

- Note that $m_u(\mathbf{0}) = \dot{Y}$.

Testing our EI on simulated data

- Data cleaning on racial breakdown by precinct in NC
- Used exit polls to calculate realistic α and β for 2016 presidential election
- 3×3:
  - Races: White/Black/Other
  - Candidates: Trump/Clinton/Other (where "other" includes both 3rd party and not voting)
- Simulated datasets from our model using three choices of σ_ν: 0.02, 0.1, 0.3
- Ran LVI using pyro, RJKT using eiPack in R.
- Reporting $\bar{Q}$ and s_Q for each race and candidate: the mean and s.d. of $Q(z_i)$ where z_i is a sample from the fitted posterior and Q gives the percent of the given race who voted for the given candidate.

EI results

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%

Point out that this is different (better!) than what you originally sent, because:
does not underestimate variance (fixed bug)
corrected alphas and betas (so that overall percentages of people of each race voting for each candidate approximate the true 2016 data, as intended)
improved amortization (optimize Y <U+2192> optimize W)

Conclusion: We are as good as RJKT, but we're just getting started

EI results

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%

Point out that this is different (better!) than what you originally sent, because:
does not underestimate variance (fixed bug)
corrected alphas and betas (so that overall percentages of people of each race voting for each candidate approximate the true 2016 data, as intended)
did NOT improve amortization (optimize Y -> optimize W)

Conclusion: We are as good as RJKT, but we're just getting started

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

Say that this is the stuff we plan to include in final paper

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

(use Ch 3 as example)

Say that this will not be in current paper, which is basically done

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:

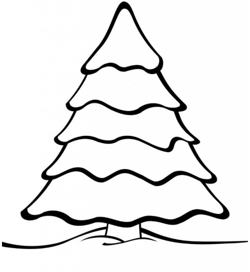
- Find approximate mode of $p(\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{\beta}; \boldsymbol{\nu} = 0)$ constrained to lie on polytope (this is linear algebra plus Stirling's approximation)
- Not yet done: One-dimensional Newton's method to find approximate mode of W. (Not the same thing, because 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!!)

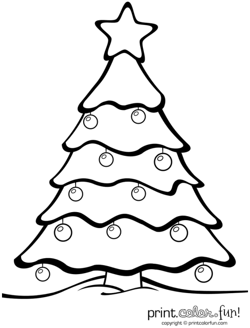
Thornburg v Gingles, 1986

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





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

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

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