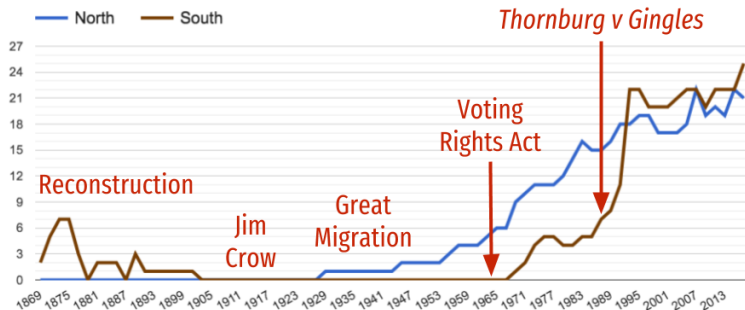


# Flexible Ecological Inference using Variational Methods

Jameson Quinn (work with Mira Bernstein)

3/13/2019

# Number of African-Americans in Congress

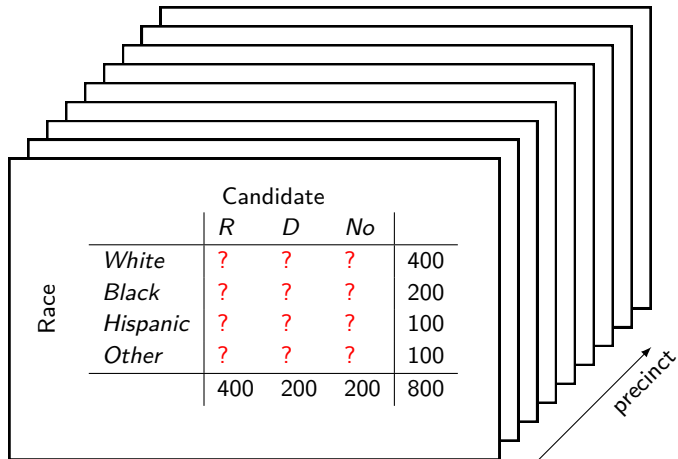


# Thornburg v Gingles, 1986

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.”

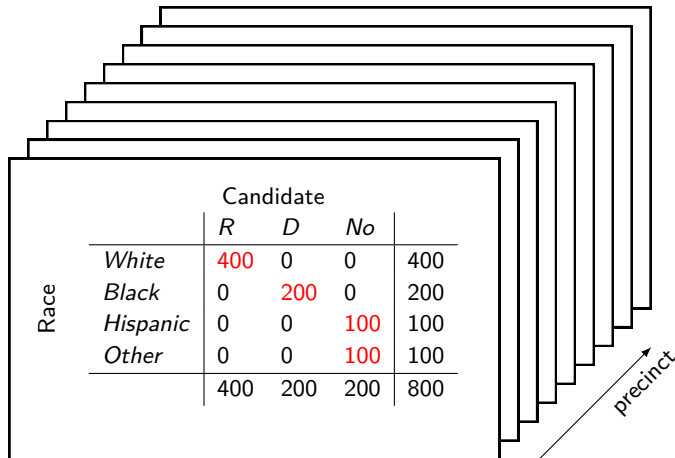
# Ecological data



The image shows a stack of 10 identical tables, each representing data for a different precinct. An arrow labeled "precinct" points from the bottom table to the right, indicating the direction of increasing precinct index. The table structure is as follows:

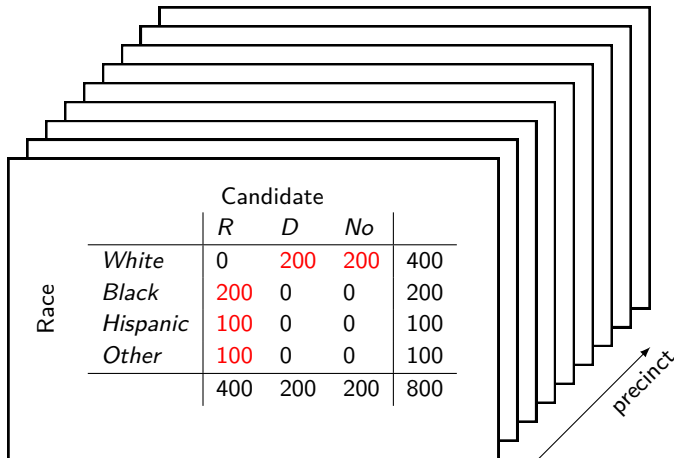
Race	Candidate			
	<i>R</i>	<i>D</i>	<i>No</i>	
	<i>White</i>	?	?	400
	<i>Black</i>	?	?	200
	<i>Hispanic</i>	?	?	100
<i>Other</i>	?	?	?	100
	400	200	200	800

# Majority=Majority?



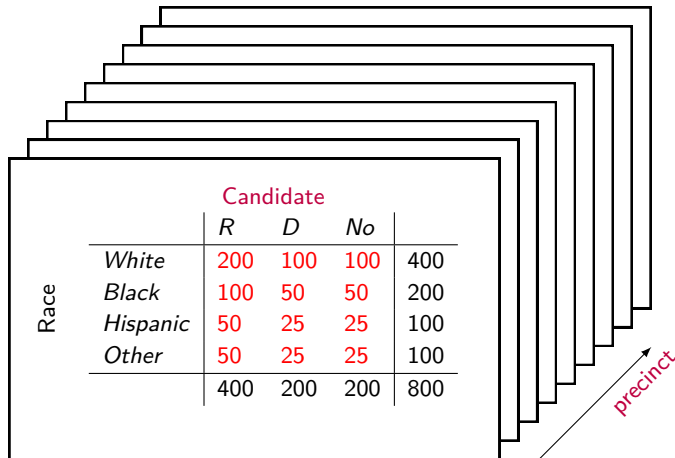
Race	Candidate				
	<i>R</i>	<i>D</i>	<i>No</i>		
	<i>White</i>	400	0	0	400
	<i>Black</i>	0	200	0	200
	<i>Hispanic</i>	0	0	100	100
<i>Other</i>	0	0	100	100	
	400	200	200		800

# Backwards?



Race	Candidate				
	<i>R</i>	<i>D</i>	<i>No</i>		
	<i>White</i>	0	200	200	400
	<i>Black</i>	200	0	0	200
	<i>Hispanic</i>	100	0	0	100
<i>Other</i>	100	0	0	100	
	400	200	200		800

# Independence?



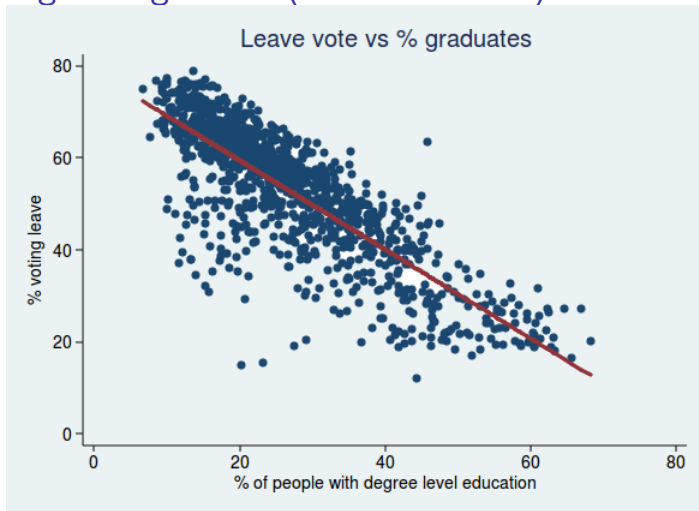
Race	Candidate				
	<i>R</i>	<i>D</i>	<i>No</i>		
	<i>White</i>	200	100	100	400
	<i>Black</i>	100	50	50	200
	<i>Hispanic</i>	50	25	25	100
<i>Other</i>	50	25	25	100	
	400	200	200		800

# Structure

- ▶ Pose the ecological problem (done)
- ▶ Quick review of prior approaches
- ▶ A basic, extensible model for EI
- ▶ Why and how to reparameterize
- ▶ Review of variational inference
- ▶ Applying variational inference to EI
- ▶ Guide (aka variational distribution) based on observed information

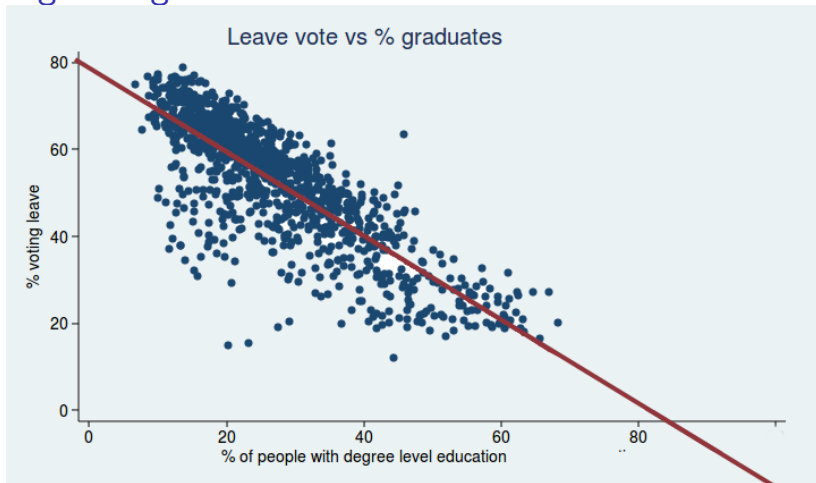


## Ecological regression (for $2 \times 2$ cases)



Brexit voting data. (Example from “The Stats Guy” blog by Adam Jacobs.) Valid under certain (strong) assumptions.

## Ecological regression: uh oh



Brexit supported by 79% of people without a degree... and -16% of those with one??? Ecological fallacy, Simpson's paradox, etc.

# Infer latents, not parameters

Insight from King, Rosen, Tanner 1999: instead of focusing on population parameters, which are not directly constrained by the data, focus on latent parameters, which are.

Refined by Rosen, King, Jiang, Tanner (2001):

- ▶ Fully Bayesian model
- ▶ extends to  $R \neq 2 \neq C$
- ▶ fast, moment-based estimator
- ▶ now widely used.

## Issues with RKJT 2001:

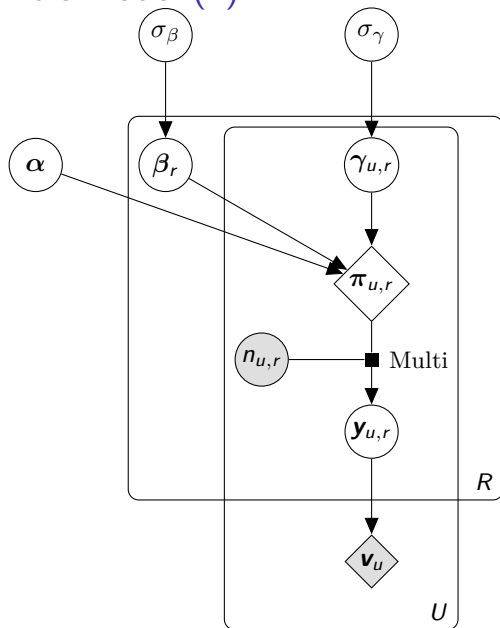
The RKJT model can, in principle, be extended to handle additional factors such as:

- ▶ inter-row or inter-column correlations
- ▶ covariates
- ▶ multiple elections
- ▶ exit polling data
- ▶ etc.

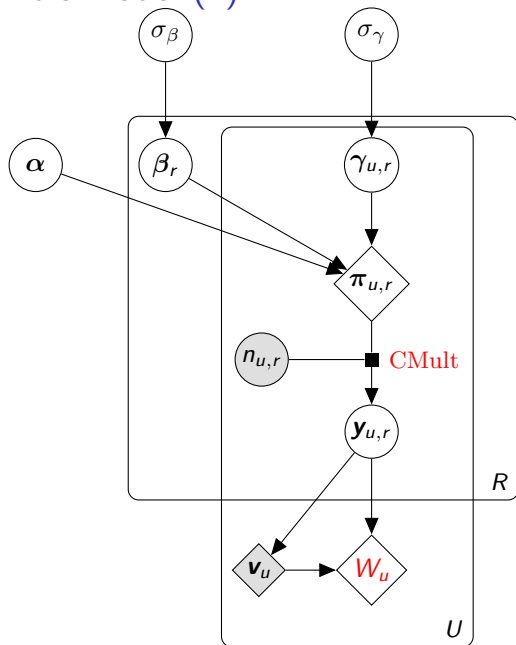
However:

- ▶ the moment-based estimator breaks down,
- ▶ MCMC on such a high-dimensional latent space can be challenging.

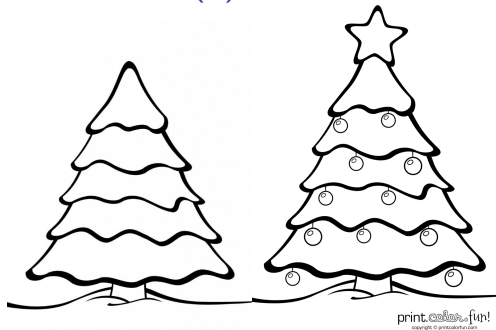
## Flexible model (1)



## Flexible model (2)



## Flexible model (3)

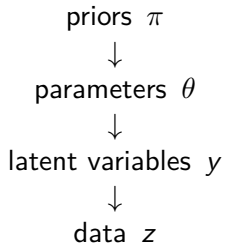


$$\vec{y}_{p,r} = y_{p,r,c} \big\|_{c=1}^C \sim \text{CMult} \left( n_{p,r}, \frac{\exp(\alpha_c + \beta_{r,c} + \lambda_{r,c,p}) \big\|_{c=1}^C}{\sum_{c=1}^C \exp(\alpha_c + \beta_{r,c} + \lambda_{r,c,p})} \right)$$

$$\alpha_c \sim \mathcal{N}(0, \sigma_\alpha) \quad \sigma_\alpha \sim \text{Expo}(5)$$

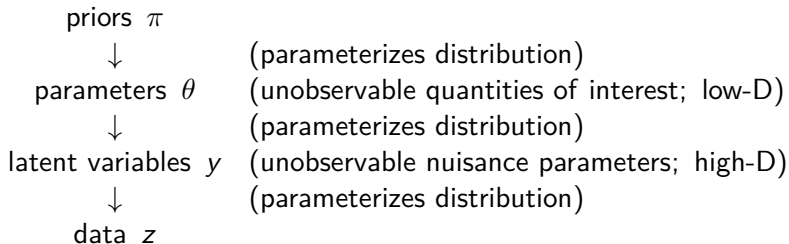
$$\beta_{r,c} \sim \mathcal{N}(0, \sigma_\beta) \quad \sigma_\beta \sim \text{Expo}(5)$$

# Standard Bayesian approach (simplified)

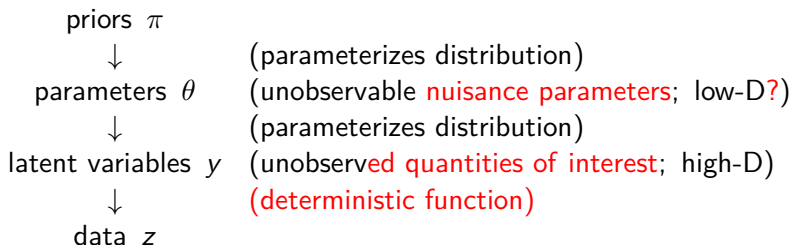




## Standard Bayesian approach (cont'd)



# Ecological Inference

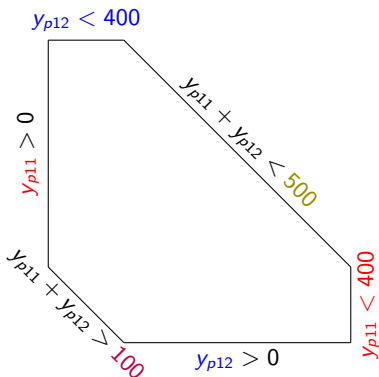


A likelihood from a deterministic function is an indicator function!

## Ecological case

Since the likelihood is just an indicator function, the posterior is just the prior, restricted to the set of values where the likelihood is 1 and renormalized. For each precinct, this set turns out to be a polytope  $\mathcal{Y}_{z_p}$  in an  $(R - 1)(C - 1)$  dimensional subspace of the full  $\mathbb{R}^{RC}$ .

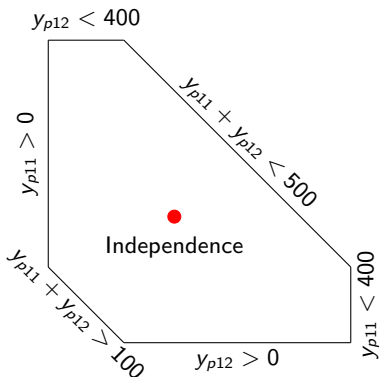
Race	Candidate			
	<i>R</i>	<i>D</i>	<i>No</i>	
	$y_{p11}$	$y_{p12}$	$y_{p13}$	
	$y_{p21}$	$y_{p22}$	$y_{p23}$	
400	400	400	400	1200



$$\mathcal{Y}_{z_p} \subset \mathbb{R}^{(R-1)(C-1)} \longleftrightarrow \mathbb{R}^{RC}$$

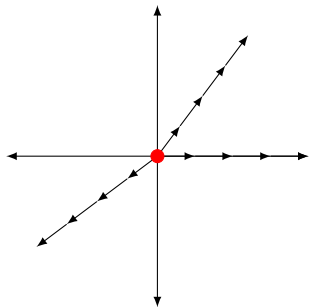
# Independence point

Candidate				
Race	<i>R</i>	<i>D</i>	<i>No</i>	
	167	167	167	500
	233	233	233	700
	400	400	400	1200



$$\mathcal{Y}_{z_p} \subset \mathbb{R}^{(R-1)(C-1)} \longleftrightarrow \mathbb{R}^{RC}$$

Diffeomorphic function  $g(y') : \mathbb{R}^{(R-1)(C-1)} \rightarrow \mathcal{Y}_{z_p}$

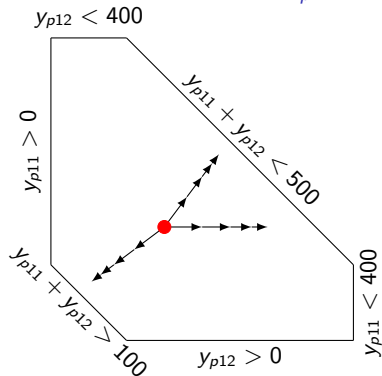


$\mathbb{R}^{(R-1)(C-1)}$

$g(y')$

$\rightarrow$

$\mathcal{Y}_{z_p} \subset \mathbb{R}^{(R-1)(C-1)} \longleftrightarrow \mathbb{R}^{RC}$



$$\frac{d\|g(y') - g(0)\|}{d\|y'\|} = y_{p11}y_{p12}(400 - y_{p11})(400 - y_{p12})(500 - y_{p11} - y_{p12}) \cdots$$

# Stochastic variational inference (Hoffman et al., 2013)

Goal: approximate **unnormalized** posterior density  $p(\theta, y|z) \propto p(z|\theta, y)p_{\pi}(\theta, y)$  with sampleable parametric distribution  $q_{\phi}(\theta, y)$ . (Called a **guide** in the pyro SVI package for python)

Maximize negative K-L divergence from guide to **normalized** posterior  $p(z|\theta, y)p_{\pi}(\theta, y)/p(z)$ :

$$E_{q_{\phi}} \left( \log \frac{p(z|\theta, y)p_{\pi}(\theta, y)}{q_{\phi}(\theta, y)p(z)} \right) < 0$$

$$E_{q_{\phi}} (\log[p(z|y)p(\theta, y)] - \log[q_{\phi}(\theta, y)] - \log(p(z))) < 0$$

$$E_{q_{\phi}} (\log[p(z|y)p(\theta, y)] - \log[q_{\phi}(\theta, y)]) < \log(p(z))$$

LHS is the **ELBO**; goal is to find  $\phi$  which maximizes it.

## ELBO terms

$E_{q_\phi}(\log[p(z|y)p(\theta, y)])$  is **energy** term. Maximized if  $q$  is a  $\delta$  (dirac mass) at MLE for  $(\theta, y|z)$ . Unboundedly negative if  $q$  has probability mass where  $p$  doesn't.

$E_{q_\phi}(-\log[q_\phi(\theta, y)])$  is **entropy** term. Maximized by making  $q$  diffuse. For example, if  $q$  is  $\mathcal{N}(\mu, \Sigma)$ , then this is inversely proportional to  $\det(\Sigma)$ . In principle unboundedly negative, but in practice, it's easier to control than energy term.

Together, they're maximized if  $q_\phi$  "imitates"  $p$ .



## El case

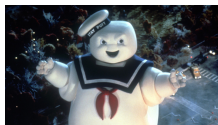
Reparameterize with  $y = g(y')$ , and approximate  $p(\theta, g(y'))$  using  $q_{\phi,z}(\theta, y')$ . ELBO over  $y'$  then becomes:

$$E_{q_{\phi,z}} \left( \log[p(\theta, g(y')) \det(J(g(y')))] - \log[q_{\phi,z}(\theta, y')] \right)$$

A common form of variational inference uses a “mean field” guide which factorizes across all parameters and latents; frequently, one that’s Normal in each dimension. This ignores the dependence induced by conditioning on the data; which is particularly strong in the case of EI.



## Choose the form of your posterior



Take  $q(\theta, y')$  to be a multivariate Normal, and assume that once the ELBO is maximized, its mode  $(\hat{\theta}, \hat{y})$  coincides with a mode of the posterior. What should its covariance matrix be?

There's an obvious way to approximate a twice-differentiable, unnormalized distribution with a Normal: a Laplace approximation.

That is, use the observed information matrix:

$$\mathcal{J}(\hat{y}', \hat{\theta}) = D^2 \left( \log[p(\hat{\theta}, g(\hat{y}')) \det(J(g(y')))] \right)$$

as the precision matrix of  $q$ .

# Computability

- ▶ Using pyro, a variational inference package for python.
- ▶  $\mathcal{J}(\hat{\mathbf{y}}', \hat{\theta})$  can be calculated using automated differentiation.
- ▶  $\mathcal{J}(\hat{\mathbf{y}}', \hat{\theta})$  is high-dimensional, but due to the structure of the model, sparse (block arrowhead format), so working with it is reasonably efficient. In practice, this means doing sampling “top down”  
(hyperparameters->parameters->hyperlatents->latents), one precinct at a time at the lower levels.

# Thanks

Thanks to Mira Bernstein, Luke Miratrix, Gary King

## Lower-D posterior (1)

Recall our model:

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

$$\alpha_c \sim \mathcal{N}(0, \sigma_\alpha) \quad \sigma_\alpha \sim \text{Expo}(5)$$

$$\beta_{r,c} \sim \mathcal{N}(0, \sigma_\beta) \quad \sigma_\beta \sim \text{Expo}(5)$$

$$\lambda_{r,c,p} \sim \mathcal{N}(0, \sigma_\lambda) \quad \sigma_\lambda \sim \text{Expo}(5)$$

Not only does the dimension of  $y$  grow linearly with the number of precincts  $P$ ; because of the latent  $\lambda$  parameters, the dimension of  $\theta$  does too. This is an issue both in estimating the ELBO and in maximizing it.

## Lower-D posterior (2)

Solution: replace  $\lambda_{p,r,c} \|_{c=1}^C$  with its MAP value conditional on all the variables connected to it (not conditionally independent):  $y_{p,r,c} \|_{c=1}^C$ ,  $\alpha_c \|_{c=1}^C$ ,  $\beta_{r,c} \|_{c=1}^C$ , and  $\sigma_\lambda$ . Because of the form of the model, this is available analytically, and we can trust that the Laplace approximation will still be reasonably good away from the MAP.

This is related to, but somewhat more aggressive than, the idea of “amortized variational inference” developed by Rezende and Mohammed (2015).