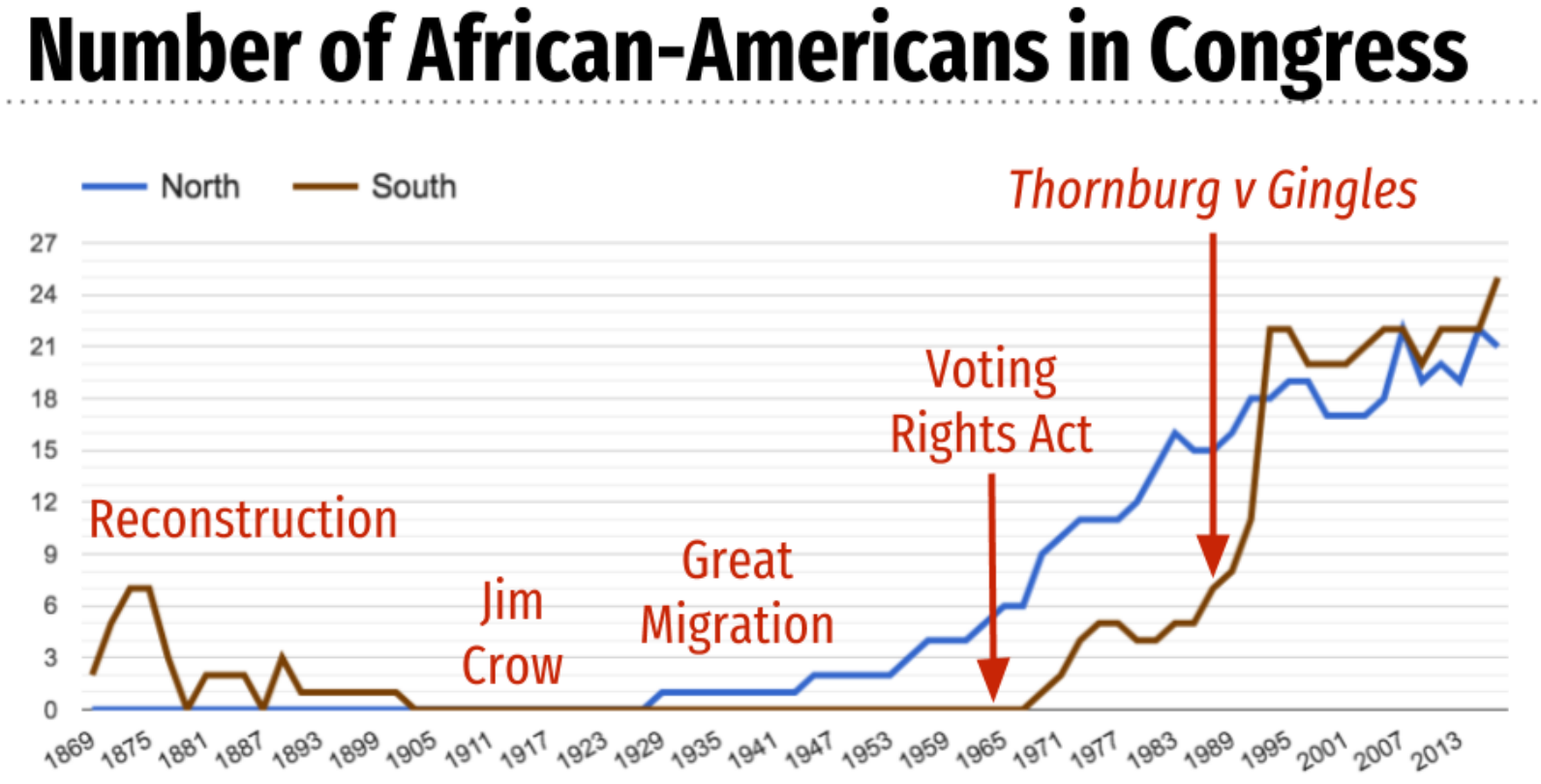
Flexible Ecological Inference using Variational Methods

Jameson Quinn (work with Mira Bernstein)

3/13/2019

Teaser

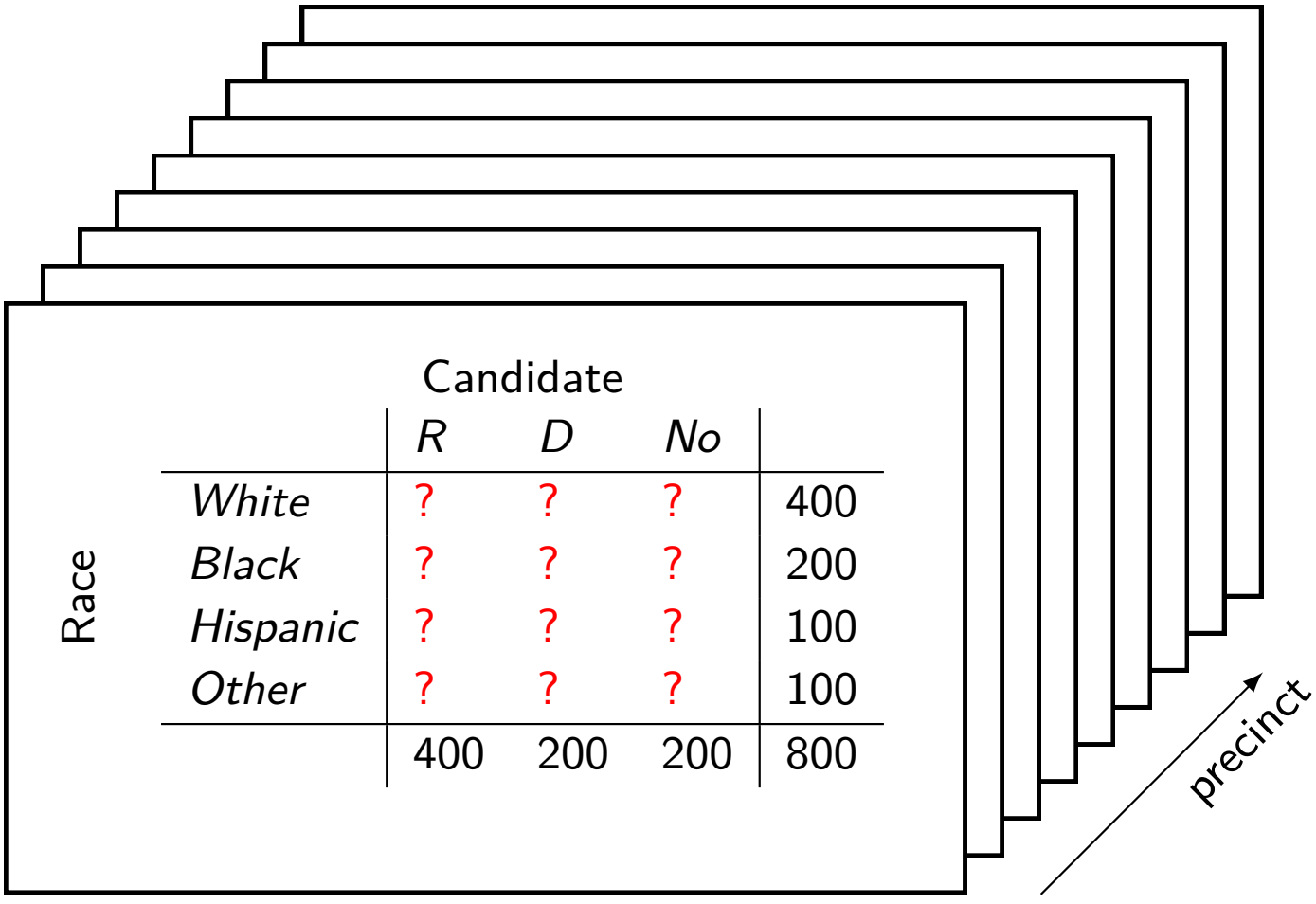

Number of African-Americans in Congress
North
South
Thornburg v Gingles
Voting Rights Act
Reconstruction
Jim Crow
Great Migration
27
24
21
18
15
12
9
6
3
0
1869
1875
1881
1887
1893
1899
1905
1911
1917
1923
1929
1935
1941
1947
1953
1959
1965
1971
1977
1983
1989
1995
2001
2007
2013

Thornburg v Gingles, 1986

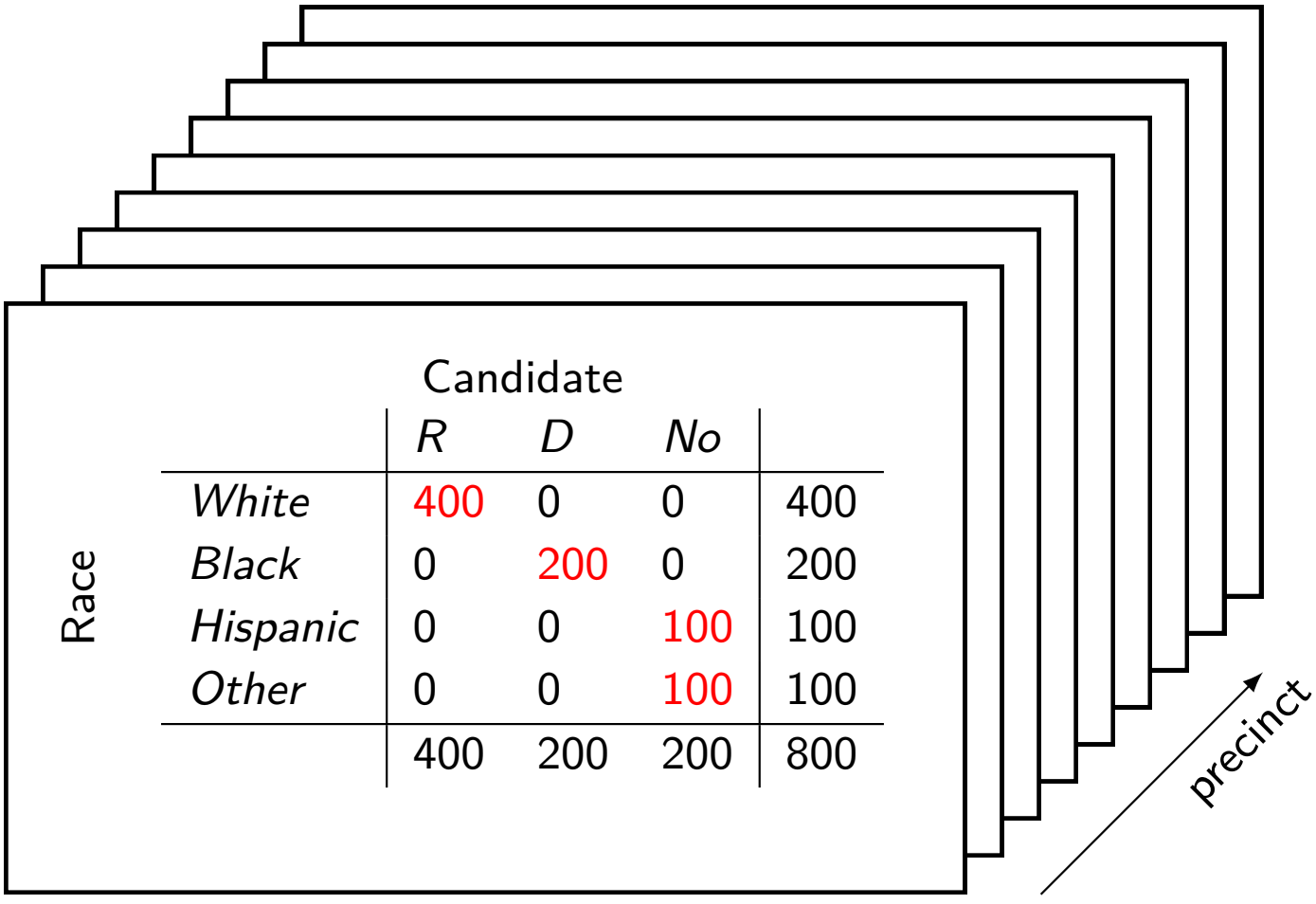
A majority-minority district must be created if:

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

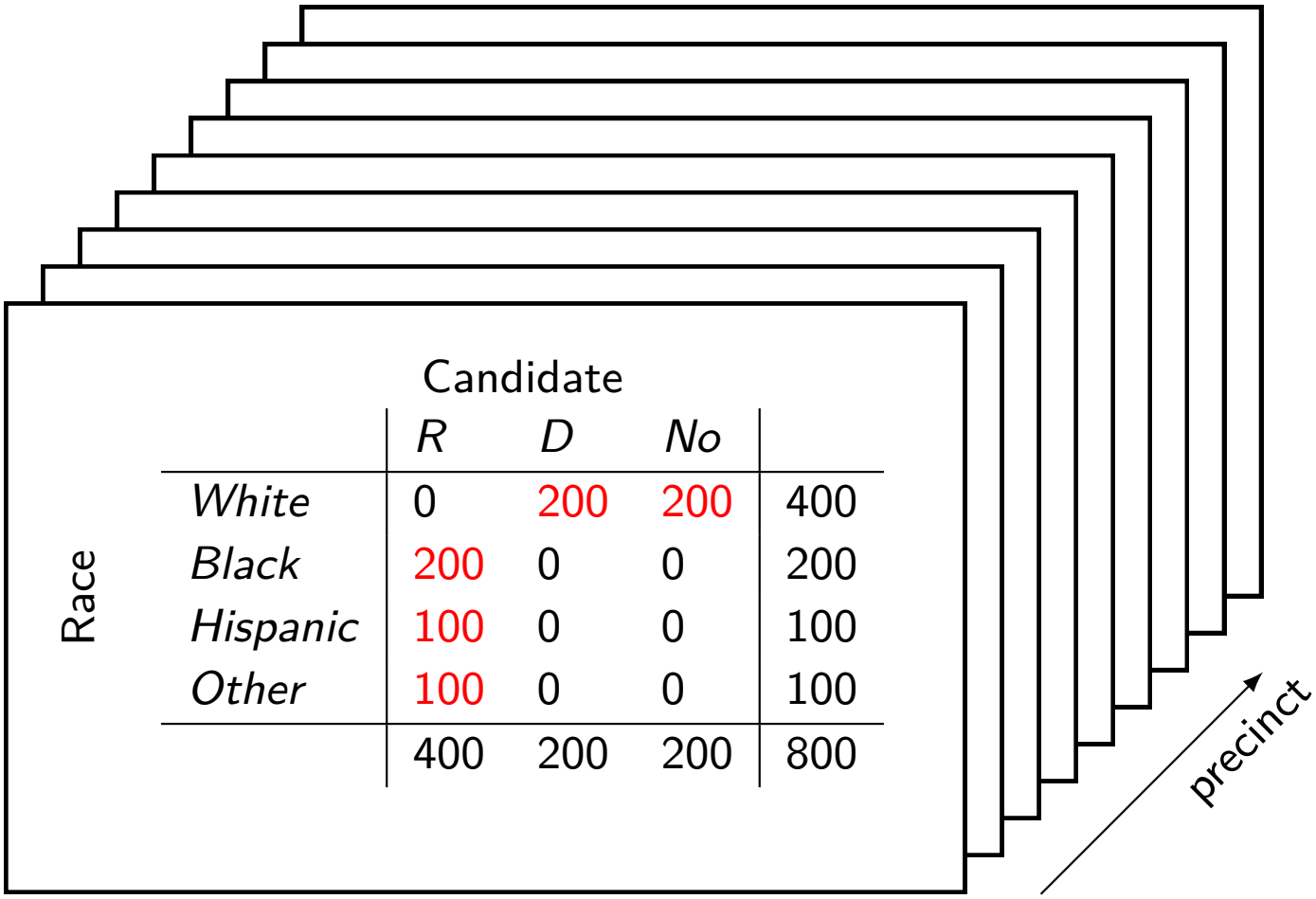


Majority=Majority?

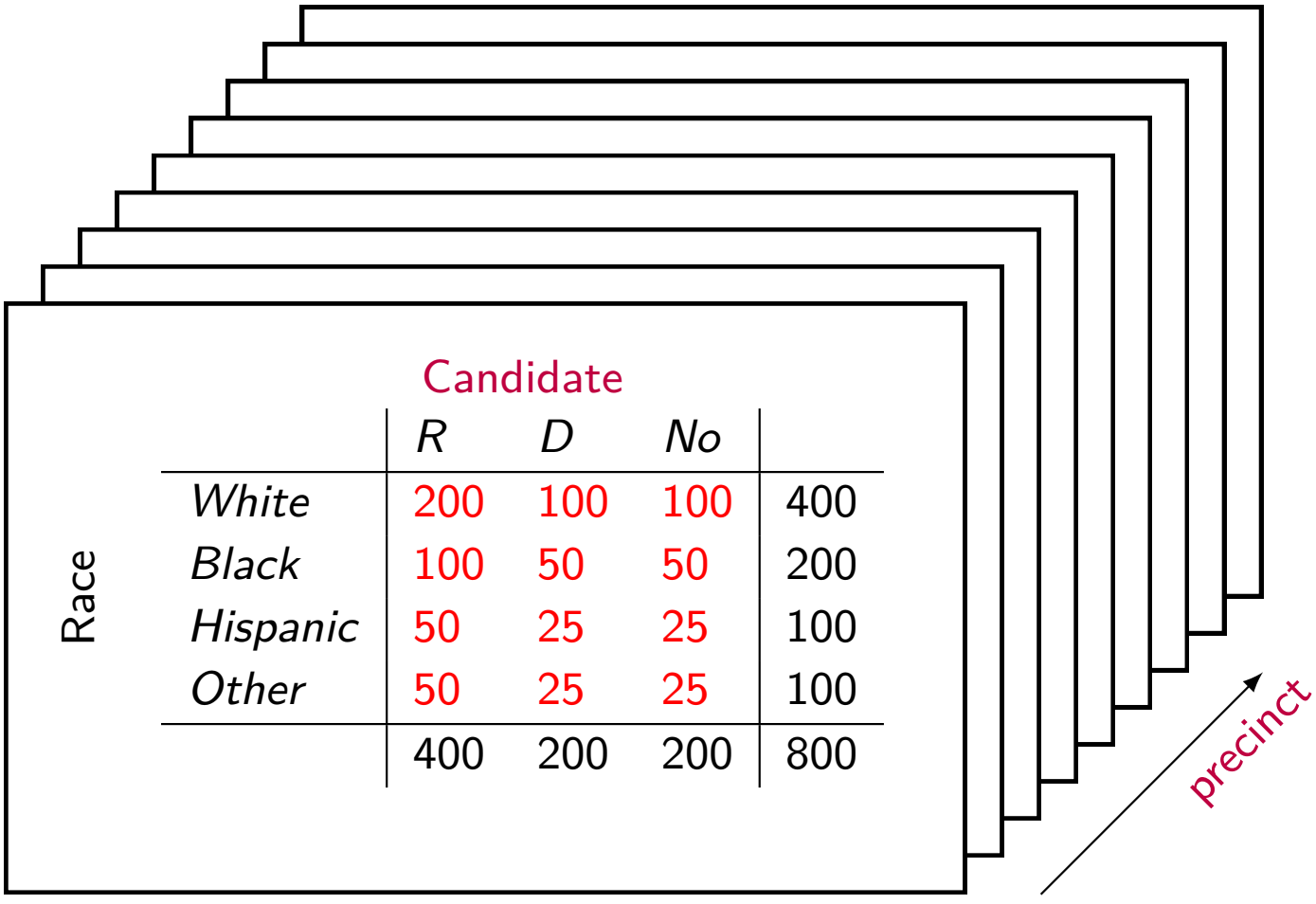


Race	R	D	No	
White	400	0	0	400
Black	0	200	0	200
Hispanic	0	0	100	100
Other	0	0	100	100
	400	200	200	800

Backwards?


Candidate
R
D
No
Race
White 0 200 200 400
Black 200 0 0 200
Hispanic 100 0 0 100
Other 100 0 0 100
400 200 200 800
precinct

Independence?

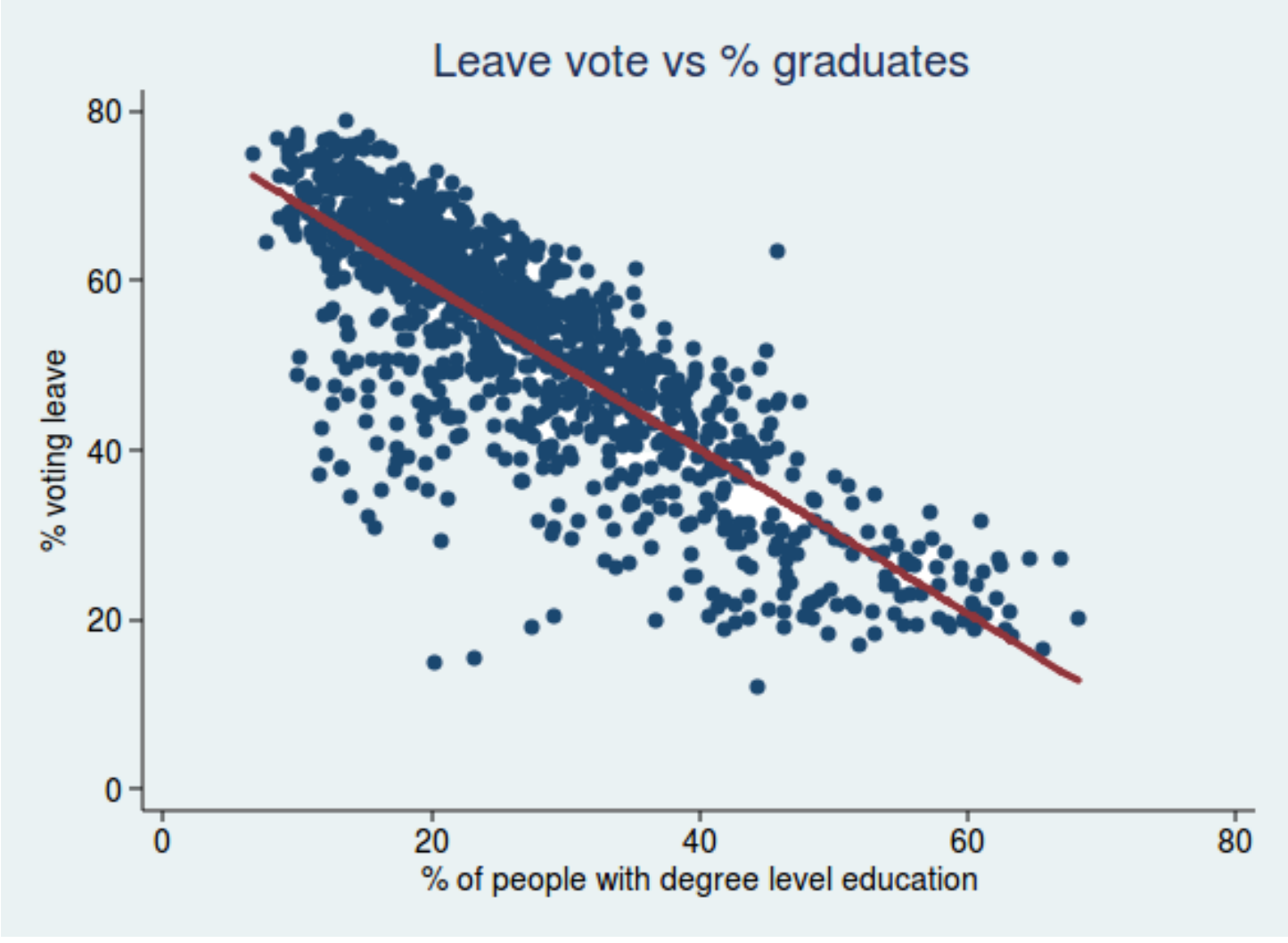


Race	R	D	No	
White	200	100	100	400
Black	100	50	50	200
Hispanic	50	25	25	100
Other	50	25	25	100
	400	200	200	800

Structure

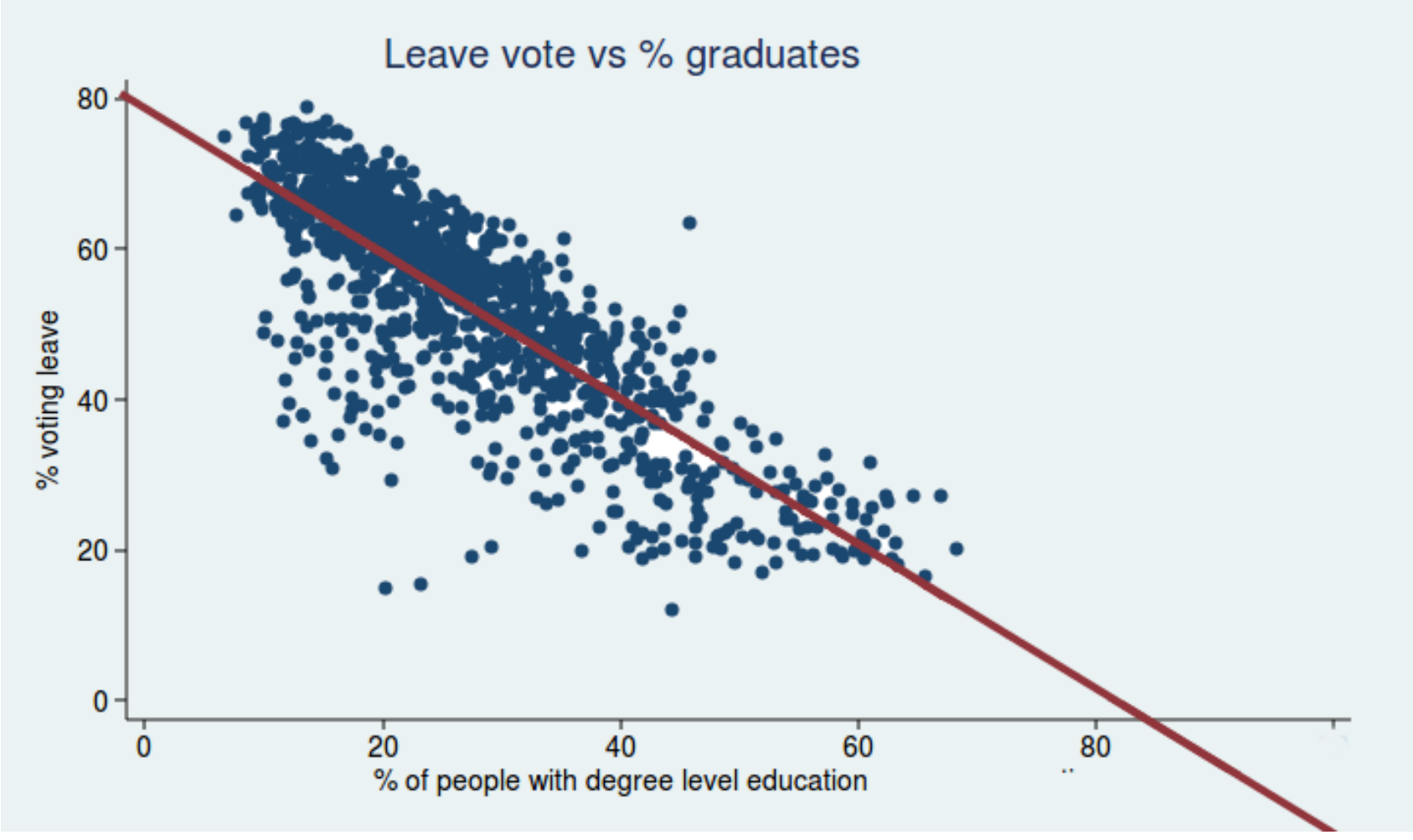
- Pose the ecological problem (done)
- Quick review of prior approaches
- A basic, extensible model for EI
- Why and how to reparameterize
- Review of variational inference
- Applying variational inference to EI
- Guide (aka variational distribution) based on observed information

Ecological regression (for 2×2 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:

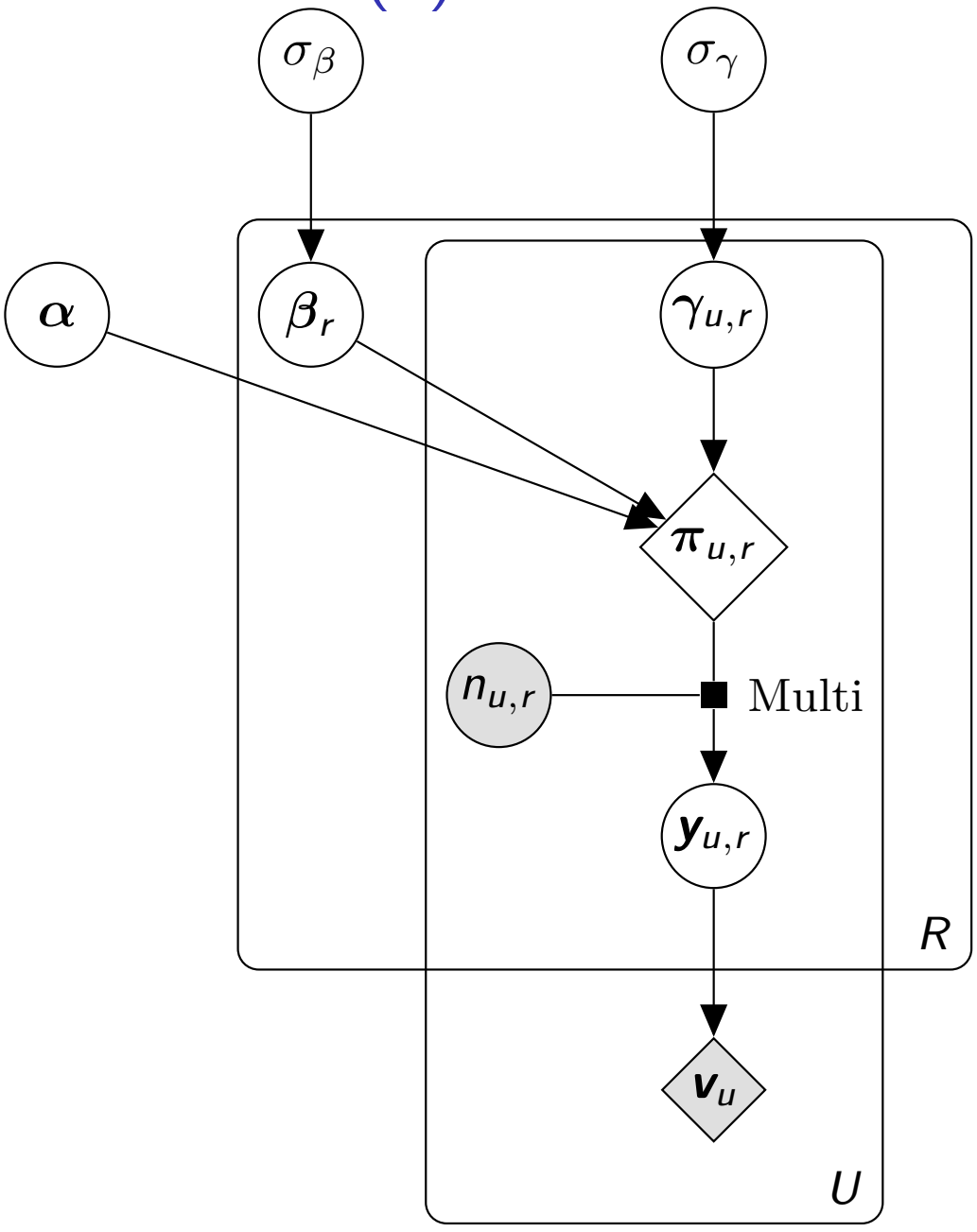
The RKJT model can, in principle, be extended to handle additional factors such as:

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

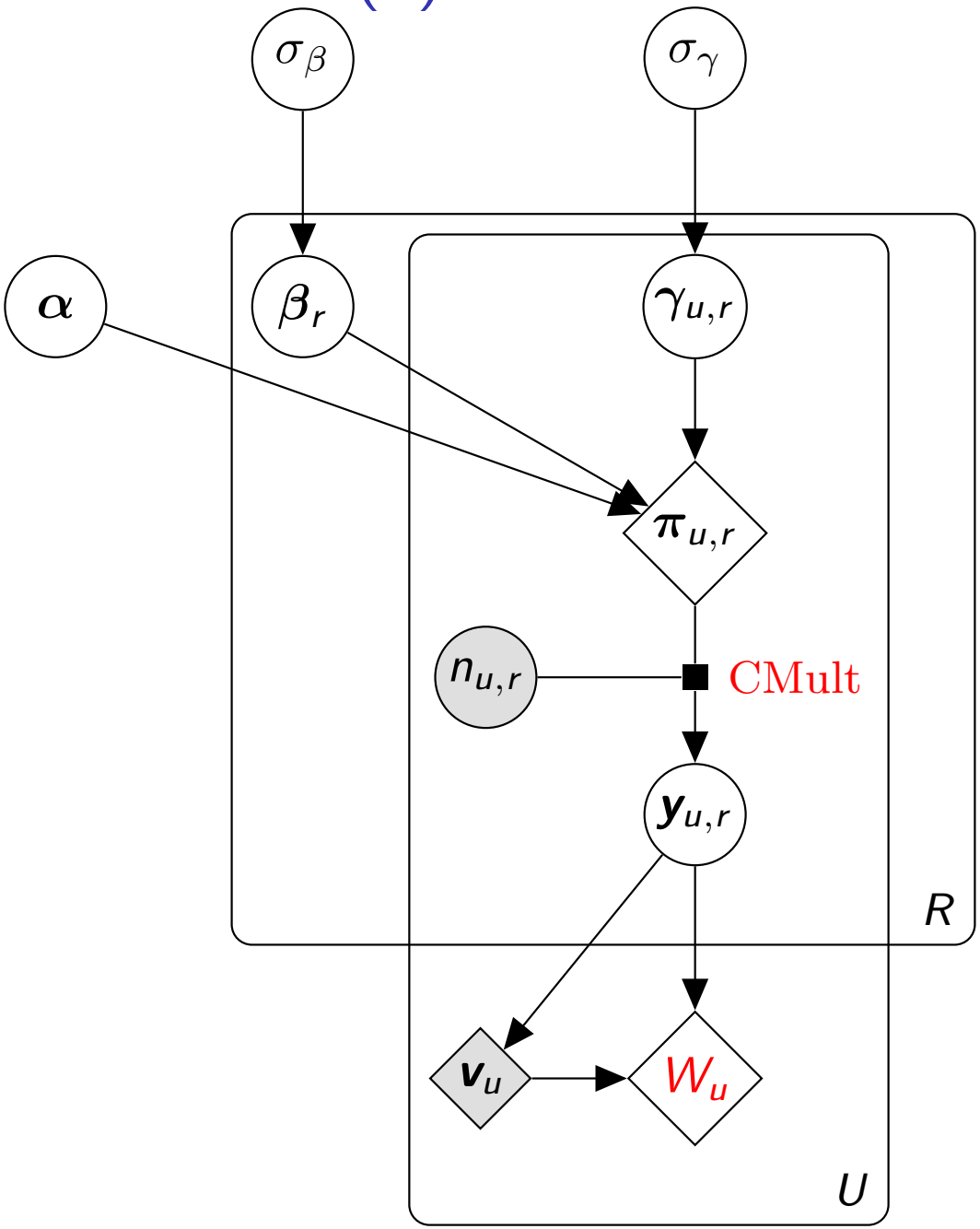
However:

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

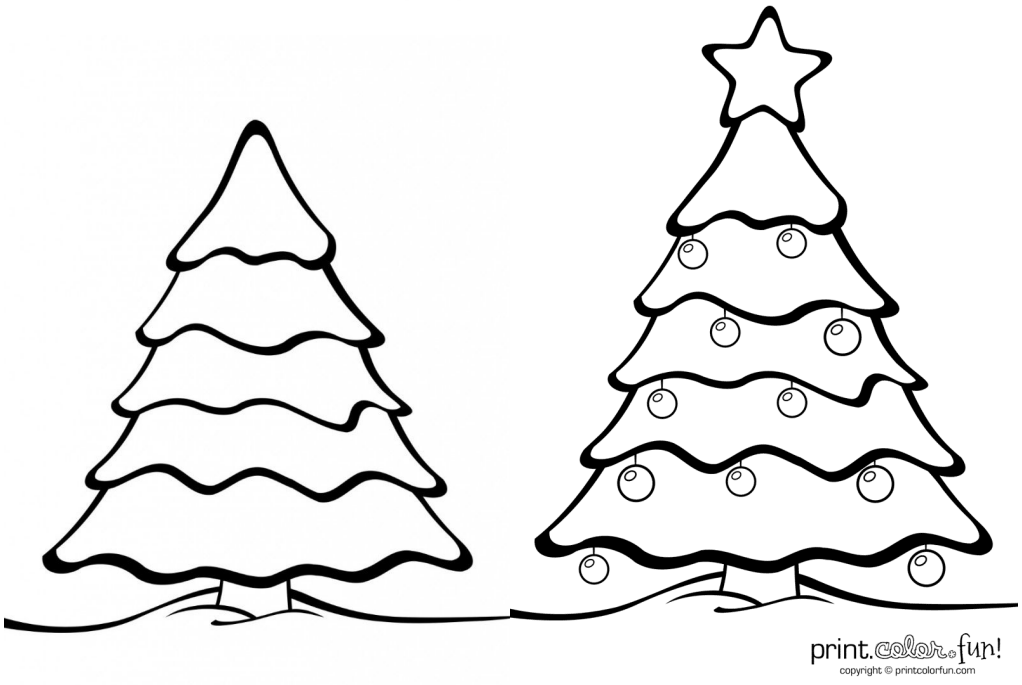
Flexible model (1)



Flexible model (2)



Flexible model (3)

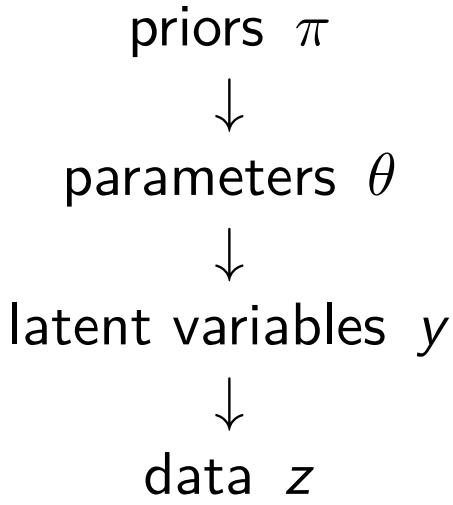


$$\vec{y}_{p,r} = y_{p,r,c}\|_{c=1}^{C} \sim \text{CMult}\left(n_{p,r}, \frac{exp(\alpha_c + \beta_{r,c} + \lambda_{r,c,p})\|_{c=1}^{C}}{\sum_{c=1}^{C} exp(\alpha_c + \beta_{r,c} + \lambda_{r,c,p})}\right)$$

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

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

Standard Bayesian approach (simplified)



Standard Bayesian approach (cont'd)

priors π	
$\downarrow$	(parameterizes distribution)
parameters θ	(unobservable quantities of interest; low-D)
$\downarrow$	(parameterizes distribution)
latent variables y	(unobservable nuisance parameters; high-D)
$\downarrow$	(parameterizes distribution)
data z	

Ecological Inference

priors π	
$\downarrow$	(parameterizes distribution)
parameters θ	(unobservable nuisance parameters; low-D?)
$\downarrow$	(parameterizes distribution)
latent variables y	(unobserved quantities of interest; high-D)
$\downarrow$	(deterministic function)
data z	

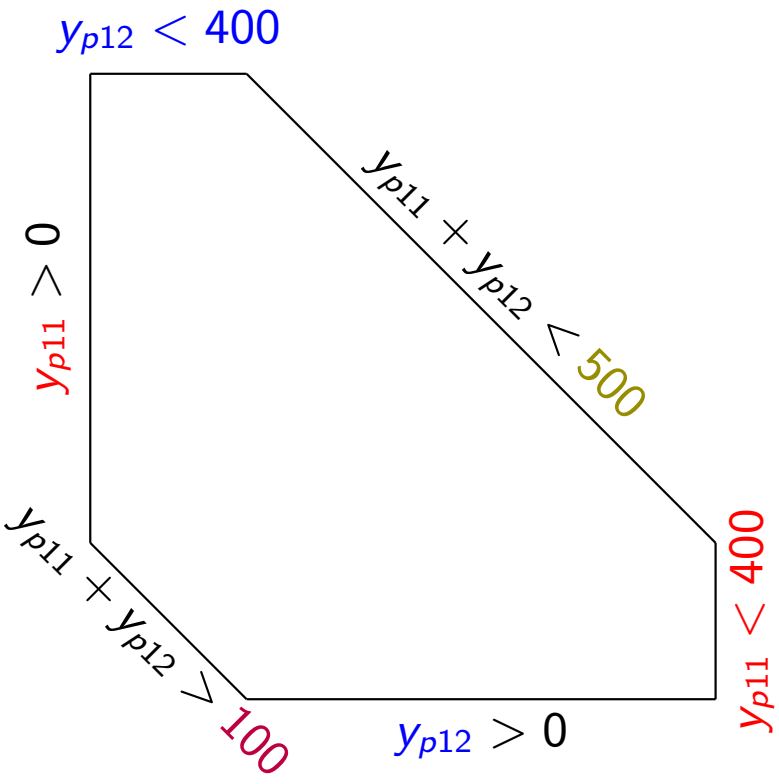
A likelihood from a deterministic function is an indicator function!

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 $\mathbb{R}^{RC}$.

Candidate

Race	R	D	No	
	y_{p11}	y_{p12}	y_{p13}	500
	y_{p21}	y_{p22}	y_{p23}	700
	400	400	400	1200

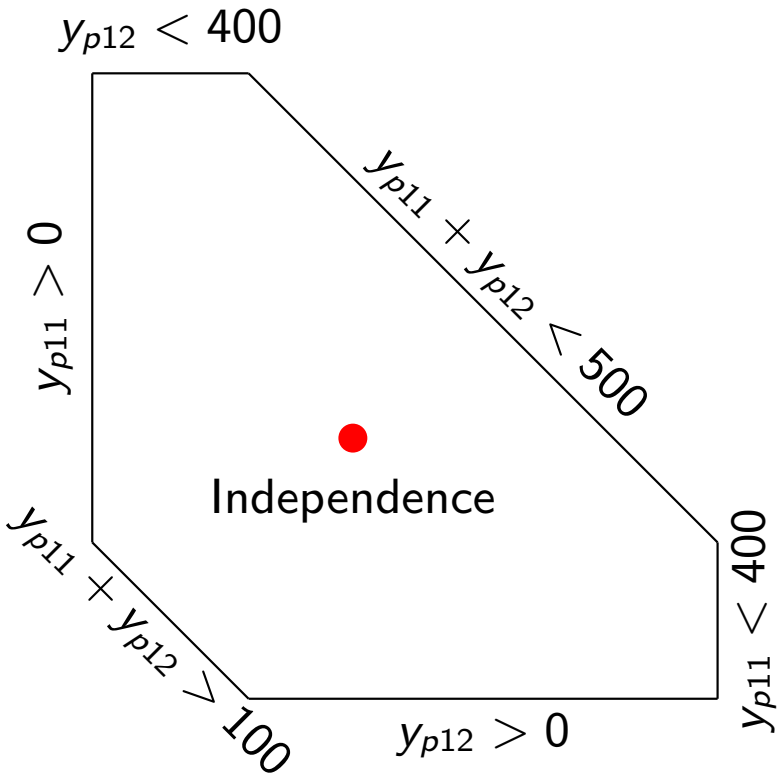


$$\mathcal{Y}_{z_p} \subset \mathbb{R}^{(R-1)(C-1)} \longleftrightarrow \mathbb{R}^{RC}$$

Independence point

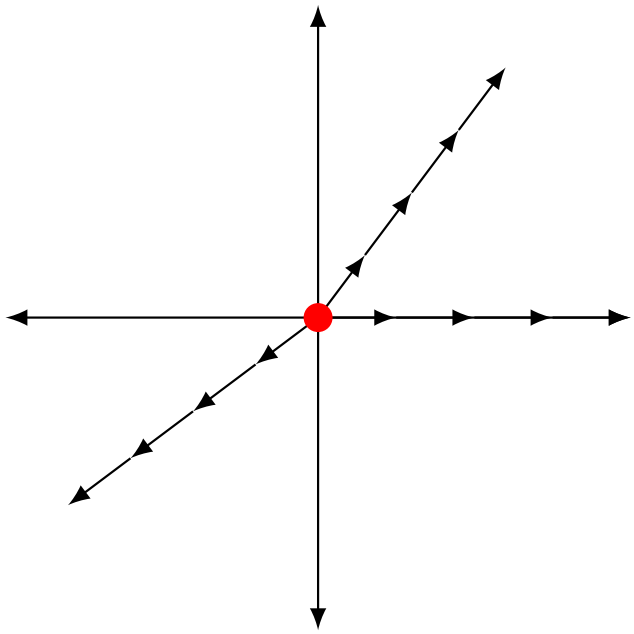
Candidate

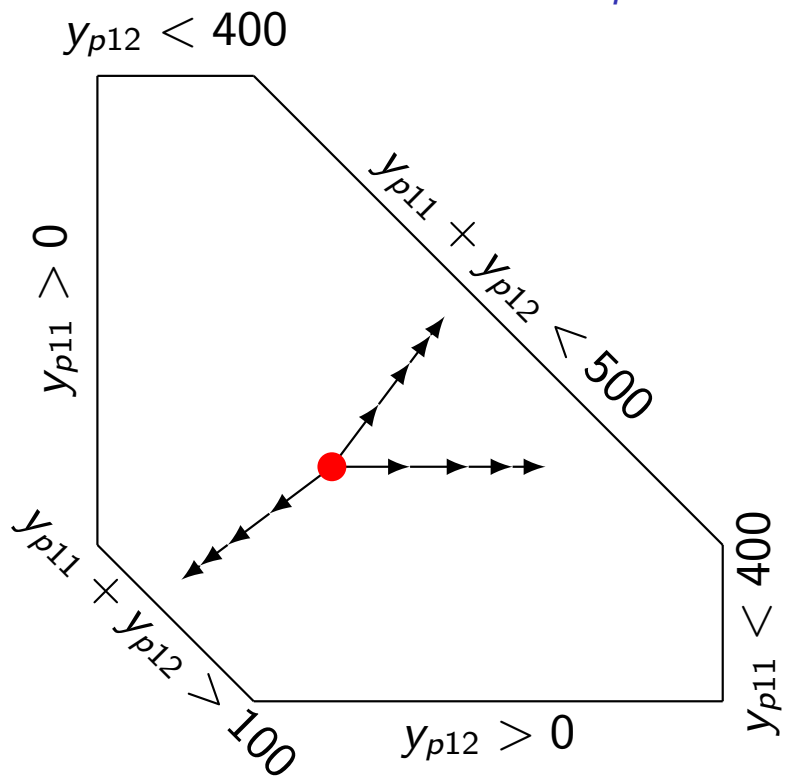
Race	*R*	*D*	*No*	
	167	167	167	500
	233	233	233	700
	400	400	400	1200



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

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





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

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

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 $p(z|\theta, y)p_\pi(\theta, y)/p(z)$:

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

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

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

LHS is the **ELBO**; goal is to find ϕ which maximizes it.

ELBO terms

$E_{q_\phi}(\log[p(z|y)p(\theta, y)])$ is **energy** term. Maximized if q is a δ (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 $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, 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_ϕ "imitates" p.

EI 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.

Laplace family

"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?

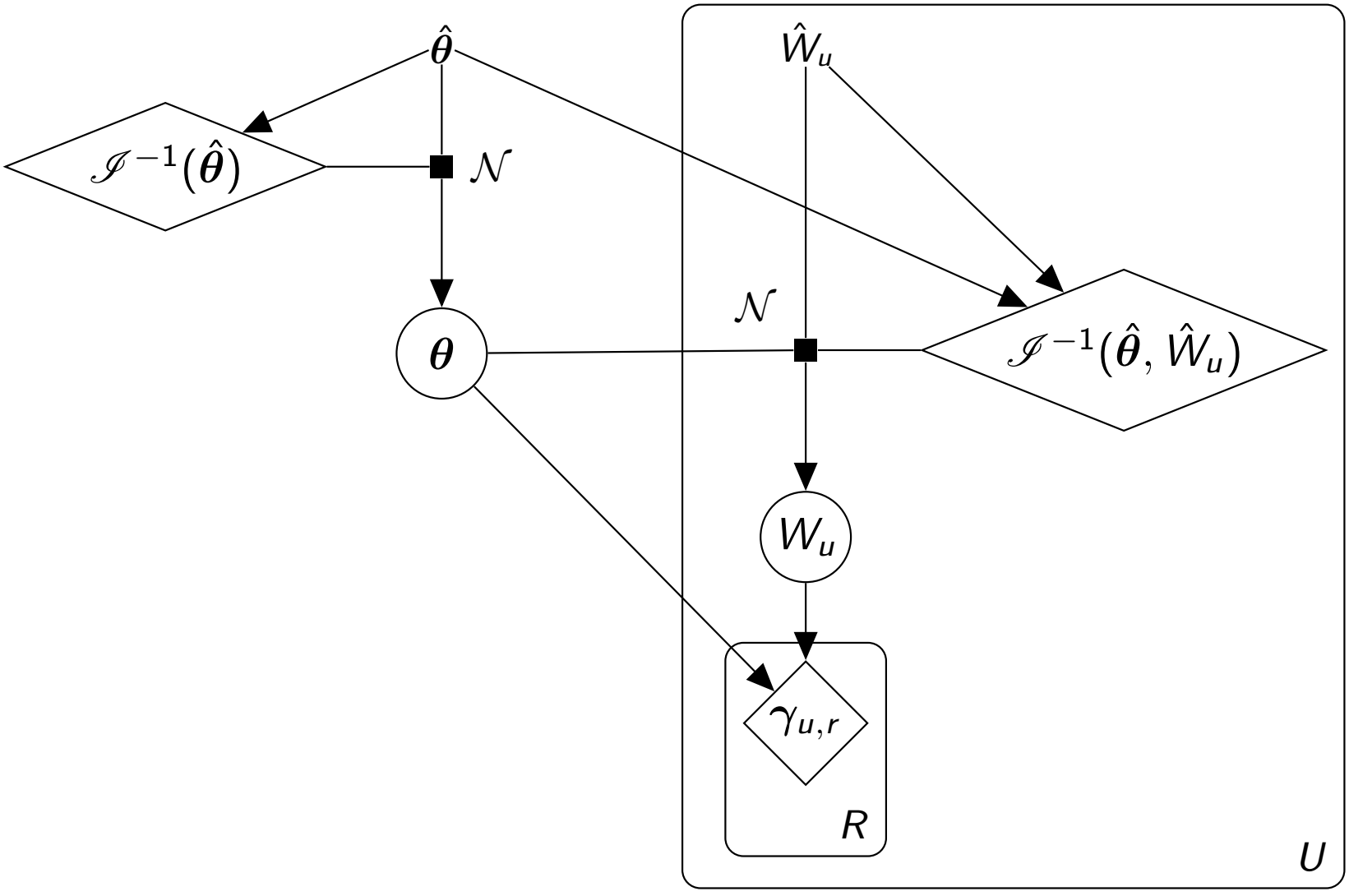
There's an obvious way to approximate a twice-differentiable, unnormalized distribution with a Normal: a Laplace approximation.

That is, use the observed information matrix:

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

as the precision matrix of q.

Graphical posterior



Computability

- Using pyro, a variational inference package for python.
- $\mathscr{I}(\hat{\boldsymbol{y}}', \hat{\theta})$ can be calculated using automated differentiation.
- $\mathscr{I}(\hat{\boldsymbol{y}}', \hat{\theta})$ 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:

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

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

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

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

Not only does the dimension of y grow linearly with the number of precincts P; because of the latent λ parameters, the dimension of θ 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 σ_λ. 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).