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

Jameson Quinn

12/9/19

Big picture overview

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

If Θ is high-D, any estimator of the denominator that amounts to numerical integration will fail; variance exponential in dim.

- Chapter 1: online data assimilation on spatiotemporal system
- Chapter 2: new method applicable to latent variable models
- Chapter 3: application of method, extension of existing models

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
Talk about Mira
the major motivating ideas (idea of Laplace family, idea of applying VI to EI) come from me
I did all the coding in Pyro
otherwise Mira is equal collaborator and coauthor

we checked that this is allowed

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 posterior with a guide distribution $q_\phi(\boldsymbol{\theta})$ and 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:

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

Emphasize conceptual part (minimizing KL-divergence)
Be sure to be clear about difference between model parameters and guide parameters
Talk about why entropy term is important

Mention EUBO

Pyro

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

- Automatic differentiation via PyTorch ML
- Stochastic optimization
- BBVI seems empirically robust

Explain automatic differentiation

All the impressive software engineering I had to do

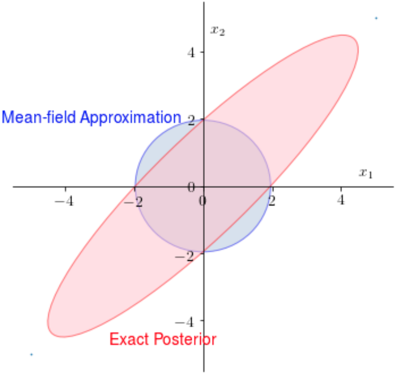
Problem with existing (mainstream) VI method

This talk will focus on MVN guide families.

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

can't capture posterior correlations

systematically underestimates posterior marginals



Introduce Laplace family VI (1)

Among MVN 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

Introduce Laplace family VI (2)

Let's guarantee that the family contains the Laplace approximation around any posterior mode.

Define covariance matrix using observed information of posterior; negative of Hessian of unnormalized log-density:

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

Boosting

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

- 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 (Fang, 2008)

Furthermore, we can parametrize f to create a boosting family f_{ψ}, for $\psi_i > 0$, s.t.:

- Each dimension of $\boldsymbol{\psi}$ corresponds to a model parameter
- As $\boldsymbol{\psi} \to \vec{\mathbf{0}}$ from above, $f(\mathcal{I}_p) \to \mathcal{I}_p$ if $\mathcal{I}_p$ already p.d.

Explain why boosting family is better than just boosting function
Version of thesis sent previously has quasi-boosting which we're no longer using

Citation for method that we're actually using

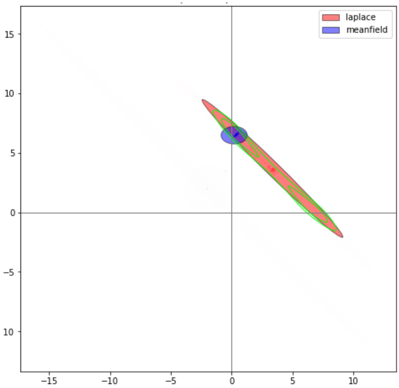
Formal definition of Laplace family (1)

Let $p(\boldsymbol{\theta})$ be a (possibly unnormalized) probability distribution on $\mathbb{R}^d$. Let $\Theta \subseteq \mathbb{R}^d$, $\Psi \subseteq \mathbb{R}_+^D$, and let f_Ψ be a boosting family.

The Laplace guide family $\mathcal{L}_{\Theta\times\Psi}(p, f_\Psi)$ is the set of d-dimesnional normal distributions $\{q_{\boldsymbol{\theta}^*,\boldsymbol{\psi}} : \boldsymbol{\theta}^* \in \Theta,\ \boldsymbol{\psi} \in \Psi\}$, where $q_{\boldsymbol{\theta}^*,\boldsymbol{\psi}}$ has mean $\boldsymbol{\theta}^*$ and precision matrix $f_{\boldsymbol{\psi}}(\mathcal{I}_p(\boldsymbol{\theta}^*))$.

Toy model results

Comparison of Laplace family fit with meanfield fit on simple model with a bimodal posterior:



Simple model with bimodal posterior

Shows several things: Importance of covariance; case where laplace of MAP isn't optimal; case where boosting is necessary, and boosting family is better than boosting function

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

A latent variable model has 3 core elements:

- Global parameters: $\boldsymbol{\gamma} \in \Gamma \cong \mathbb{R}^g$,
- “iid” latent parameter vectors: $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_N \in \Lambda \cong \mathbb{R}^l$
- Observation vectors: $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, also independent conditional on globals and relevant locals.

In other words,

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

This is why we need a high-dimensional solution.

Block arrowhead matrices

$$\mathcal{I}_p(\boldsymbol{\theta}^*) = \begin{pmatrix} G & C_1 & C_2 & \dots & C_N \\ C_1^T & U_1 & 0 & \dots & 0 \\ C_2^T & 0 & U_2 & \dots & 0 \\ \vdots & 0 & 0 & \ddots & 0 \\ C_N^T & 0 & 0 & \dots & U_N \end{pmatrix}$$

Useful fact: if this is viewed as a precision matrix, then the "marginal precision" $([\mathcal{I}_p(\boldsymbol{\theta}^*)^{-1}]_{\Gamma,\Gamma})^{-1} = G - \sum_i C_i U_i^{-1} C_i^T$

Call this marginal precision matrix $\mathcal{G}_p(\boldsymbol{\theta}^*)$.

easy to sample from and easy to boost.

Additional methods for LV models 1: SVI

Two methods that are useful with LV models. Conceptually independent, but go well together. First is:

SVI (Stochastic Variational Inference)

Basic idea: since unnormalized log density is a sum of unit-level terms, use only a randomly-sampled subset of those terms. For a given sampling scheme (possibly weighted), there are obvious estimators for:

- Posterior log density; unbiased
- Hessian thereof. Up to boosting, unbiased for both conditional precision G and marginal precision $\mathcal{G}_p$.
- ELBO and ELBO gradient wrt guide parameters, using the above. Not unbiased, even without boosting; but seem to work well.

Mention weights, say we haven't yet implemented them anywhere

think about exactly what you want to say about its qusi-unbiasedness (have hidden slides with slide with details)

Additional methods for LV models 2: amortization

An amortized guide family is one that's restricted to a subspace where the latent variables are a function of the globals.

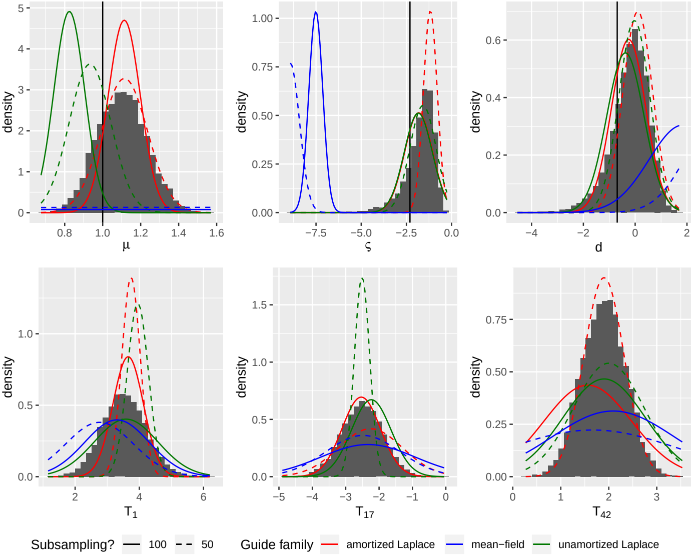
The aim is to find an analytic a priori function that sets the latents to their conditional MAP values, under the assumption that this is the value that will approximately maximize the ELBO given the globals.

If you can get into the right neighborhood, you can take 1 step of Newton's method for free.

guide parameters are now just gamma, psi-gamma, psi-lambda

Note that it's MAP not MLE

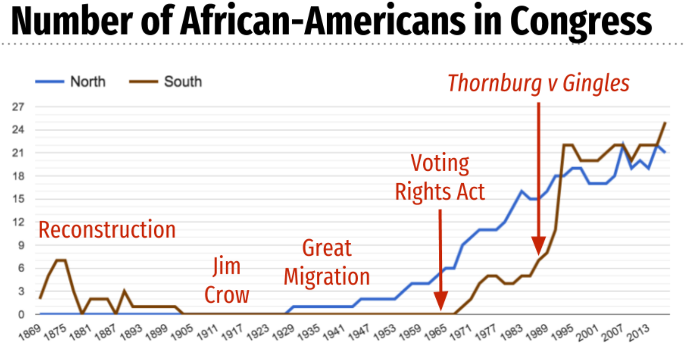
Ch 1 Results



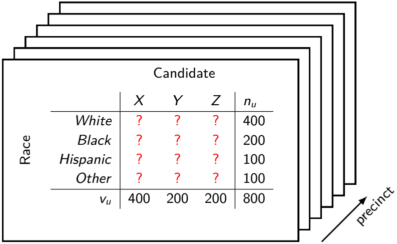
Changes: df; boosting

Ch 2: Ecological inference. Relevance: Thornburg v. Gingles

Since the 1986 Supreme Court decision in Thornburg v. Gingles, EI is key to proving the need for redistricting under the Voting Rights Act. Which seems important:



EI: Problem statement (1)

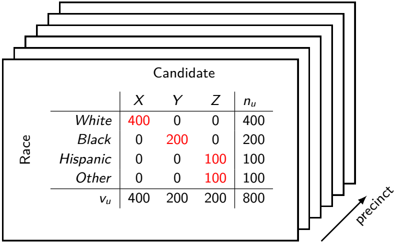


	X	Y	Z	n_u
White	?	?	?	400
Black	?	?	?	200
Hispanic	?	?	?	100
Other	?	?	?	100
v_u	400	200	200	800

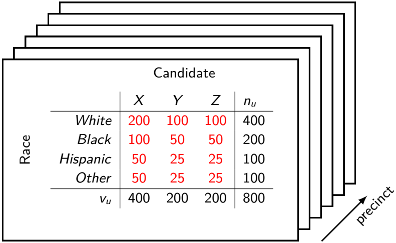
Mainly comes up in voting rights cases, so I'll talk about it in this setting.

Z could represent not voting.

EI: Problem statement (2)

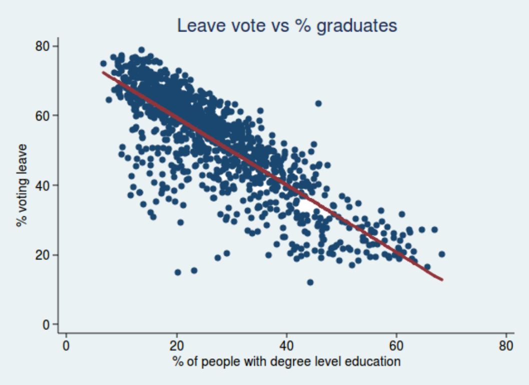


		Candidate			
		X	Y	Z	n_u
Race	*White*	400	0	0	400
	Black	0	200	0	200
	Hispanic	0	0	100	100
	Other	0	0	100	100
	v_u	400	200	200	800



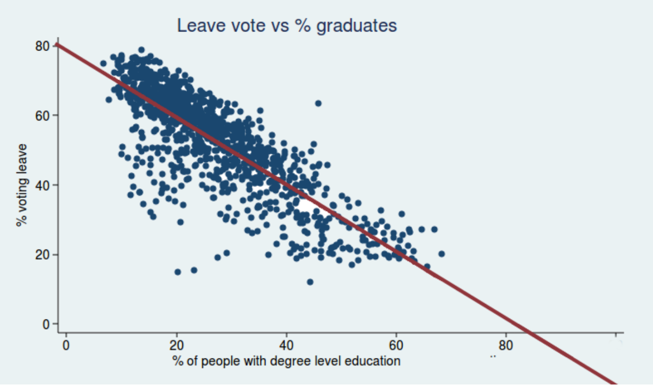
		Candidate			
		X	Y	Z	n_u
Race	*White*	200	100	100	400
	Black	100	50	50	200
	Hispanic	50	25	25	100
	Other	50	25	25	100
	v_u	400	200	200	800

History of attempted solutions: ER



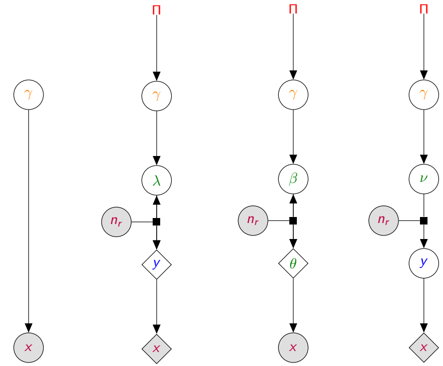
Brexit voting data. (Example by Adam Jacobs.)

History of attempted solutions: ERrrrr. . .



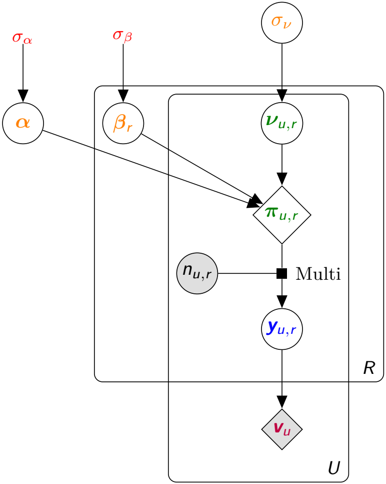
Brexit support: -16% of those with a degree???

Comparison of models



Level	ER (1953)	King (1997)	RJKT (2001)	Us (2019)
Prior		Π	Π	Π
Global voting propensities (for each race)	γ	γ	γ	γ
Precinct-level variation (for each race)		λ n_r	β n_r	ν n_r
Vote totals by race & candidate (unobserved)		y	θ	y
Vote totals by candidate (observed)	x	x	x	x
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.	

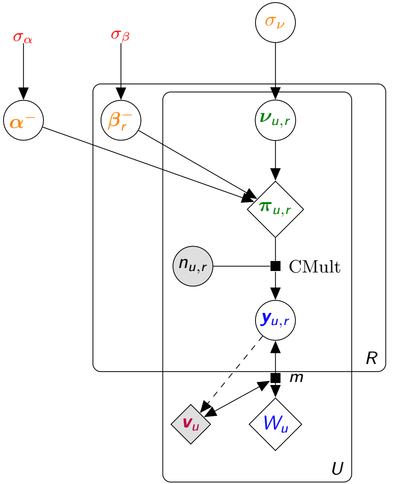
Our model



Talk briefly about how you could add other Christmas tree ornaments

Talk about why this is hard to make a guide for

Modified model

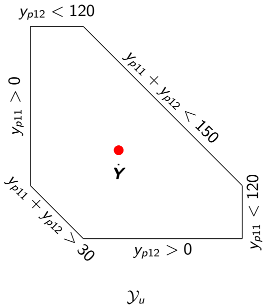


Cmult; Polytopize: ae smooth bijective map from R^n to polytope
Pseudovoters because of boundary issues that arise from Cmult and polytopize

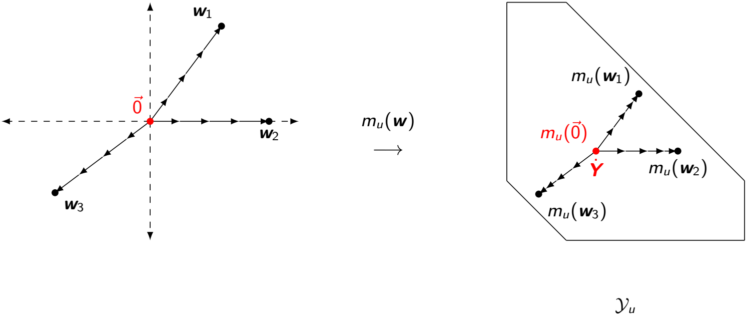
All of these make the model itself slightly less-realistic, but make VI work.

Polytopize (1)

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



Polytopize (2)



Guide, with amortization

too ugly

Testing our EI on simulated data

Describe how we got the simulated NC data actual demographics realistic alpha and beta we get to experiment sigma_nu

EI results (1)

Show updated tables from paper

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 (2)

Discussion/future work (Ch. 3)

- Including the covariate
- 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 (use Ch 3 as example)
  - maybe some theory to help choose sample size for SVI
- Replace normal with T in guide

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

Thanks

Directory of extra slides

Non-meanfield prior work

Just the list from the paper Give example of actual theorem you can prove when you have conjugate model structure

Details on toy model

Just the model

More on block-arrowhead matrices

Formulas for boosting Formulas for sampling

(basically just the stuff in the appendix)

What we expect from subsampling

HARD

Details of ECHS

The actual model Result tables

More details on RJKT

Possible extensions to our EI model

Boundary issues with polytope; pseudovoters

HARD

How our EI amortization works (1)

Which variables are we amortizing: Y, nu, sigma_nu Steps: Find approximate mode of p(Y|alpha, beta, nu) constrained to lie on polytope (this is linear algebra plus stirling's approximation) 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(nu, sigma_nu| gamma, W) Newton's method (for free!!)

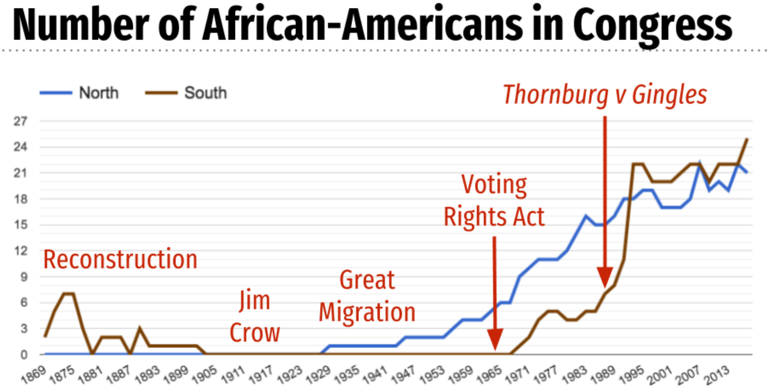
How our EI amortization works (2)

Details on how we get nu and sigma_nu

More EI results

END DEFENSE, START OLD PRESENTATION

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

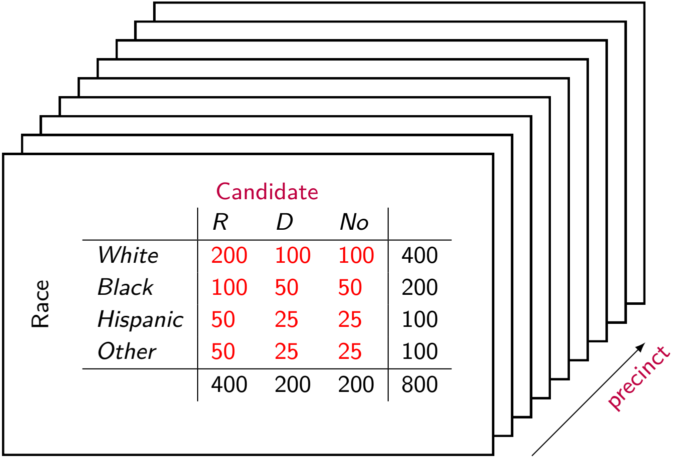
Thornburg v Gingles, 1986

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

Independence?

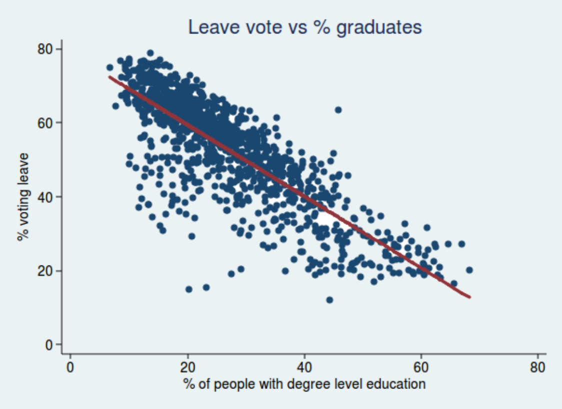


		Candidate			
		R	*D*	*No*	
Race	*White*	200	100	100	400
	Black	100	50	50	200
	Hispanic	50	25	25	100
	Other	50	25	25	100
		400	200	200	800

Structure

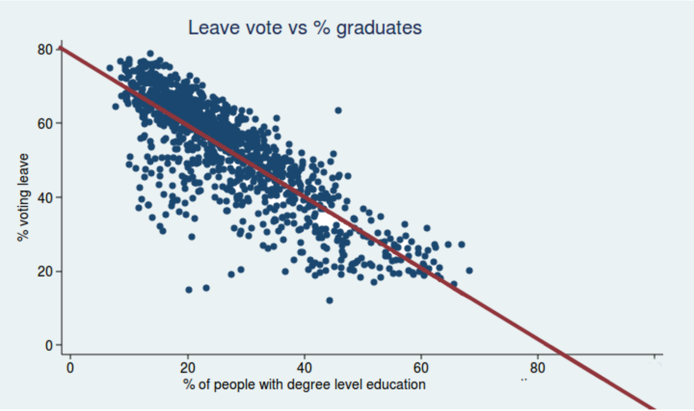
- Pose the ecological problem (done)
- Quick review of prior approaches
- A basic, extensible model for EI
- Why and how to reparameterize
- Review of variational inference
- Applying variational inference to EI
- Guide (aka variational distribution) based on observed information

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:

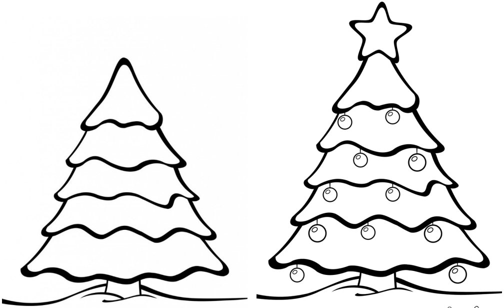
- inter-row or inter-column correlations
- covariates
- multiple elections
- exit polling data
- etc.

However:

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

Flexible model (1)

Flexible model (3)



print.color.fun!
copyright © printcolorfun.com

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

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

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

Standard Bayesian approach (simplified)

priors π

↓

parameters θ

↓

latent variables y

↓

data z

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

↓ (parameterizes distribution)

parameters θ (unobservable nuisance parameters; low-D?)

↓ (parameterizes distribution)

latent variables y (unobserved quantities of interest; high-D)

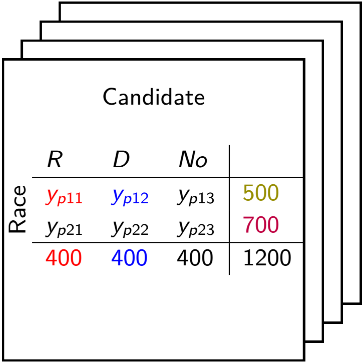
↓ (deterministic function)

data z

A likelihood from a deterministic function is an indicator function!

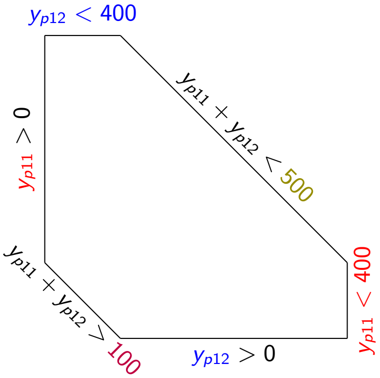
Ecological case

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

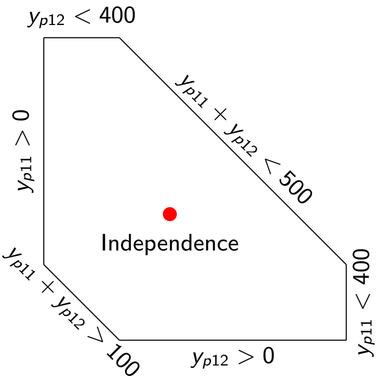
Race	R	D	No	
	y_{p11}	y_{p12}	y_{p13}	500
	y_{p21}	y_{p22}	y_{p23}	700
	400	400	400	1200



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

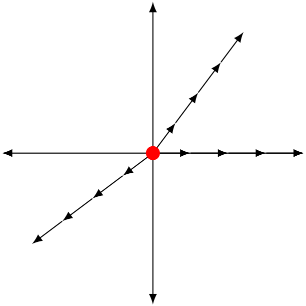
Independence point

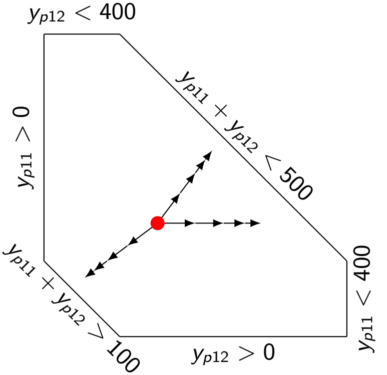
	Candidate			
	R	D	No	
Race	167	167	167	500
	233	233	233	700
	400	400	400	1200



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

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

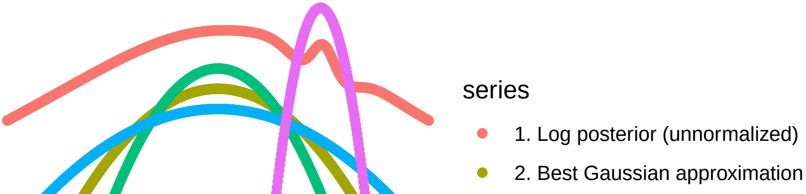
$E_{q_\phi}(\log[p(z|y)p(\theta, y)])$ is **energy** term. Maximized if q is a δ (dirac mass) at MLE for $(\theta, y|z)$. Unboundedly negative if q has probability mass where p doesn't.

$E_{q_\phi}(-\log[q_\phi(\theta, y)])$ 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.

```
## Warning: package 'ggplot2' was built under R version 3.6
```

```
## Warning: package 'data.table' was built under R version
```



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

staypuft *"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?

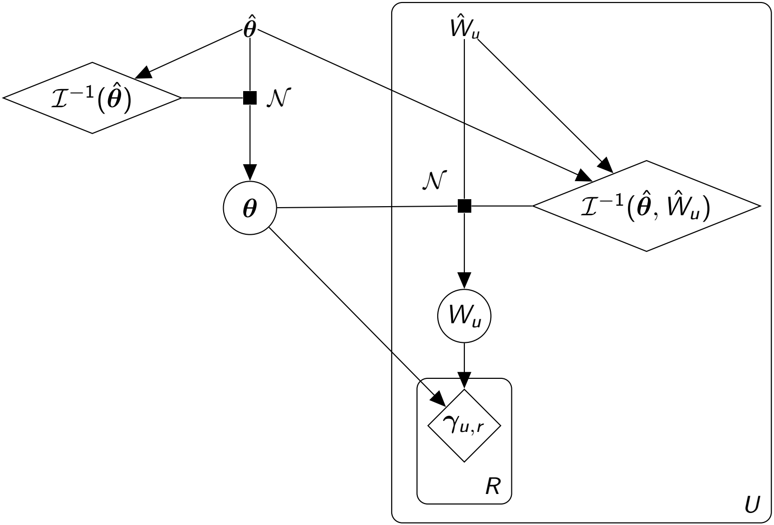
There's an obvious way to approximate a twice-differentiable, unnormalized distribution with a Normal: a Laplace approximation.

That is, use the observed information matrix:

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

as the precision matrix of q.

Graphical posterior


$\hat{\boldsymbol{\theta}}$
$\mathcal{I}^{-1}(\hat{\boldsymbol{\theta}})$
$\mathcal{N}$
$\boldsymbol{\theta}$
$\hat{W}_u$
$\mathcal{N}$
$\mathcal{I}^{-1}(\hat{\boldsymbol{\theta}}, \hat{W}_u)$
W_u
$\gamma_{u,r}$
R
U

Computability

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