

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

Jameson Quinn

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

Variational Inference

Approximate posterior with a guide distribution $q_{\phi}(\boldsymbol{\theta})$ and choose ϕ to minimize KL:

$$\hat{\phi} = \operatorname{argmin}_{\phi} [D_{\text{KL}} (q_{\phi}(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta}|\mathbf{x}))].$$

Equivalent to maximizing ELBO:

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

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 ELBO

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.....with picture!

Introduce Laplace family VI (1)

Among MVN guide families: * Set of all normals, unrestricted covariance, is too big * Meanfield subset 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:

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

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 ψ corresponds to a model parameter * As $\psi \rightarrow \vec{0}$ from above, $f(\mathcal{I}_p) \rightarrow \mathcal{I}_p$ 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

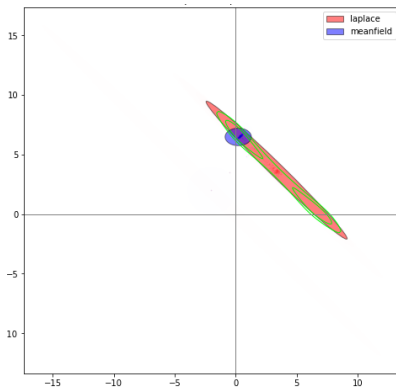
Formal definition of Laplace family (1)

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 -dimensional normal distributions $\{q_{\boldsymbol{\theta}^*, \psi} : \boldsymbol{\theta}^* \in \Theta, \psi \in \Psi\}$, where $q_{\boldsymbol{\theta}^*, \psi}$ has mean $\boldsymbol{\theta}^*$ and precision matrix $f_\psi(\mathcal{I}_p(\boldsymbol{\theta}^*))$.

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?)

Define, including defining Gamma and Theta This is why we need a high-dimensional solution.

Block arrowhead matrices

Show block-arrowhead matrix, say it's easy to sample from and easy to boost (don't write anything down for this, just point)

Additional methods for LV models: subsampling

Explain subsampling 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) Explain amortization: restricting to subspace of the guide where latents are approximate MLEs relative to globals, guide parameters are now just γ , ψ_γ , ψ_λ (possibly mention Newton's method, but Mira thinks not) (Say that they go together nicely, but don't go into detail)

Additional methods for LV models: boosting

Ch 1 Results

Improved version from thesis Very brief explanation only, but mention transforming parameters

Ch 2: Ecological inference. Introduction

Mainly comes up in voting rights cases, so I'll talk about it in this setting Explain problem in the abstract by showing pictures of matrices and talking about them (you don't need much else on this slide)

El: example matrices

Relevance: Thornburg v. Gingles

Mention Gingles SCOTUS case briefly show Gingles graph

History of attempted solutions: ER

Picture of Brexit ER Explain problem with this method

King's El: 2x2 case (1)

Describe Point out innovation of looking at latents rather than globals

King's El: 2x2 case (2)

Rosen, Jiang, King, & Tanner: RxC case (1)

Explain model

Rosen, Jiang, King, & Tanner: RxC case (1)

Limitations of RJKT...

it's cheating (applies racial constraints at wrong level of hierarchy, because they don't know how to apply both vote-total and racial constraints at same level; I'll solve this with polytopize) hard to extend model (give examples of how you might extend it) relies on MCMC, so if you extend by a lot will be too slow

... and our contribution

No cheating (polytope) More flexible model VI instead of MCMC
(make this slide parallel to previous one)

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: a smooth map from \mathbb{R}^n to polytope

Pseudovoters because of boundary issues that arise from Cmult and polytopize; don't go into details

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

Polytopize (1)

two nice pictures from thesis formula written down

Polytopize (2)

Guide, with amortization

picture from paper, but remade to look more sane!!!! Maybe a list of things to note (changes form model)

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

El 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 $Y \rightarrow$ optimize W)

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

El results (2)

Discussion/future work (Ch. 3)

Including the covariate Multiple elections Actual NC data Compare hierarchical 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 , ν , σ_ν 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!!)

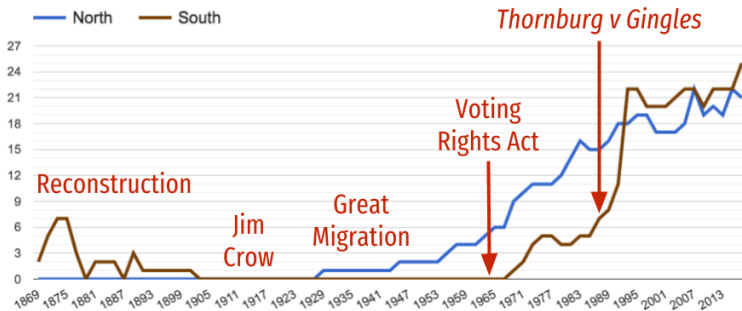
How our EI amortization works (2)

Details on how we get ν and σ_{ν}

More EI results

END DEFENSE, START OLD PRESENTATION

Number of African-Americans in Congress

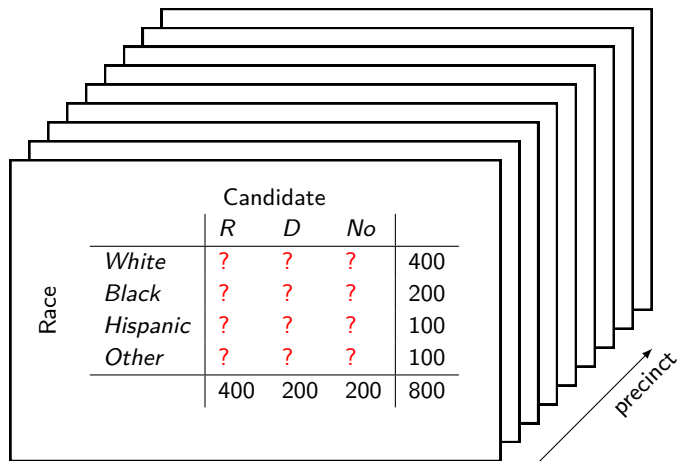


Thornburg v Gingles, 1986

A majority-minority district must be created if:

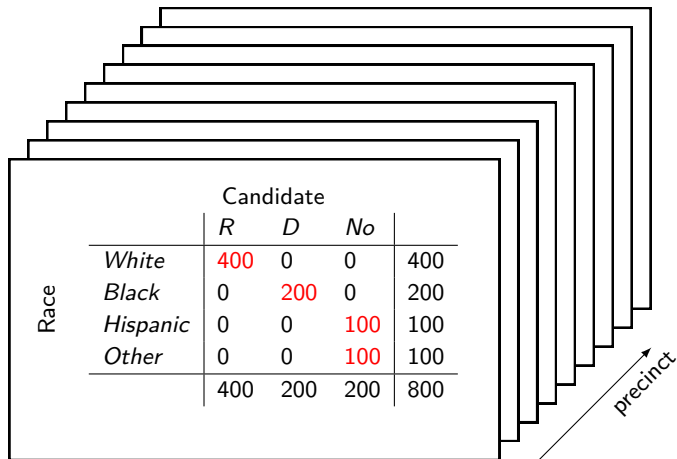
1. A minority group is “sufficiently numerous and compact to form a majority in a single-member district”; and
2. The minority group is **"politically cohesive"**; and
3. The “majority **votes sufficiently as a bloc** to enable it . . . usually to defeat the minority’s preferred candidate.”

Ecological data



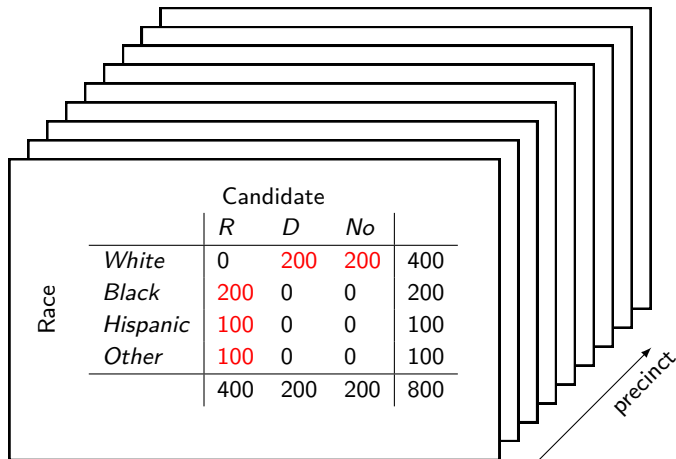
Race	Candidate			
	<i>R</i>	<i>D</i>	<i>No</i>	
	<i>White</i>	?	?	400
	<i>Black</i>	?	?	200
	<i>Hispanic</i>	?	?	100
<i>Other</i>	?	?	?	100
	400	200	200	800

Majority=Majority?



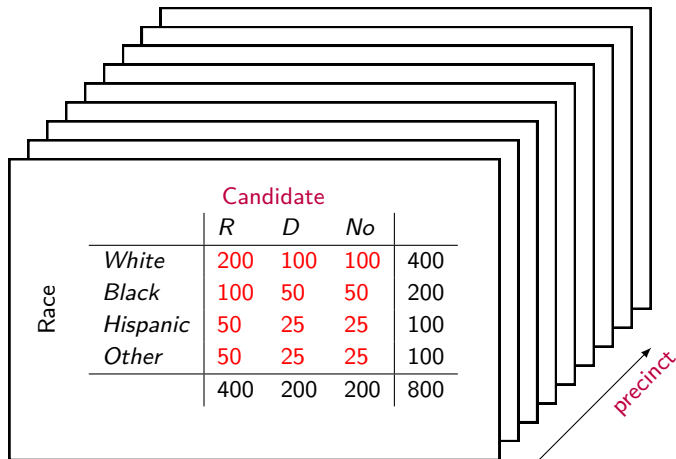
Race	Candidate				
	<i>R</i>	<i>D</i>	<i>No</i>		
	<i>White</i>	400	0	0	400
	<i>Black</i>	0	200	0	200
	<i>Hispanic</i>	0	0	100	100
<i>Other</i>	0	0	100	100	
	400	200	200	800	

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?



The diagram shows a stack of 10 identical tables, representing different precincts. The front table contains the following data:

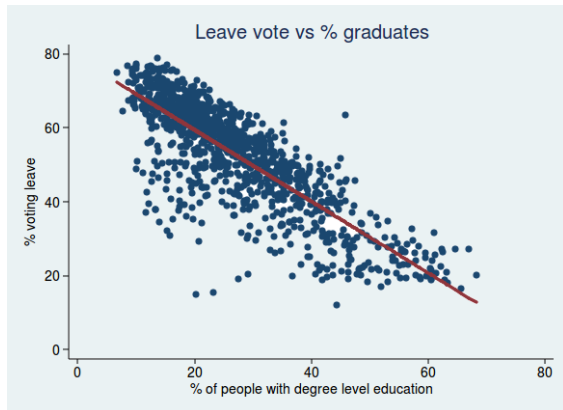
Race	Candidate				
	<i>R</i>	<i>D</i>	<i>No</i>		
	<i>White</i>	200	100	100	400
	<i>Black</i>	100	50	50	200
	<i>Hispanic</i>	50	25	25	100
<i>Other</i>	50	25	25	100	
	400	200	200	800	

An arrow points from the bottom right of the stack to the word "precinct" written in red.

Structure

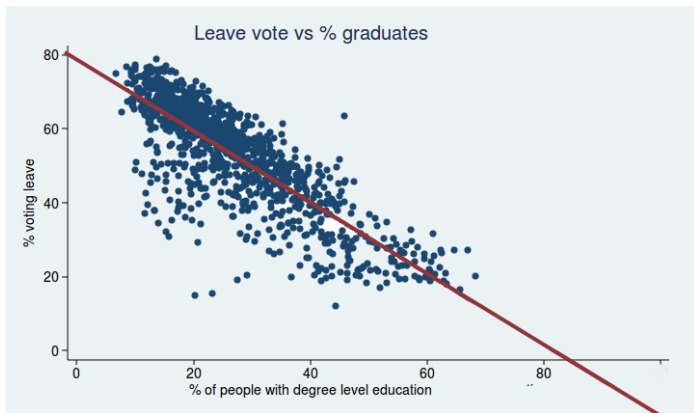
- ▶ Pose the ecological problem (done)
- ▶ Quick review of prior approaches
- ▶ A basic, extensible model for EI
- ▶ Why and how to reparameterize
- ▶ Review of variational inference
- ▶ Applying variational inference to EI
- ▶ Guide (aka variational distribution) based on observed information

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:

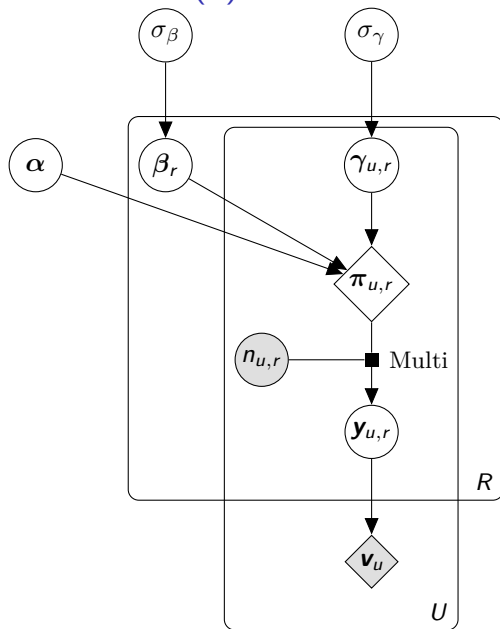
The RKJT model can, in principle, be extended to handle additional factors such as:

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

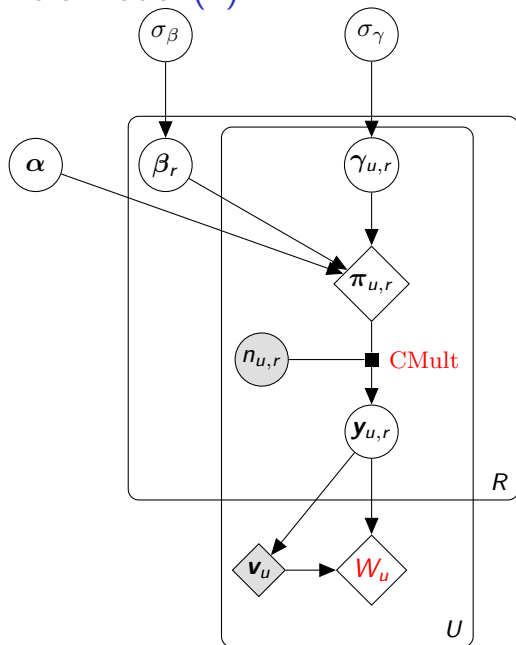
However:

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

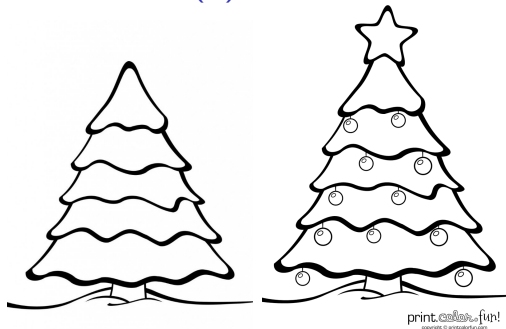
Flexible model (1)



Flexible model (2)



Flexible model (3)

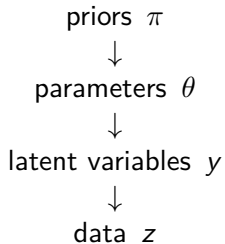


$$\vec{y}_{p,r} = y_{p,r,c} \Big|_{c=1}^C \sim \text{CMult} \left(n_{p,r}, \frac{\exp(\alpha_c + \beta_{r,c} + \lambda_{r,c,p}) \Big|_{c=1}^C}{\sum_{c=1}^C \exp(\alpha_c + \beta_{r,c} + \lambda_{r,c,p})} \right)$$

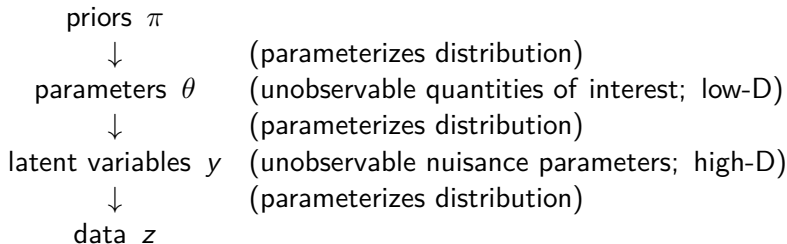
$$\alpha_c \sim \mathcal{N}(0, \sigma_\alpha) \quad \sigma_\alpha \sim \text{Expo}(5)$$

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

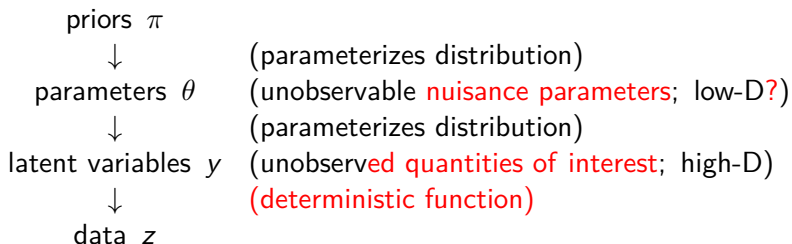
Standard Bayesian approach (simplified)



Standard Bayesian approach (cont'd)



Ecological Inference

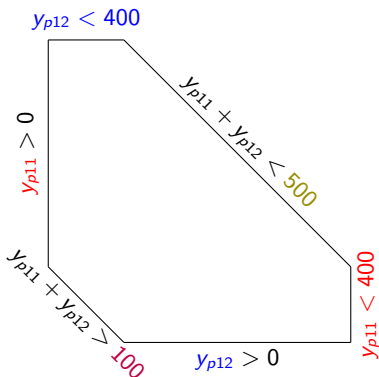


A likelihood from a deterministic function is an indicator function!

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} .

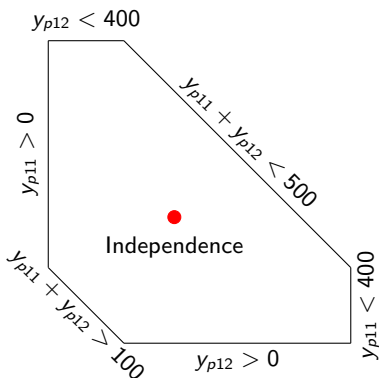
Race	Candidate			
	<i>R</i>	<i>D</i>	<i>No</i>	
	y_{p11}	y_{p12}	y_{p13}	500
	y_{p21}	y_{p22}	y_{p23}	700
400	400	400	400	1200



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

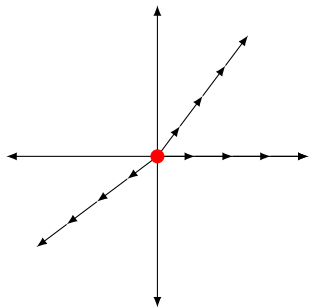
Independence point

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

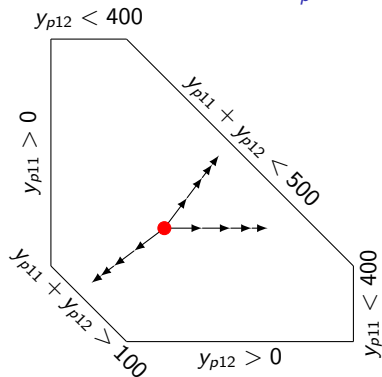


$\mathbb{R}^{(R-1)(C-1)}$

$g(y')$

\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 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 $\mathcal{N}(\mu, \Sigma)$, 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
```



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.

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?

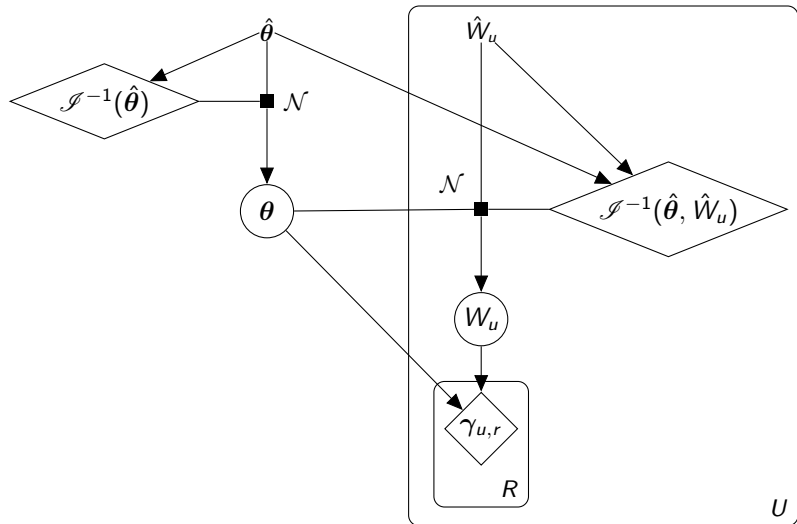
There's an obvious way to approximate a twice-differentiable, unnormalized distribution with a Normal: a Laplace approximation.

That is, use the observed information matrix:

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

as the precision matrix of q .

Graphical posterior



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 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 $\lambda_{p,r,c} \parallel_{c=1}^C$ with its MAP value conditional on all the variables connected to it (not conditionally independent): $y_{p,r,c} \parallel_{c=1}^C$, $\alpha_c \parallel_{c=1}^C$, $\beta_{r,c} \parallel_{c=1}^C$, 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).