

45th Scientific Meeting of the Italian Statistical Society
University of Padua, June 2010

Estimation in Dirichlet Process Random Effects Models

Minjung Kyung	Jeff Gill	George Casella
Department of Statistics University of Florida	Center for Applied Statistics Washington University	Department of Statistics University of Florida

Supported by NSF Grants: DMS-0631632 and SES-0631588.

Introduction

► The Beginning

Prior distributions in the social sciences

► Transition

After the data analysis: model properties

► Dirichlet Process
Random Effects

Likelihood, subclusters, precision parameter

► MCMC

Parameter expansion, convergence, optimality

► Example

Scottish election, normal random effects

► Some Theory

Why are the intervals shorter?

► Sampling Schemes

Logistic, Loglinear, ...

► Conclusions

And other remarks

————But First————
Here is the Big Picture

► Usual Random Effects Model

$$\mathbf{Y}|\psi \sim N(X\beta + \psi, \sigma^2 I), \quad \psi_i \sim N(0, \tau^2)$$

▷ Subject-specific random effect

► Dirichlet Process Random Effects Model

$$\mathbf{Y}|\psi \sim N(X\beta + \psi, \sigma^2 I), \quad \psi_i \sim \mathcal{DP}(\textcolor{red}{m}, N(0, \tau^2))$$

► Results in

▷ Fewer Assumptions

▷ Better Estimates

▷ Shorter Credible Intervals

How This All Started

The Use of Prior Distributions in the Social Sciences

Can more flexible
priors help us
recover latent
hierarchical
information?

- ▶ When do priors matter in social science research?
- ▶ How to specify known prior information?
- ▶ Bayesian social scientists like uninformed priors
- ▶ Reviewers often skeptical about informed priors

▶ **Survey of Political Executives** ([Gill and Casella 2008 JASA](#))

- ▷ Outcome Variable: **stress**
- ▷ surrogate for self-perceived effectiveness and job-satisfaction
- ▷ five-point scale from “not stressful at all” to “very stressful.”
- ▷ **Ordered probit model**

Survey of Political Executives

Some Coefficient Estimates

Posterior	Mean	95% HD Interval
Government Experience	0.120	[-0.086 : 0.141]
Republican	0.076	[-0.031 : 0.087]
Committee Relationship	-0.181	[-0.302 : -0.168]
Confirmation Preparation	-0.316	[-0.598 : -0.286]
Hours/Week	0.447	[0.351 : 0.457]
President Orientation	-0.338	[-0.621 : -0.309]
<i>Cutpoints:</i> (None) (Little)	-1.488	[-1.958 : -1.598]
(Little) (Some)	-0.959	[-1.410 : -1.078]
(Some) (Significant)	-0.325	[-0.786 : 0.454]
(Significant) (Extreme)	0.844	[0.411 : 0.730]

► Intervals are very tight

► Most do not overlap zero

► Seems typical of Dirichlet Process random effects model (later)

► Reasonable Subject Matter Interpretations

Transition

What Did We Learn?

Analyzing Social Science Data

- ▶ Dirichlet Process Random Effects Models
 - ▷ Accepted by Social Scientists
 - ▷ Computationally Feasible
 - ▷ Provides good estimates
- ▶ “Off the shelf ” MCMC ▷ can we do better?
- ▶ Precision parameter m ▷ arbitrarily fixed
- ▶ Answers insensitive to m ???

Understanding the Methodology

- ▶ Next: Better understanding of MCMC and estimation of m .
- ▶ Performance evaluations and wider applications

A Dirichlet Process Random Effects Model

Estimating the Dirichlet Process Parameters

- A general random effects Dirichlet Process model can be written

$$(Y_1, \dots, Y_n) \sim f(y_1, \dots, y_n \mid \theta, \psi_1, \dots, \psi_n) = \prod_i f(y_i \mid \theta, \psi_i)$$

- ▷ $\psi_1, \dots, \psi_n$ iid from $G \sim \mathcal{DP}$
- ▷ $\mathcal{DP}$ is the Dirichlet Process
 - ▷ Base measure ϕ_0 and precision parameter m
 - ▷ The vector θ contains all model parameters

- Blackwell and MacQueen (1973) proved

$$\psi_i \mid \psi_1, \dots, \psi_{i-1} \sim \frac{m}{i-1+m} \phi_0(\psi_i) + \frac{1}{i-1+m} \sum_{l=1}^{i-1} \delta(\psi_l = \psi_i)$$

- ▷ Where δ denotes the Dirac delta function.

Some Distributional Structure

- ▶ Freedman (1963), Ferguson (1973, 1974) and Antoniak (1974)
 - ▷ Dirichlet process prior for nonparametric G
 - ▷ Random probability measure on the space of all measures.

- ▶ Notation
 - ▷ G_0 , a **base distribution** (finite non-null measure)
 - ▷ $m > 0$, a **precision parameter** (finite and non-negative scalar)
 - ▷ Gives spread of distributions around G_0 ,
 - ▷ Prior specification $G \sim \mathcal{DP}(m, G_0) \in \mathcal{P}$.

- ▶ For *any* finite partition of the parameter space, $\{B_1, \dots, B_K\}$,
$$(G(B_1), \dots, G(B_K)) \sim \mathcal{D}(mG_0(B_1), \dots, mG_0(B_K)),$$

A Mixed Dirichlet Process Random Effects Model

Likelihood Function

- The likelihood function is integrated over the random effects

$$L(\theta \mid \mathbf{y}) = \int f(y_1, \dots, y_n \mid \theta, \psi_1, \dots, \psi_n) \pi(\psi_1, \dots, \psi_n) d\psi_1 \cdots d\psi_n$$

- From Lo (1984 Annals) Lemma 2 and Liu (1996 Annals)

$$L(\theta \mid \mathbf{y}) = \frac{\Gamma(m)}{\Gamma(m+n)} \sum_{k=1}^n m^k \left[\sum_{C:|C|=k} \prod_{j=1}^k \Gamma(n_j) \int f(\mathbf{y}_{(j)} \mid \theta, \psi_j) \phi_0(\psi_j) d\psi_j \right],$$

- ▷ The [partition](#) C defines the subclusters
- ▷ $\mathbf{y}_{(j)}$ is the vector of y_i s in subcluster j
- ▷ ψ_j is the common parameter for that subcluster

A Mixed Dirichlet Process Random Effects Model

Matrix Representation of Partitions

- Associate a **binary matrix** $A_{n \times k}$ with a partition C

$$C = \{S_1, S_2, S_3\} = \{\{3, 4, 6\}, \{1, 2\}, \{5\}\} \leftrightarrow A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

- ▷ **Rows**: a_i is a $1 \times k$ vector of all zeros except for a 1 in its subcluster
- ▷ **Columns**: The column sums of A are the number of observations in the groups
- ▷ **Variables**: $\psi_i \in S_j \Rightarrow \psi_i = \eta_j$ (constant in subclusters)

A Mixed Dirichlet Process Random Effects Model

Underlying Random Effects

$$C = \{S_1, S_2, S_3\} = \{\{3, 4, 6\}, \{1, 2\}, \{5\}\} \leftrightarrow A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

► Matrix Representation

$$\boldsymbol{\psi} = A\boldsymbol{\eta} \quad \text{where } A = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \quad \text{so} \quad \begin{pmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_6 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix}.$$

► Only need to generate three random variables

Linear Mixed Model Effect of the A Matrix

► Start with

$$\mathbf{Y}|\psi \sim N(X\beta + \psi, \sigma^2 I), \text{ where } \psi_i \sim \mathcal{DP}(m, N(0, \tau^2)), \quad i = 1, \dots, n$$

► Introduce the A matrices to get

$$\mathbf{Y}|\mathbf{A}, \eta \sim N(X\beta + A\eta, \sigma^2 I), \quad \eta \sim N_k(0, \tau^2 I),$$

► Marginalize over η :

$$\mathbf{Y}|\mathbf{A} \sim N(X\beta, \Sigma^*), \quad \Sigma^* = \left(I + \frac{\tau^2}{\sigma^2} AA' \right)$$

Estimating the Precision Parameter

Maximum Likelihood Estimates

- The likelihood function

$$L(\theta \mid \mathbf{y}) = \frac{\Gamma(m)}{\Gamma(m+n)} \sum_{k=1}^n m^k \mathcal{L}_k(\theta \mid \mathbf{y}),$$

where

$$\mathcal{L}_k(\theta \mid \mathbf{y}) = \sum_{A \in \mathcal{A}_k} \prod_{j=1}^k \Gamma(n_j) \int f(\mathbf{y} \mid \theta, A) \phi_0(\eta_j) d\eta_j.$$

- The MLE $\hat{m}$ solves

$$\frac{\sum_{k=1}^n k m^k \mathcal{L}_k(\theta \mid \mathbf{y})}{\sum_{k=1}^n m^k \mathcal{L}_k(\theta \mid \mathbf{y})} = \sum_{i=1}^n \frac{m}{m+i-1},$$

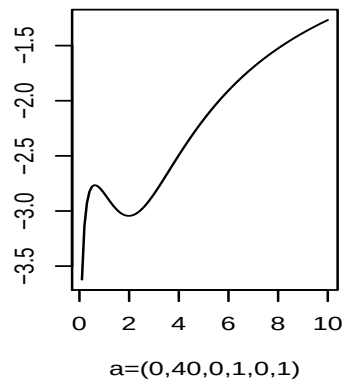
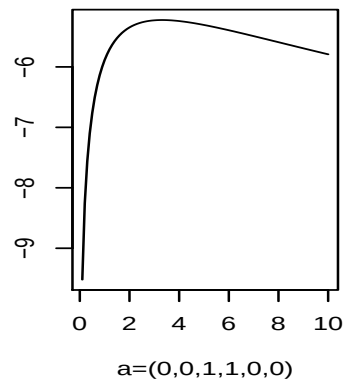
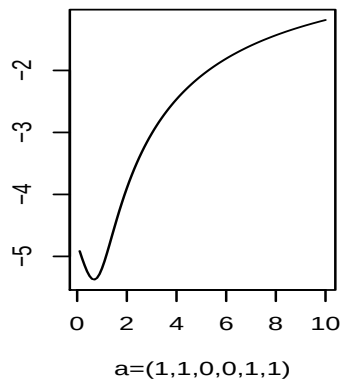
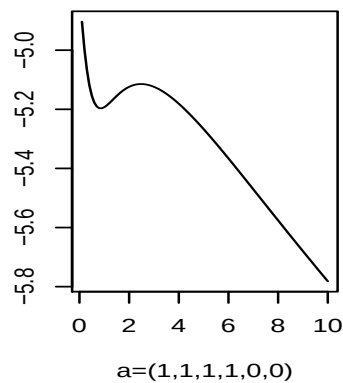
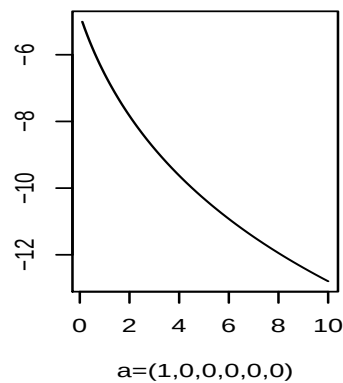
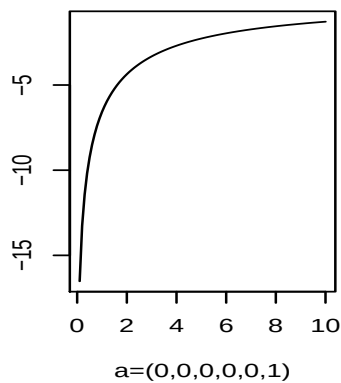
- ▷ Equates posterior and prior mean number of clusters (Liu 1996)

- Many surprises here!

Estimating the Precision Parameter

Shapes of the Likelihood Function

- Log likelihoods for a selection of configurations of component likelihoods



- Not Pleasant

Estimating the Precision Parameter Posterior Mode Estimates

- ▶ Given the potential problems with using and MLE for m
- ▶ Prior input can be helpful here

Gamma priors previously used

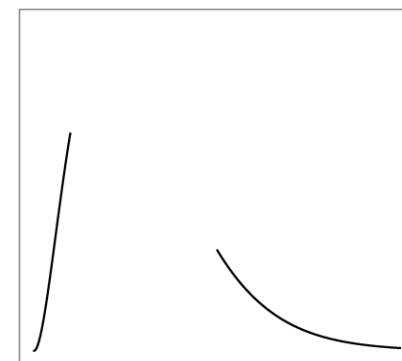
- ▷ Escobar and West (1995)
- ▷ Teh *et al.* (2006)

▶ **Theorem:**

Gamma prior results in

$$\begin{aligned} \pi(m \mid \mathbf{y}) &\uparrow \text{ from } 0 \\ \pi(m \mid \mathbf{y}) &\downarrow \text{ to } \infty \end{aligned}$$

- ▶ Can force interior modal estimate
- ▶ Can reflect expected number of clusters



MCMC Sampling Scheme

Posterior Distribution

- The joint posterior distribution

$$\pi(\theta, A \mid \mathbf{y}) = \frac{m^k f(\mathbf{y} \mid \theta, A) \pi(\theta)}{\int_{\Theta} \sum_A m^k f(\mathbf{y} \mid \theta, A) \pi(\theta) d\theta}.$$

Model

Model parameters θ

→ sampling is straightforward

Random effects

Dirichlet Process parameters

A : the subclusters

m : the precision parameter

MCMC Sampling Scheme

Model Parameters and Dirichlet Process Parameters

► For $t = 1, \dots, T$, at iteration t

Model Parameters

► Starting from $(\theta^{(t)}, A^{(t)})$,

$$\theta^{(t+1)} \sim \pi(\theta \mid A^{(t)}, \mathbf{y}),$$

► Given $\theta^{(t+1)}, A^{(t+1)}$

$$\mathbf{q}^{(t+1)} \sim \text{Dirichlet}(\underbrace{n_1^{(t)} + 1, \dots, n_k^{(t)} + 1, 1, \dots, 1}_{\text{length } n})$$

Dirichlet Process Parameters

$$A^{(t+1)} \propto m^k f(\mathbf{y} \mid \theta^{(t+1)}, A) \binom{n}{n_1 \dots n_n} \prod_{j=1}^n [q_j^{(t+1)}]^{n_j}$$

► where $n_j \geq 0$, $n_1 + \dots + n_n = n$.

MCMC Sampling Scheme – Generating the Partitions

► For Illustration take $n = 7$. At iteration t

$$A_{7 \times 3}^{(t)} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \Rightarrow q^{(t+1)} | A_{7 \times 3}^{(t)} \sim \text{Dirichlet}(4, 3, 3, 1, 1, 1, 1)$$

$$q^{(t+1)} \Rightarrow A_{7 \times 7}^{(t+1)} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \Rightarrow A_{7 \times 5}^{(t+1)} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

► No Matrices Needed

► $A_{7 \times 7}^{(t+1)}$ represented by $(2, 3, 6, 2, 1, 7, 1)$

MCMC Sampling Scheme

Convergence of Dirichlet Process

- Neal (2000) describes 8 algorithms: All use “stick-breaking” conditionals
- The full conditionals of the two chains

Our chain

$$P(a_j = 1 | A_{-j}) \propto \begin{cases} \left(\frac{n_j}{n-1+m} \right) \left(\frac{q_j}{n_j+1} \right) & j = 1, \dots, k \\ \frac{m}{n-1+m} q_{k+1} & j = k+1, \dots, n \end{cases}$$

Stick-breaking chain

$$P(a_j = 1 | A_{-j}) \propto \begin{cases} \frac{n_j}{n-1+m} & j = 1, \dots, k \\ \frac{m}{n-1+m} & j = k+1 \end{cases}$$

- This is a Parameter Expansion

(Liu/Wu 1999 JASA, vanDyk/Meng 2001 JCGS)

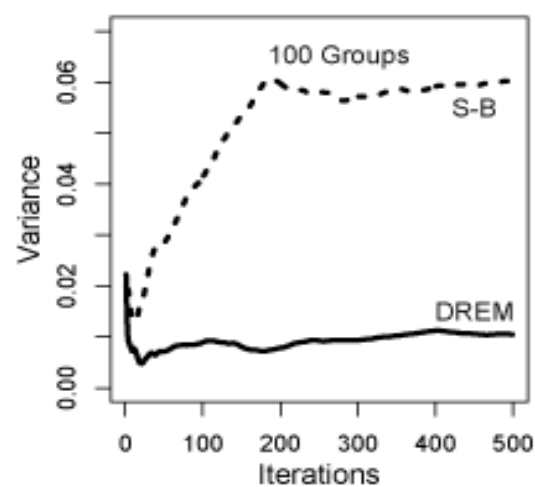
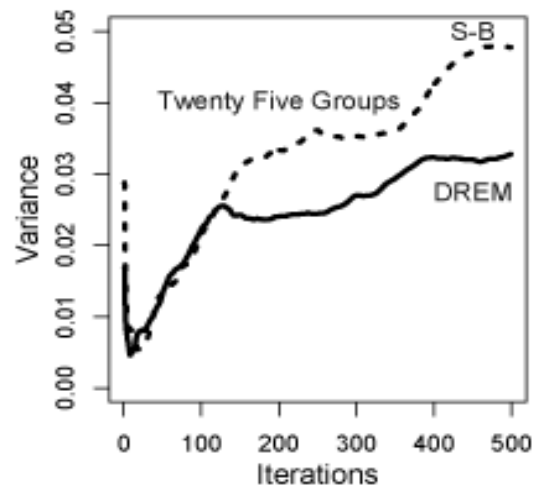
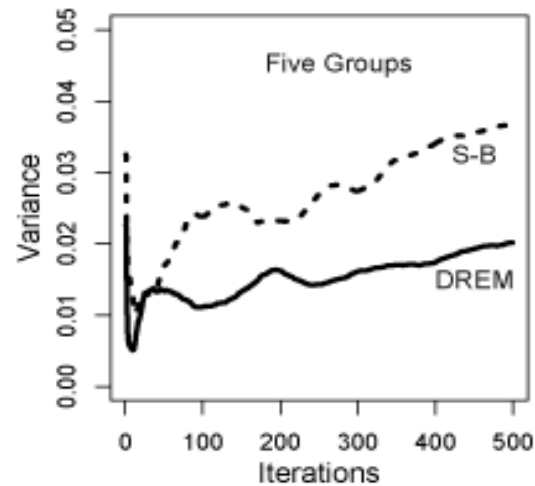
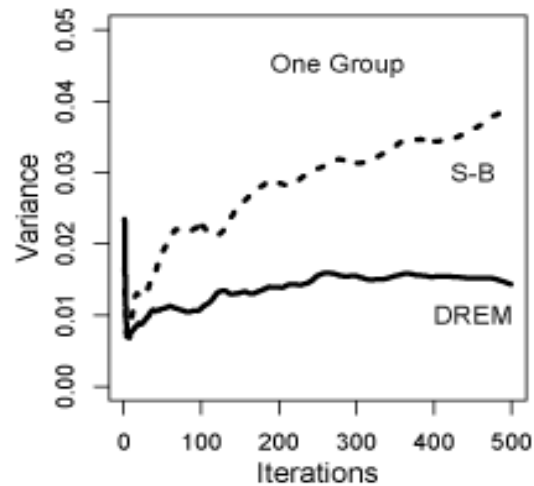
MCMC Sampling Scheme Optimality

- ▶ Hobert and Marchev (2008)
 - ▷ Established superiority of **parameter expanded kernels**

 - ▶ Properties
 1. **Parameter expansion** dominates in **Operator Norm**
 2. **Parameter expansion** dominates in **Efficiency Ordering**

 - ▶ $\text{Var } h(Y)$ is **smaller** for any square-integrable function h .
- (Mira and Geyer, 1999 and Mira, 2001)

MCMC Sampling Scheme – Variance Comparisons



- ▷ $n = 100$
- ▷ 20 Markov chains
- ▷ 1, 5, 25, 100 true groups
- ▷ Variance at each iteration
- ▷ Improvement $\approx 50\%$

Scottish Election Data - History

1997: Scottish voters overwhelmingly (74.3%) approved the creation of the first Scottish parliament

The voters gave strong support, (63.5%), to granting this parliament taxation powers

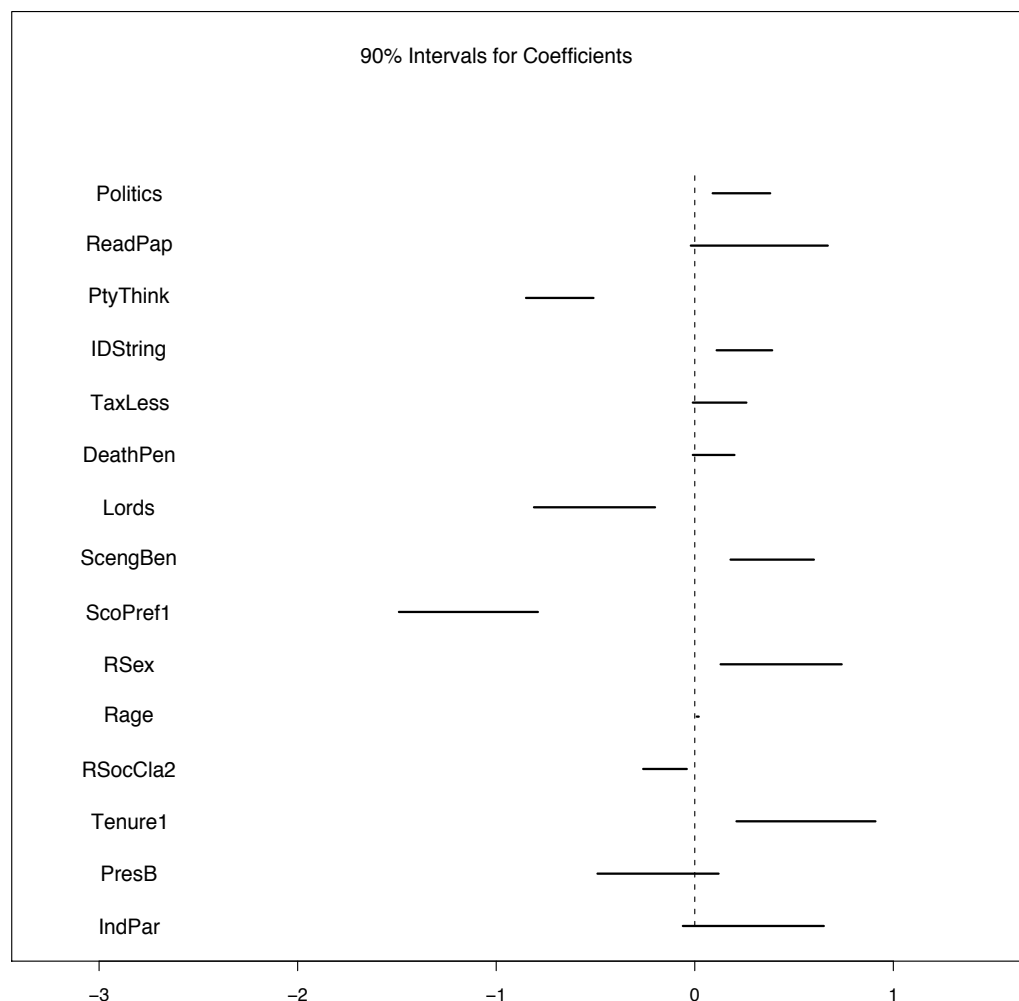
Our Interest:

- ▶ Who subsequently voted conservative in Scotland?

The Data:

- ▶ British General Election Study of 880 Scottish nationals
- ▶ Outcome: party choice (conservative or not) in UK general election
- ▶ Independent variables: political and social measures
- ▶ Probit model

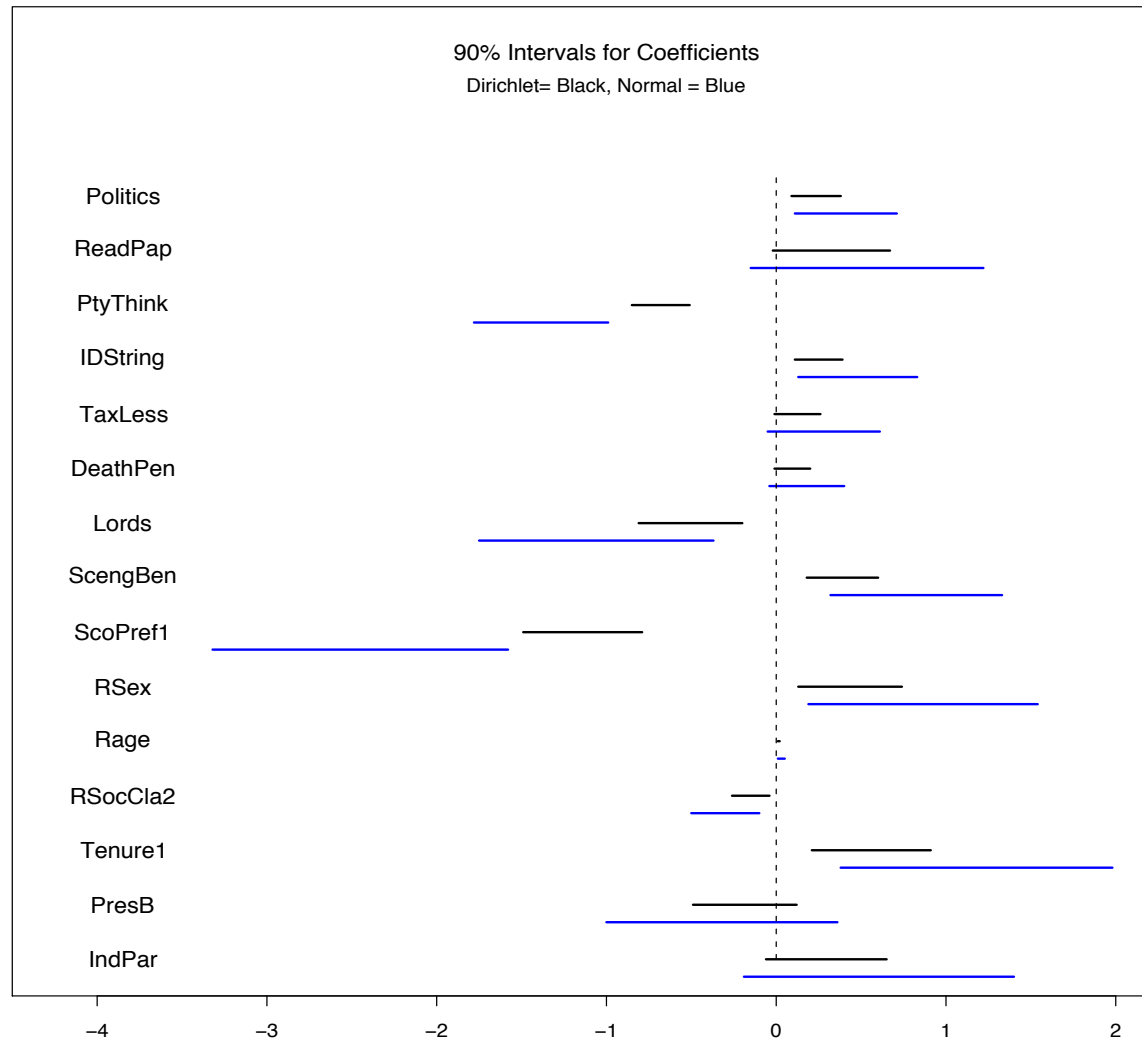
Scottish Election Data - Dirichlet Process Credible Intervals



Probability of Voting
Conservative $\uparrow$ with:

- ▷ Interest in politics (Politics)
- ▷ Read newspapers (ReadPap)
- ▷ Supports fewer taxes (TaxLess)
- ▷ Return death penalty (DeathPen)
- ▶ Some Other Surprising Results

Scottish Election Data - Credible Interval Comparison



Dirichlet Process
vs.
Normal
Random
Effects

Dirichlet Process
Intervals
Uniformly
Shorter

Investigating the Intervals

Why are they shorter?

Kyung, *et al.* (2009)
Stat. and Prob. Letters

- Simpler Model
- Posterior Variance Domination

► Linear Mixed Model

$$Y_{ij} = \mu + \psi_i + \varepsilon_{ij},$$

- Where $\boldsymbol{\psi} = \mathbf{A}\boldsymbol{\eta}$,

$$\mathbf{Y}|\mu, \boldsymbol{\eta}, \sigma^2, \mathbf{A} \sim \mathcal{N}(\mu\mathbf{1} + \mathbf{A}\boldsymbol{\eta}, \sigma^2\mathbf{I}) \quad \boldsymbol{\eta}|\sigma^2 \sim \mathcal{N}_k(\mathbf{0}, c\sigma^2\mathbf{I}_k)$$

$$\mu|\sigma^2 \sim \mathcal{N}(0, v\sigma^2) \quad \sigma^2 \sim \mathcal{IG}(a, b),$$

- ▷ and the hyperparameters are assumed known.

Investigating the Intervals
Why are they shorter?

► Marginal posterior variance distribution $\pi(\sigma^2 | \mathbf{Y}, \mathbf{A})$

▷ Inverted Gamma

▷ Mean proportional to

$$\frac{b + \frac{1}{2} \mathbf{y}' [\mathbf{I} + c \mathbf{A} \mathbf{A}' + v \mathbf{1} \mathbf{1}']^{-1} \mathbf{y}}{\frac{n}{2} + a - 1}$$

► $\mathbf{A} = \mathbf{I}$ is the normal random effects model

► When is

$$\frac{\mathbf{y}' [\mathbf{I} + c \mathbf{I} + v \mathbf{1} \mathbf{1}']^{-1} \mathbf{y}}{\mathbf{y}' [\mathbf{I} + c \mathbf{A} \mathbf{A}' + v \mathbf{1} \mathbf{1}']^{-1} \mathbf{y}} > 1?$$

Investigating the Intervals

Why are they shorter?

► Eigenvalues of $n \times n$ matrix $[\mathbf{I} + c\mathbf{I} + v\mathbf{1}\mathbf{1}']^{-1} [\mathbf{I} + c\mathbf{A}\mathbf{A}' + v\mathbf{1}\mathbf{1}']$

► $\mathbf{A}_{n \times k}$

► There are $n - k$ eigenvalues < 1

Eigenvectors must contain
within cluster contrast

► Look at posterior distribution of the variance

The mean from the
Dirichlet Process model

is
smaller
than

The mean from the
normal model

▷ For all $\mathbf{y}$ not containing a within-subcluster contrast

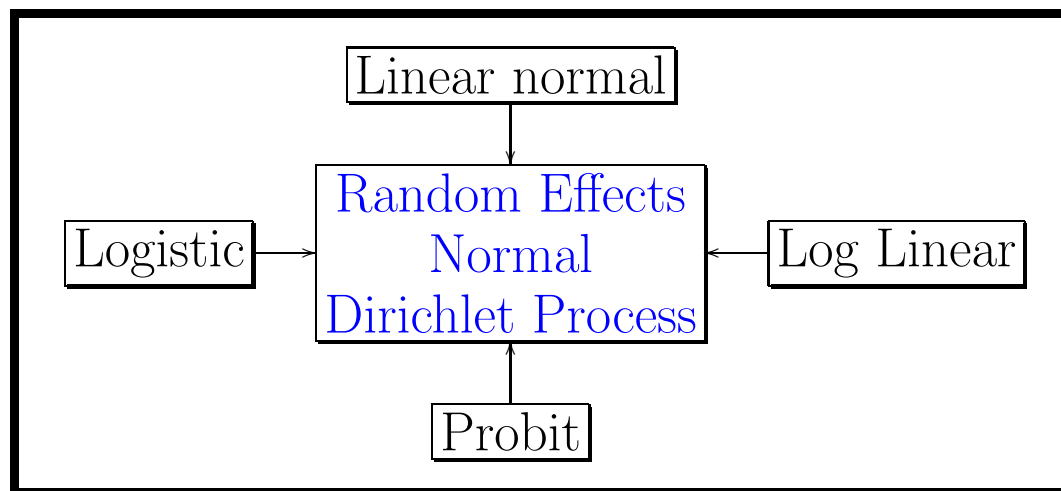
► Implications

▷ The set of $\mathbf{y}$ containing a within-subcluster contrast has measure zero

▷ So the dominance occurs almost surely.

Application to Other Models

Logit and Log Linear



► Can Mix and Match

Model Parameters:

- ▷ Focus on Logistic Here
- ▷ Slice Sampling
- ▷ Mixtures

Dirichlet Process Parameters:

- ▷ Same as before

Application to Other Models

Logistic Regression

- Likelihood function with logit link:

$$L_k(\boldsymbol{\beta}, \tau^2, \boldsymbol{\eta} | A, \mathbf{y}) = \prod_{i=1}^n \left[\frac{\exp(\mathbf{X}_i \boldsymbol{\beta} + (\mathbf{A} \boldsymbol{\eta})_i)}{1 + \exp(\mathbf{X}_i \boldsymbol{\beta} + (\mathbf{A} \boldsymbol{\eta})_i)} \right]^{y_i} \left[\frac{1}{1 + \exp(\mathbf{X}_i \boldsymbol{\beta} + (\mathbf{A} \boldsymbol{\eta})_i)} \right]^{1-y_i} \\ \times \prod_{j=1}^k \left(\frac{1}{2\pi\tau^2} \right)^{1/2} \exp \left(-\frac{1}{2\tau^2} \eta_j^2 \right),$$

- Slice Sampler (Damien *et al.* 1999 JRSSB)

$$L_k(\boldsymbol{\beta}, \tau^2, \boldsymbol{\eta}, \mathbf{U}, \mathbf{V} | \mathbf{A}, \mathbf{y}) \\ = \prod_{i=1}^n I \left[u_i < \left\{ \frac{1}{1 + \exp(-\mathbf{X}_i \boldsymbol{\beta} - (\mathbf{A} \boldsymbol{\eta})_i)} \right\}^{y_i}, v_i < \left\{ \frac{1}{1 + \exp(\mathbf{X}_i \boldsymbol{\beta} + (\mathbf{A} \boldsymbol{\eta})_i)} \right\}^{1-y_i} \right] \\ \times \prod_{j=1}^k \left(\frac{1}{2\pi\tau^2} \right)^{1/2} \exp \left(-\frac{1}{2\tau^2} \eta_j^2 \right)$$

- $\mathbf{U}, \mathbf{V}$ latent; Implement Gibbs sampler

Application to Other Models

Logistic Regression

► Logistic is a Mixture of Normals

▷ Kolmogorov-Smirnov density:

$$f_{KS}(x) = 8 \sum_{\alpha=1}^{\infty} (-1)^{\alpha+1} \alpha^2 x e^{-2\alpha^2 x^2} \quad x \geq 0$$

▷ Mixture of normals is logistic (Andrews and Mallows 1974)

$$\int_0^{\infty} \frac{1}{2x\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left(\frac{y}{2x} \right)^2 \right\} f_{KS}(x) dx = \frac{e^{-y}}{(1 + e^{-y})^2}$$

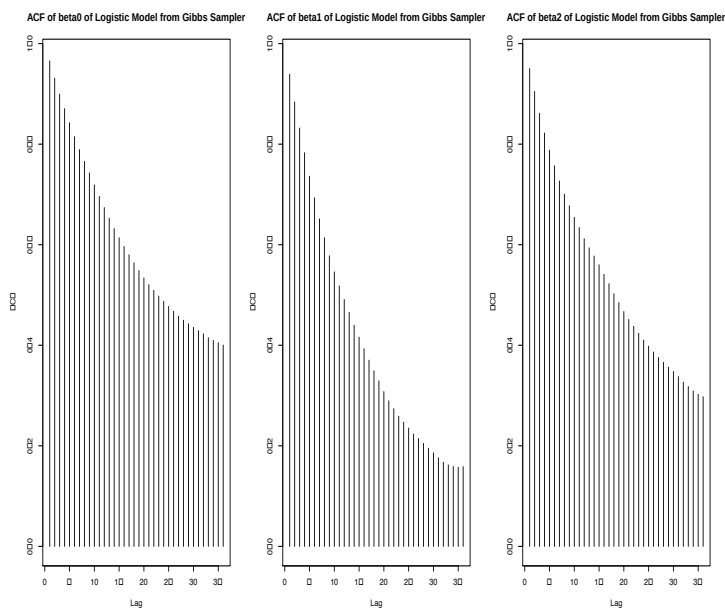
► Easy to simulate (Devroye's (1986) Accept-Reject Algorithm)

Application to Other Models Comparisons

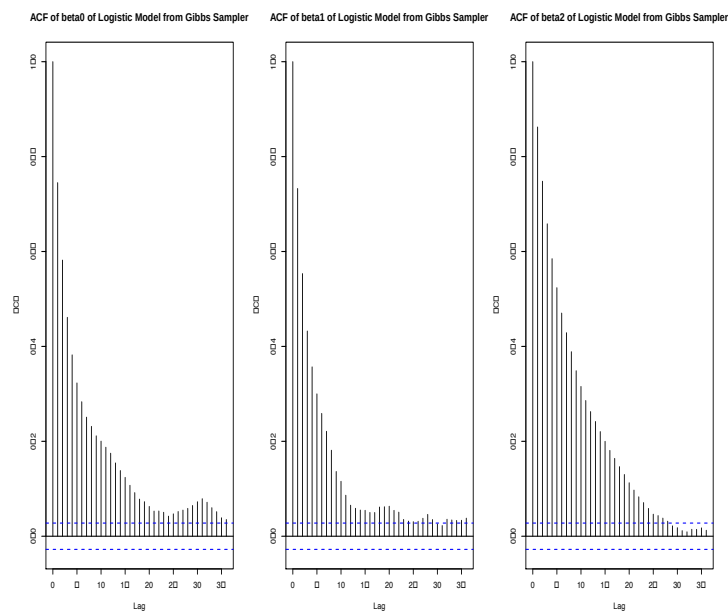
► Estimation of the coefficients and K

▷ True: $K = 40$ and $\boldsymbol{\beta} = (1, 2, 3)$. ▷ Standard errors in parentheses.

Estimation Method	β_0	β_1	β_2	K
Slice	1.4870(0.3139)	1.1512(0.3919)	2.6212(0.4632)	43.4722(4.0477)
K-S Mixture	1.3410(0.1841)	2.1565(0.1999)	2.7205(0.2245)	43.5494(3.9708)



Slice Sampler



K-S Mixture Representation

Conclusions

Modelling the Random Effects

Why is the
Dirichlet Process
a better model
for random effects?

- ▶ “Noninformative”
- ▶ Richer model for random effects
 - ▷ Normality is unverifiable
 - ▷ Dirichlet captures extra variation
- ▶ Shorter Credible Intervals
 - ▷ More precise inference for fixed effects

Conclusions

Estimation and MCMC

Improvements to the
estimation procedure
and the MCMC

Beyond the
Linear Model

- ▶ Matrix representation
 - ▷ Allows simplification
- ▶ Better precision parameter estimation
- ▶ Improved Gibbs sampler
 - ▷ Exploits properties of multinomial
 - ▷ Better mixing
 - ▷ Better Monte Carlo variances
- ▶ Logistic, Loglinear
 - ▷ Can use Dirichlet error model
 - ▷ Retains estimation properties

Thank You for Your Attention

kyung@stat.ufl.edu

jgill@wustl.edu

casella@stat.ufl.edu

