OBayes 09, University of Pennsylvania, June 2009

Estimation in Dirichlet Random Effects Models

Minjung Kyung
Department of Statistics
University of Florida

Jeff Gill
Center for Applied Statistics
Washington University

George Casella
Department of Statistics
University of Florida

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

Introduction

▶ How this all Started	Prior distributions in the social sciences, survey of political executives
▶ Transition	After the data analysis, look at model properties
▶ Dirichlet Random Effects Model	Likelihood, subclusters, linear and probit models
▶ Precision Parameter	Problems with the MLE and how to fix them
▶ MCMC Sampling Scheme	Parameter expansion, convergence, optimality, generating subclusters, precision parameter
▶ Example and Conclusions	Scottish election data, comparison with normal random effects, future work

———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

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

- Results in
  - Fewer Assumptions
  - Better Estimates

How This All Started
The Use of Prior Distributions in the Social Sciences

SubsequentQuestions

- When do priors matter in social science research?
- How to specify known prior information?
- Why do Bayesian social scientists tlike uninformed priors?
- Why are reviewers typically skeptical about informed priors, even with lots of supporting evidence?
- Topically:

Can more flexible priors help us recover latent hierarchical information?

Ordinal Model for 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

$$Y_i \sim \text{Multinomial}(1, (p_1, p_2, \dots, p_C)), \quad i = 1, \dots n$$

▷ The p_j are ordered by the probit model

$$p_j = P(\theta_{j-1} \leq U_i \leq \theta_j)$$

where $-\infty = \theta_0 < \theta_1 < \cdots < \theta_C = \infty$.

▷ Defining the random quantity:

$$U_i \sim \mathcal{N}(\mathbf{X}_i \boldsymbol{\beta} + \psi_i, \sigma^2)$$

• $\mathbf{X}_i$: covariates • $\boldsymbol{\beta}$: coefficient vector • ψ_i: subject-specific random effect

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, $\{B_1, \ldots, B_K\}$,

$$(G(B_1), \ldots, G(B_K)) \sim \mathcal{D}(mG_0(B_1), \ldots, mG_0(B_K)),$$

Ordinal Model for Survey of Political Executives
Specifying the Full Model

- Ordered probit model

$$
\begin{aligned}
Y_i &\sim \text{Multinomial}(1, (p_1, p_2, \ldots, p_C)), \quad i = 1, \ldots n \\
p_j &= P(\theta_{j-1} \leq U_i \leq \theta_j) \\
U_i &\sim \mathcal{N}(\mathbf{X}_i \boldsymbol{\beta} + \psi_i, \sigma^2) \\
\psi_i &\sim G \\
G &\sim \mathcal{DP}(m, G_{\mu,\tau^2}).
\end{aligned}
$$

Model Parameters: $\theta, \mathbf{p}, \beta, \sigma^2$

- ▷ Treated in the usual way
- ▷ Few details here

Dirichlet Parameters: ψ, m

- ▷ Focus of the talk

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]
Cutpoints: **(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 random effects model (later)
- Reasonable Subject Matter Interpretations

Transition
What Did We Learn?

Analyzing Social Science Data

- ▶ Dirichlet Random Effects Models
  - ▷ Accepted by Social Scientists
  - ▷ Computationally Feasible
  - ▷ Provide good estimates
- ▶ Model fitting was cavalier

Understanding the Methodology

- ▶ "Off the shelf" MCMC ▷ can we do better?
- ▶ Precision parameter m ▷ arbitrarily fixed
- ▶ Answers insensitive to m???

- ▶ Next: Better understanding of MCMC and estimation of m.
- ▶ This is what we really want to talk about!

A Mixed Dirichlet Random Effects Model
Estimating the Dirichlet Parameters

- A general random effects Dirichlet model can be written

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

  - $\mathcal{DP}$ is the Dirichlet Process with base measure ϕ_0 and precision parameter m.
  - The vector θ contains all model parameters ⇒ Later

- Blackwell and MacQueen (1973) proved that for $\psi_1, \ldots, \psi_n$ iid from $G \sim \mathcal{DP}$,

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

  - where δ denotes the Dirac delta function.

A Mixed Dirichlet Random Effects Model
Likelihood Function

- The likelihood function is integrated over the random effects

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

$$\pi(\psi_1, \ldots, \psi_n) = \prod_{i=1}^{n} \frac{m\phi_0(\psi_i) + \sum_{j=1}^{i-1} I(\psi_j = \psi_i)}{m + i - 1}.$$

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

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

where

- C defines the subclusters
- $\mathbf{y}_{(j)}$ is the vector of y_is in subcluster j
- ψ_j is the common parameter for that subcluster

A Mixed Dirichlet Random Effects Model
Subclusters

A subcluster C is a partition of the sample of size n into k groups

- Grouping done nonparametrically rather than on substantive criteria
- Likely that any real underlying clusters would be broken up into multiple subclusters
- Little penalty for over-separation

- Regular GLMM assumes that the random effect ψ_i's $\sim N(0, \sigma^2_\psi)$, iid.
- Subclustering assigns different ψ_i across groups, common ψ_i within groups.

A Mixed Dirichlet Random Effects Model
Matrix Representation of Subclusters

- Associate a binary matrix $A_{n\times k}$ with of subcluster 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 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

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

- Only need to generate three random variables

A Mixed Dirichlet Random Effects Model
Components of the Likelihood Function

- For each k, $\mathcal{S}_{n,k}$ different subclusters C, the *Stirling Number of the Second Kind*

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

- $\sum_{k=1}^{n} \mathcal{S}_{n,k} =$ *Bell Number* $= \mathcal{B}_n$
- For example
  - $\mathcal{S}_{25,3} = 141,197,991,025$
  - $\mathcal{B}_{25} = 4,638,590,332,229,999,353$
- Way Cool!

Linear Mixed Model
Effect of the A Matrix

- Start with

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

- 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:

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

Proof: Fun with Matrix Algebra

- Woodbury's Formula: Inverses
- Sylvester's Theorem: Determinants

Probit Mixed Model
Straightforward Extension

- $Y_i \sim$ Bernoulli$(p_i), \quad i = 1, \ldots, n$
- $g(p_i) = X_i\beta + \psi_i$
- Define the Latent Variable

$$U_i \sim N(X_i\beta + \psi_i, \sigma^2),$$

$Y_i = I(U_i > 0)$
results in probit model

- Adds another step to the Gibbs sampler
- Generate U_i conditional on the other parameters
- Truncated normal random variable generation

- Example Later

Estimating the Precision Parameter
Maximum Likelihood Estimates

- The likelihood function;

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

where

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

- The MLE $\hat{m}$ solves

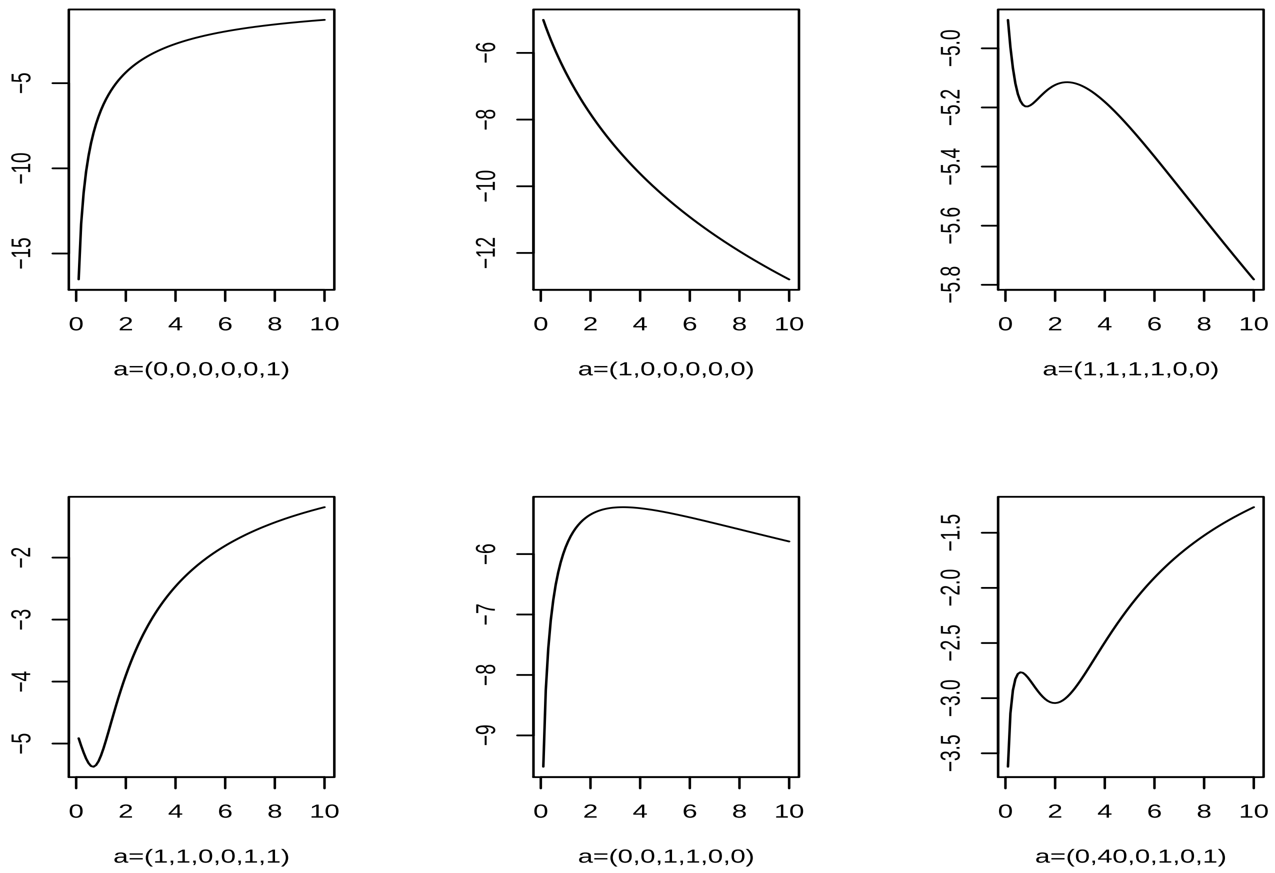
$$\frac{\sum_{k=1}^{n} k m^k \mathcal{L}_k(\theta|\mathbf{y})}{\sum_{k=1}^{n} m^k \mathcal{L}_k(\theta|\mathbf{y})} = \sum_{i=1}^{n} \frac{m}{m+i-1},$$

▷ 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

  $\pi(m \mid \mathbf{y}) \uparrow$ from 0

  $\pi(m \mid \mathbf{y}) \downarrow$ to ∞

- Can force interior modal estimate
- Prior can reflect expected number of clusters
- $\widehat{m}$ from $\frac{\partial}{\partial m} \log \pi(m \mid \mathbf{y}) = 0$

MCMC Sampling Scheme

Posterior Distribution

- The joint posterior distribution

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

Model	Random effects
Model parameters θ $\longrightarrow$ sampling is straightforward	Dirichlet parameters A : the subclusters m : the precision parameter

MCMC Sampling Scheme
Model Parameters and Dirichlet Parameters

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

Model Parameters

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

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

Dirichlet Parameters

- Given $\theta^{(t+1)}$,

$$\mathbf{q}^{(t+1)} \sim \text{Dirichlet}(n_1^{(t)} + \beta_1, \dots, n_k^{(t)} + \beta_k, \beta_{k+1}, \dots, \beta_n)$$

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

- where $n_j \geq 0$, $n_1 + \cdots + n_n = n$.

MCMC Sampling Scheme – Generating the Subclusters

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

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

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

MCMC Sampling Scheme – Marginalization

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

Multinomial Marginalizes Correctly

- No Matrices Needed
- $A^{(t+1)}_{7\times 7}$ represented by $(2, 3, 6, 2, 1, 7, 1)$

MCMC Sampling Scheme
The Transition Kernel

- The joint posterior distribution

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

- The transition kernel of this Markov chain is

  ▷ $P(A|q,\theta) \propto m^k f(\boldsymbol{y}|\theta, A)\binom{n}{n_1 \cdots n_n}\prod_{j=1}^{n} q_j^{n_j}$

  ▷ $q|A \sim \text{Dirichlet}(n_1+\beta_1, \ldots, n_k+\beta_k, \beta_{k+1}, \ldots, \beta_n)$

- $\beta_j = 1$ for all $j \Rightarrow \pi(\theta, A|\boldsymbol{y})$ stationary distribution

MCMC Sampling Scheme
Convergence Properties

- Convergence of Dirichlet :
  - Ignoring the model parameters
  - Concern with convergence of the chain to the stationary distribution

$$\pi(A) = \pi(n_1, \ldots, n_k) = \frac{\Gamma(n)}{\Gamma(n+m)} m^k \prod_{j=1}^{k} \Gamma(n_j),$$

- Neal (2000) describes 8 algorithms
  - Some developed by others
  - All use stick-breaking conditionals

MCMC Sampling Scheme
Parameter Expansion

- The full conditionals of the two chains

Our chain

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

Stick-breaking chain

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

- This is a Parameter Expansion

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

MCMC Sampling Scheme
Optimality

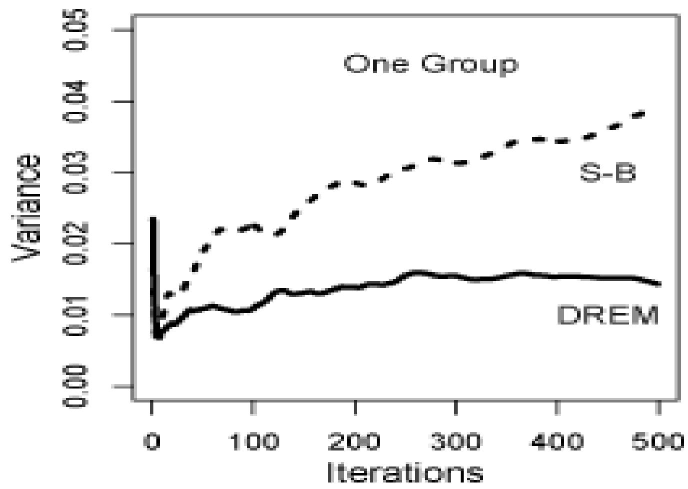
- Hobert and Marchev (2008)
  - ▷ Established superiority of parameter expanded kernels

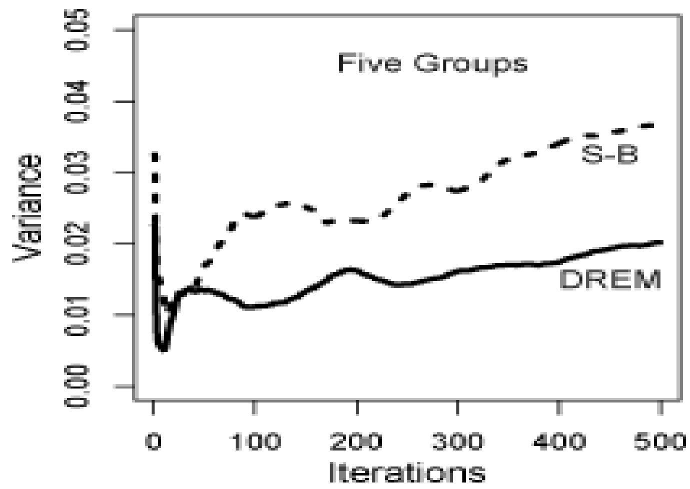
- Properties
  1. Parameter expansion dominates in Operator Norm
  2. Parameter expansion dominates in Efficiency Ordering

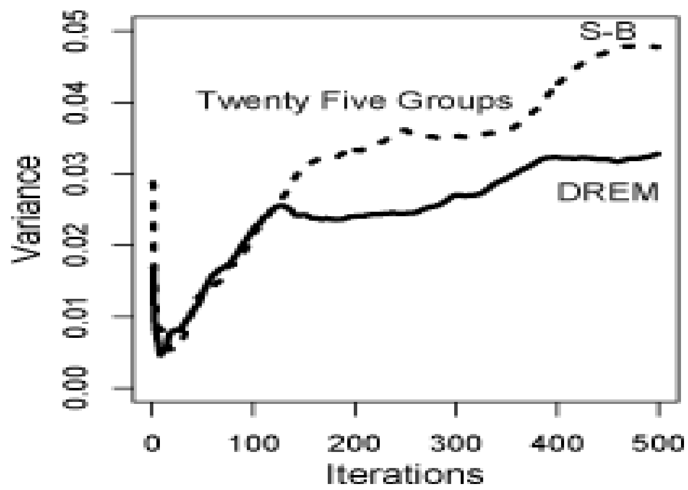
  - Var$h(Y)$ is smaller for any square-integrable function h.

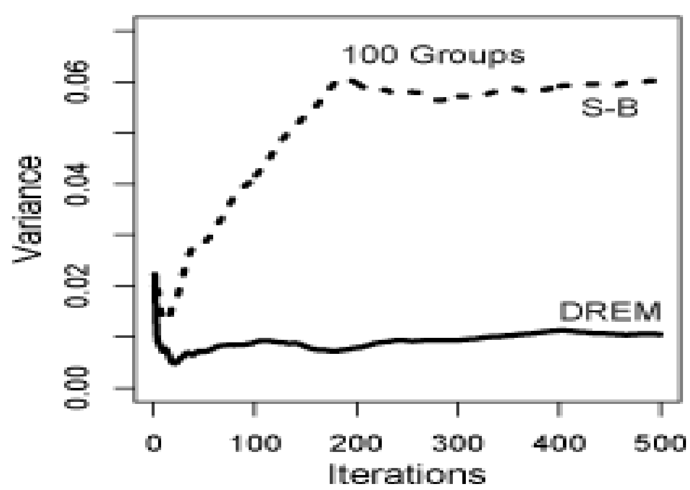
(Mira and Geyer, 1999 and Mira, 2001)

MCMC Sampling Scheme – Variance Comparisons

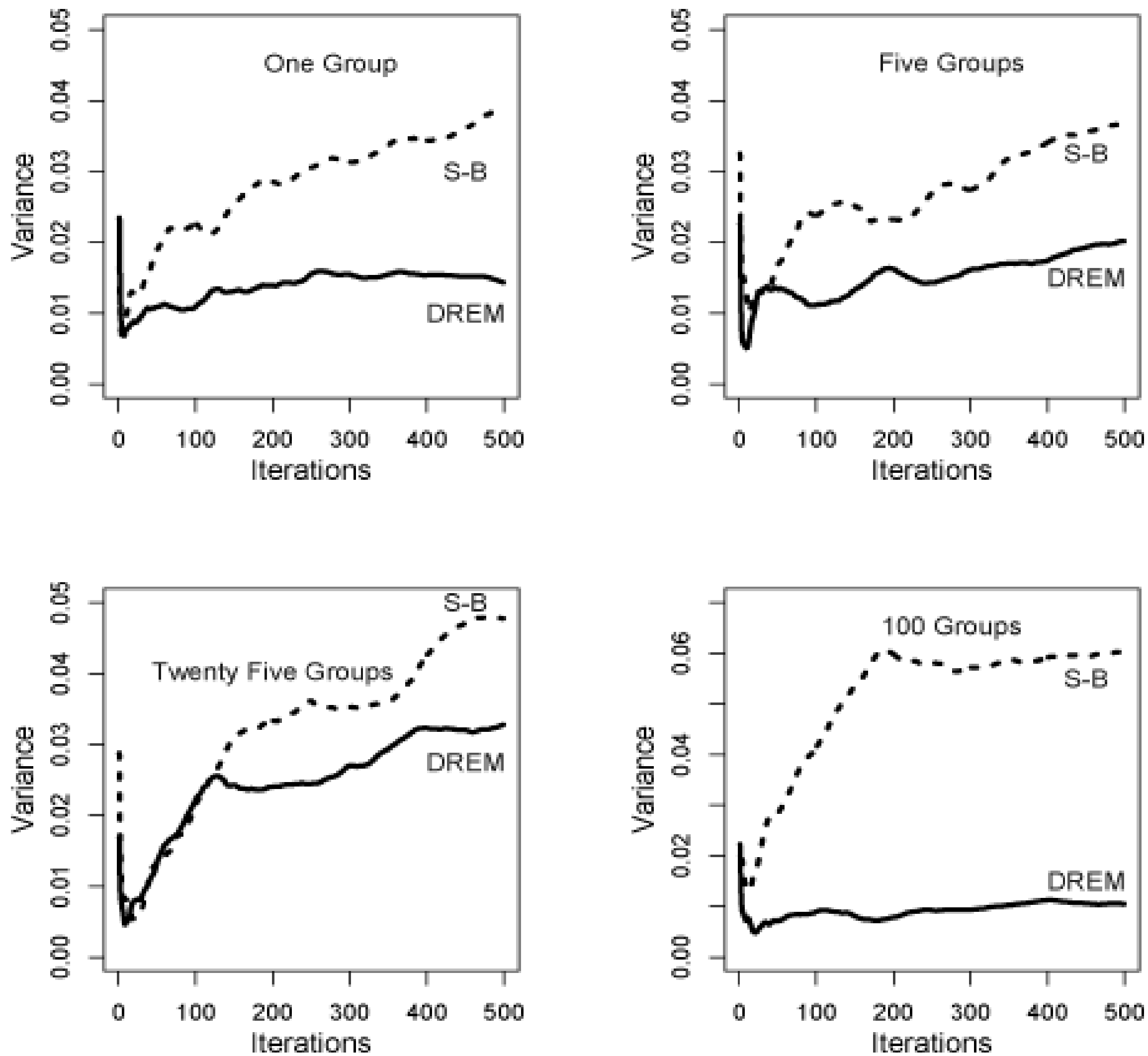








MCMC Sampling Scheme – Variance Comparisons-2



- ▷ $n = 100$
- ▷ 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

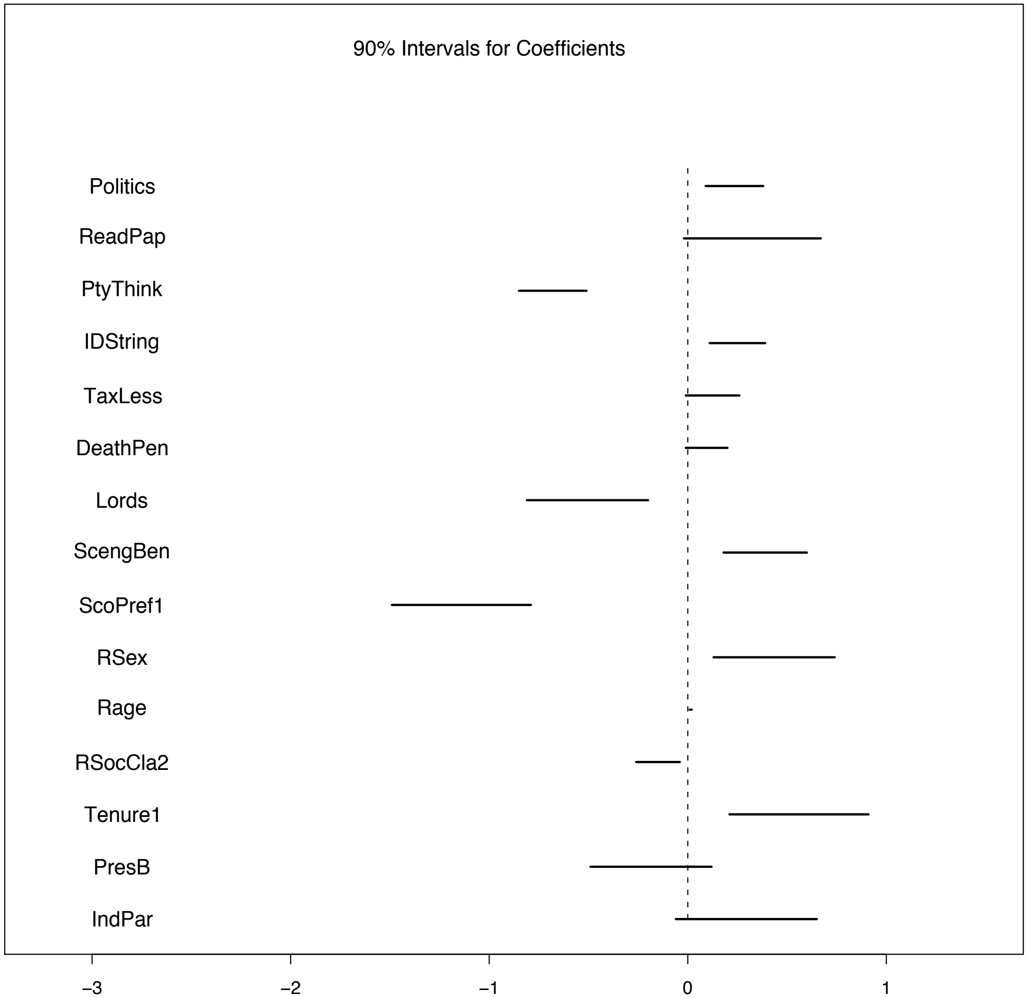
Our Interest:

- 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

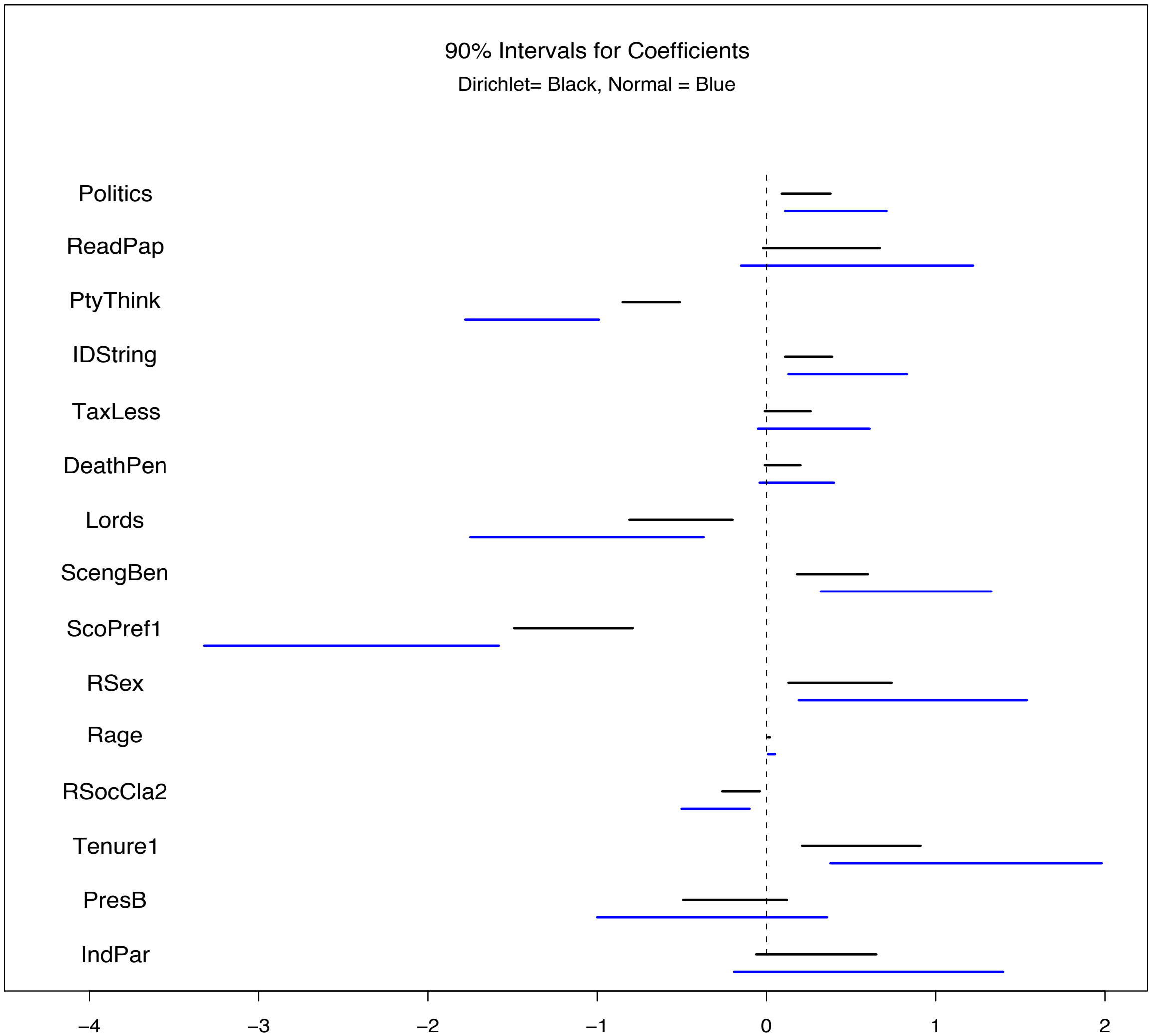
Scottish Election Data - Dirichlet Confidence Intervals



Probability of Voting Conservative ↑ with:

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

Scottish Election Data - Confidence Interval Comparison



Dirichlet vs. Normal Random Effects

Dirichlet Intervals Uniformly Shorter

Conclusions

Modelling the Random Effects

Why is the Dirichlet a better model for random effects?

- Balances the data/prior information
- Burr and Doss (2005): No "residual check" for random effects
  - Normality is unverifiable
- Subclustering captures extra variation
  - Shorter Confidence Intervals

Conclusions
Estimation and MCMC

Improvements to the estimation procedure and the MCMC

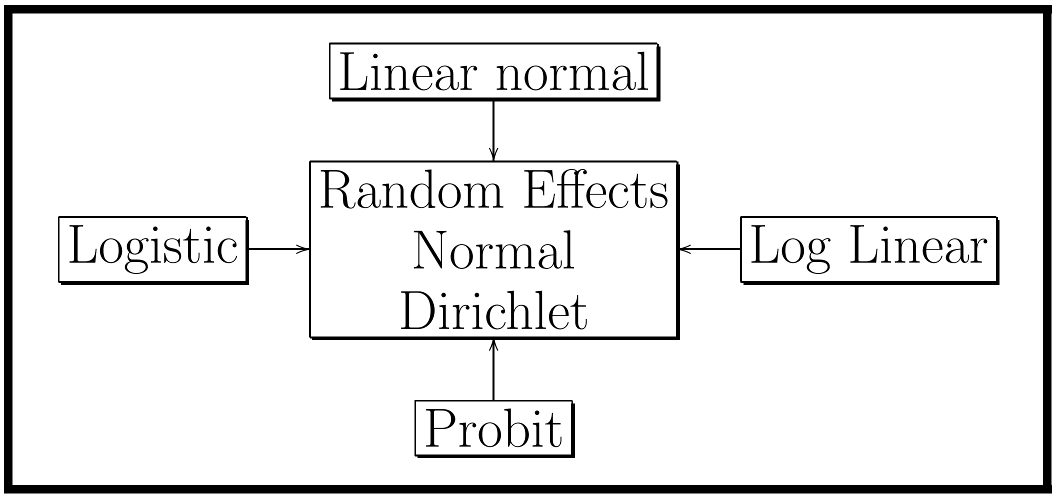
- Matrix representation
  - Allows simplification
- Better precision parameter estimation
- Improved Gibbs sampler
  - Exploits properties of multinomial
  - Better mixing

Conclusions

What Comes Next

Other Models,
Other Methods

- Extend to other models
  - Logit, Loglinear
- More on estimation of m
  - Add to Gibbs Sampler



- Can Mix and Match

Thank You for Your Attention

`casella@ufl.edu`

University of Florida Gators