The Fourth Erich L. Lehmann Symposium May 9 - 12, 2011

Estimation in Mixed Models with Dirichlet Process Random Effects

Both Sides of the Story

George Casella
Department of Statistics
University of Florida

Chen Li
Department of Statistics
University of Florida

Minjung Kyung
Center for Applied Statistics
Washington University

Jeff Gill
Center for Applied Statistics
Washington University

Supported by NSF Grants: SES-0958982 & SES-0959054.

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?
▶ Classical Mixed Models	OLS, BLUE
▶ 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

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

▶ Results in

▷ Fewer Assumptions

▷ Better Estimates

▷ Shorter Credible Intervals

▷ Straightforward Classical Estimation

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]
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 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

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.
- ▶ 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

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

  - $\psi_1, \dots, \psi_n$ 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

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

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

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

A Mixed Dirichlet Process Random Effects Model

Likelihood Function

- The likelihood function is integrated over the random effects

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

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

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

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

A Mixed Dirichlet Process Random Effects Model
Matrix Representation of Partitions

- Start with the model

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

- With Likelihood Function

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

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

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

A Mixed Dirichlet Process Random Effects Model
Matrix Representation of Partitions

- $\boldsymbol{\psi} = A\boldsymbol{\eta}$, $\eta \sim N_k(0, \sigma^2 I)$

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

  - 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)
  - Monte Carlo: Only need to generate k normal random variables

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 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 $(\theta^{(t)}, A^{(t)})$,

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

Dirichlet Process Parameters

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

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

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

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

MCMC Sampling Scheme
Convergence of Dirichlet Process

- Neal (2000) describes 8 algorithms: All use "stick-breaking" conditionals

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

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

- Ours is a Parameter Expansion
- Parameter expansion dominates
- Var $h(Y)$ is smaller for any square-integrable function h.

(Liu/Wu 1999; vanDyk/Meng 2001; Hobert/Marchev 2008; Mira/ Geyer 1999; Mira, 2001)

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

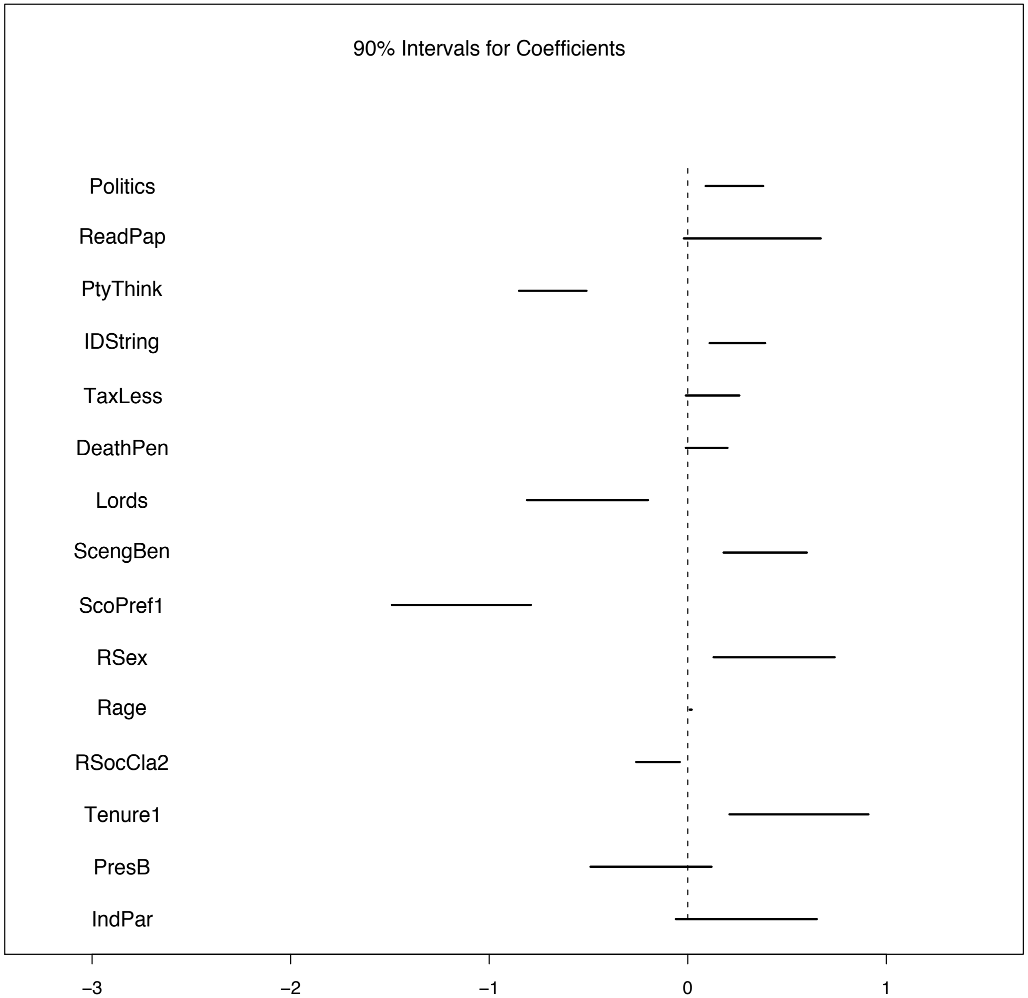
Our Interest:

- 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

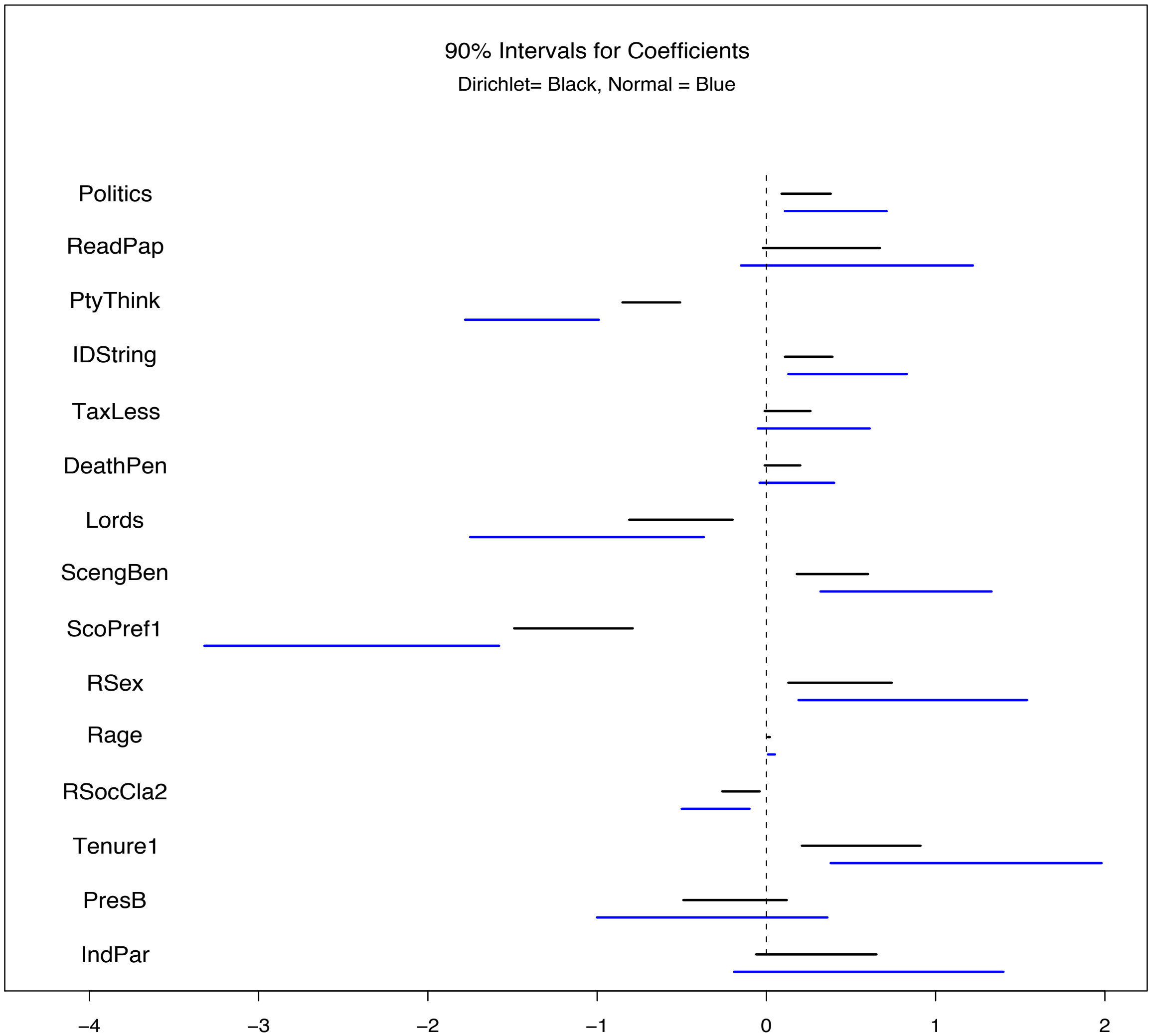
Scottish Election Data - Dirichlet Process Credible Intervals



Probability of Voting Conservative ↑ 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 $\boldsymbol{\psi} = \mathbf{A}\boldsymbol{\eta}$,

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

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

▷ and the hyperparameters are assumed known.

Investigating the Intervals
Why are they shorter?

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

- We can show that

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.

And Now for Something Completely Different
Gauss-Markov Theorem

- Start with the Classic Linear Mixed Model

$$Y = X\boldsymbol{\beta} + Z\boldsymbol{\psi} + \boldsymbol{\varepsilon}$$

  - ▷ $\boldsymbol{\psi} \sim \mathcal{DP}(m, N(0, \tau^2))$
  - ▷ $\varepsilon \sim N(0, \sigma^2 I)$

- Conditional on $\mathbf{A}$, $\boldsymbol{\psi} = \mathbf{A}\boldsymbol{\eta}$, $\boldsymbol{\eta} \sim N(0, \tau^2 I)$, and

$$Y = X\boldsymbol{\beta} + Z\mathbf{A}\boldsymbol{\eta} + \boldsymbol{\varepsilon}$$

- With Mean $\mathrm{E}Y = \mathrm{E}[\mathrm{E}(Y|\mathbf{A})] = X\boldsymbol{\beta}$
- And Variance

$$\mathbf{V} = \mathrm{Var}(Y) = \mathrm{E}[\mathrm{Var}(Y|\mathbf{A})] + \mathrm{Var}[\mathrm{E}(Y|\mathbf{A})] = \mathrm{E}[\mathrm{Var}(Y|\mathbf{A})]$$

Gauss-Markov Theorem
First Application

- Straightforward Application of theorem
  - Zyskind and Martin (1969); Harville (1976)
- BLUE
$$\widetilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$$
- BLUP
$$\widetilde{\boldsymbol{\psi}} = \mathbf{C}\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}),$$
  - $\mathbf{C} = \text{Cov}(Y, \boldsymbol{\psi})$
  - $\mathbf{V} = \text{Var}(Y)$
- Neat Theory
  - What is $\mathbf{C}$?
  - What is $\mathbf{V}$?

Using the Gauss-Markov Theorem
Calculating the Variance

- $\mathbf{V} = \text{Var}(Y) = \text{E}[\text{Var}(Y|\mathbf{A})]$, where

$$\mathbf{V} = \sigma^2 I_n + \text{E}[\tau^2 Z\mathbf{A}\mathbf{A}'Z'] = \sigma^2 I_n + \tau^2 \sum_{\mathbf{A}} P(\mathbf{A}) Z\mathbf{A}\mathbf{A}'Z'.$$

  - with

$$P(\mathbf{A}) = \pi(r_1, r_2, ..., r_k) = \frac{\Gamma(m)}{\Gamma(m+r)} m^k \prod_{j=1}^{k} \Gamma(r_j).$$

  - $r_1, r_2, ..., r_k$ are the column sums

- The sum is over all possible $\mathbf{A}$ matrices
  - Lots of terms in the sum
  - But we can do it (almost - in a special case)

Calculating the Variance
A Special Case

- We can handle the model

$$Y_{ij} = \mathbf{x}_i'\boldsymbol{\beta} + \psi_i + \varepsilon_{ij}, \;\; 1 \leq i \leq r, \; 1 \leq j \leq t,$$

  - which is the previous model with $Z = B$ where

$$B = \begin{bmatrix} \mathbf{1}_t & 0 & \cdots & 0 \\ 0 & \mathbf{1}_t & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & \mathbf{1}_t \end{bmatrix}_{n\times r},$$

- Resulting in

$$d = Cor(Y_{i,j}, Y_{i',j'}) = \tau^2 \sum_{\mathbf{A}} P(\mathbf{A}) a_i' a_j$$

Covariance Matrix
A Special Case

- For the model

$$Y = X\boldsymbol{\beta} + B\boldsymbol{\psi} + \boldsymbol{\varepsilon}$$

- The covariance matrix is

$$\mathbf{V} = \begin{bmatrix} \sigma^2\mathbf{I} + \tau^2\mathbf{J} & d\mathbf{J} & d\mathbf{J} & \cdots & d\mathbf{J} \\ d\mathbf{J} & \sigma^2\mathbf{I} + \tau^2\mathbf{J} & d\mathbf{J} & \cdots & d\mathbf{J} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ d\mathbf{J} & d\mathbf{J} & \cdots & d\mathbf{J} & \sigma^2\mathbf{I} + \tau^2\mathbf{J} \end{bmatrix},$$

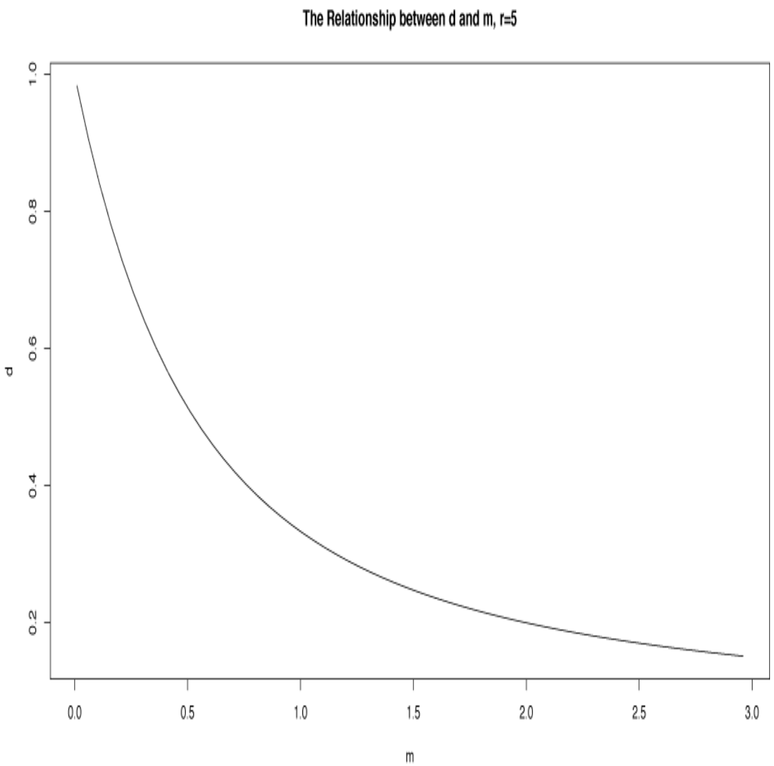
where $\mathbf{I}$ is the $t \times t$ identity matrix, $\mathbf{J}$ is a $t \times t$ matrix of ones,

- And

$$d = Cor(Y_{i,j}, Y_{i',j'}) = \tau^2 \sum_{i=1}^{r-1} im \frac{\Gamma(m+r-1-i)\Gamma(i)}{\Gamma(m+r)}.$$

Examining the Covariance
Dirichlet Precision Parameter

Corr.



m

- Precision parameter m related to correlation in the observations
- Relationship not previously known
- $m \downarrow$ yields more clusters
  - Decreased correlation
- $m \uparrow$ yields fewer clusters
  - Increased correlation

Alternatively
OLS - Least Squares

- For the model

$$Y = X\boldsymbol{\beta} + B\boldsymbol{\psi} + \boldsymbol{\varepsilon}$$

- The OLS Estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'Y$$

- When is OLS=BLUE?
  - ▷ This is “Fun with Matrix Algebra”
  - ▷ Relationship between X, B, and $\mathbf{V}$
  - ▷ Zyskind (1967); Puntanen and Styan (1989)

$$\mathbf{HV} = \mathbf{VH} \text{ where } \mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'.$$

  - ▷ Alternative eigenvector/eigenvalue conditions

OLS=BLUE
Some Conditions

- For the model

$$Y = X\boldsymbol{\beta} + B\boldsymbol{\psi} + \boldsymbol{\varepsilon}$$

- OLS=BLUE for
  - Balanced anova models
  - Some slight extensions

- In particular, for the oneway random effects model

$$\mathbf{Y} = \mathbf{1}\mu + \mathbf{B}\boldsymbol{\psi} + \boldsymbol{\varepsilon},$$

we have

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} = \overline{\mathbf{Y}}.$$

Distribution of the BLUE $\overline{Y}$
Oneway Model

- Here we look at

$$\mathbf{Y} = \mathbf{1}\mu + \mathbf{B}\boldsymbol{\psi} + \boldsymbol{\varepsilon},$$

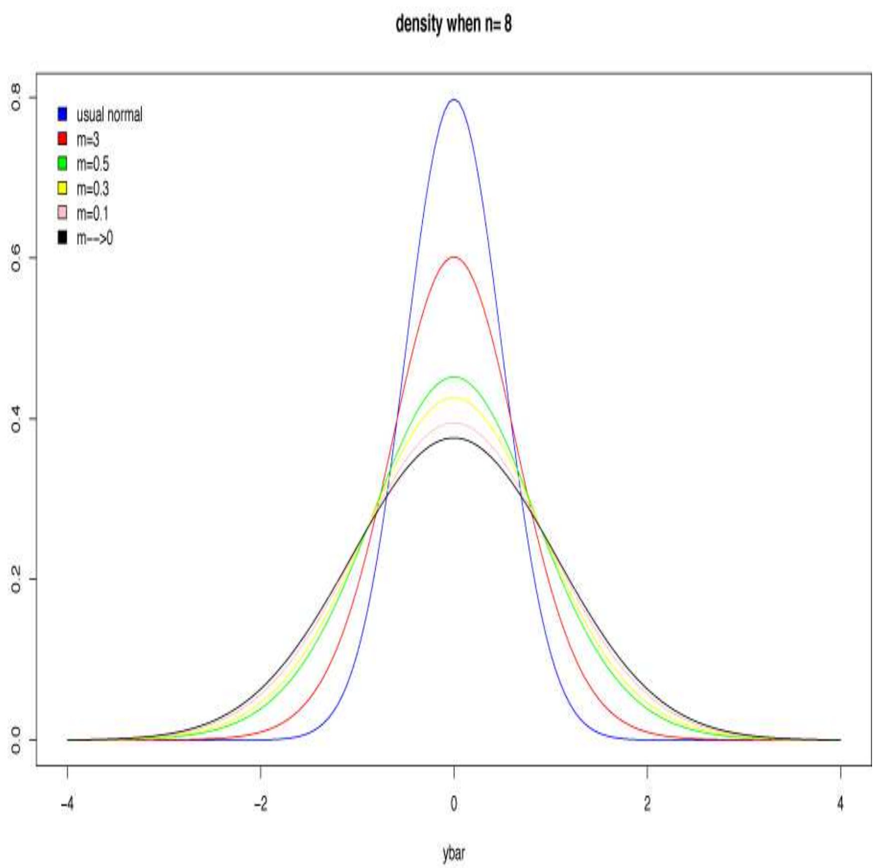
  - Some results generalize (in paper)

- The BLUE $\overline{Y}$ has density

$$f_m(\bar{y}) = \sum_A f(\bar{y}|\mathbf{A})P(\mathbf{A})$$

  - $f(\bar{y}|\mathbf{A}) = N(\mathbf{1}\mu, \sigma^2 I + \frac{\tau^2}{\sigma^2}\mathbf{BAA'B'})$
  - $P(\mathbf{A}) = \pi(r_1, r_2, ..., r_k) = \frac{\Gamma(m)}{\Gamma(m+r)} m^k \prod_{j=1}^k \Gamma(r_j).$
  - m is the precision parameter

Properties of $f_m(y)$
Oneway Model



- Unimodal
- $m \to 0$, $\overline{Y} \sim N(\mu, \frac{1}{n}\sigma^2 + \tau^2))$
  - One Cluster
- $m \to \infty$, $\overline{Y} \sim N(\mu, \frac{1}{n}(\sigma^2 + \tau^2 t))$
  - n Clusters
  - Classical oneway model
- $\underbrace{F_0(\bar{y})}_{\text{Fattest Tails}} < F_m(\bar{y}) < \underbrace{F_\infty(\bar{y})}_{\text{Thinnest Tails}}$

Distribution of the BLUE $\overline{Y}$
Example Cutoff Points

- 95% Confidence Bounds
- $Y_{ij} = \mu + \psi_i + \varepsilon_{ij},\;\; 1 \le i \le 6,\; 1 \le j \le 6,\; ,\; \sigma^2 = \tau^2 = 1$

m							
0	.1	.5	1	2	5	20	∞
1.987	1.917	1.706	1.566	1.355	1.145	0.952	0.864

- Conservative Confidence Bounds
- Can also estimate σ^2 and τ^2

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

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

Beyond the Linear Model

- Logistic, Loglinear
  - Can use Dirichlet error model
  - Retains estimation properties

Conclusions

Classical Approach

Point Estimation

- Covariance Matrix
  - ▷ Calculable
  - ▷ Interpretation of precision parameter
- Estimates
  - ▷ OLS and BLUE reasonable

Confidence Intervals

- Next
  - ▷ Variance Comparisons?
  - ▷ Coverage of Bayes Intervals?

Thank You for Your Attention

`casella@ufl.edu`

Findings So Far

- Gill and Casella(2009). "Nonparametric Priors For Ordinal Bayesian Social Science Models: Specification and Estimation." *JASA*, 104, 453-464
  DPP on RE can uncover latent clustering.

- Kyung *et al.* (2009) "Characterizing the Variance Improvement in Linear Dirichlet Random Effects Models." *Stat. Prob. Letters*, 79, 2343-2350
  DPP on RE can produce lower SE for regression parameters on average.

- Kyung, Gill and Casella(2010) "Estimation in Dirichlet Random Effects Models." *Annals of Statistics*, 38, 979-1009
  Estimation of the precision parameter; improved Gibbs sampler.

- Kyung *et al.* (2011) "Sampling Schemes for Generalized Linear Dirichlet Process Random Effects Models." *Stat. Methods & Applications, to appear*.
  Slice sampling worse than KS mixture representation or MH algorithm.

- Kyung *et al.* (2011) "New Findings from Terrorism Data: Dirichlet Process Random Effects Models for Latent Groups." *JRSSC, to appear*.
  Logistic model, uncovering latent information with difficult data.

- Li, Chen (2011). "Classical Estimation in Linear Mixed Models with Dirichlet Process Random Effects". PhD Thesis, University of Florida
  OLS, BLUE, and comparisons with Bayes estimates