Estimation in Mixed Models  
with  
Dirichlet Process Random Effects  
Both Sides of the Story  

George Casella  
Department of Statistics  
University of Florida  

Minjung Kyung  
Center for Applied Statistics  
Washington University  

Chen Li  
Department of Statistics  
University of Florida  

Jeff Gill  
Center for Applied Statistics  
Washington University  

Supported by NSF Grants: SES-0958982 & SES-0959054.
<table>
<thead>
<tr>
<th>Section</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Beginning</td>
<td>Prior distributions in the social sciences</td>
</tr>
<tr>
<td>Transition</td>
<td>After the data analysis: model properties</td>
</tr>
<tr>
<td>Dirichlet Process Random Effects</td>
<td>Likelihood, subclusters, precision parameter</td>
</tr>
<tr>
<td>MCMC</td>
<td>Parameter expansion, convergence, optimality</td>
</tr>
<tr>
<td>Example</td>
<td>Scottish election, normal random effects</td>
</tr>
<tr>
<td>Some Theory</td>
<td>Why are the intervals shorter?</td>
</tr>
<tr>
<td>Classical Mixed Models</td>
<td>OLS, BLUE</td>
</tr>
<tr>
<td>Conclusions</td>
<td>And other remarks</td>
</tr>
</tbody>
</table>
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
  \[ \mathbf{Y} | \psi \sim N(X\beta + \psi, \sigma^2 I), \quad \psi_i \sim \text{DP}(m, N(0, \tau^2)) \]

- 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

<table>
<thead>
<tr>
<th>Posterior</th>
<th>Mean</th>
<th>95% HD Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Government Experience</td>
<td>0.120</td>
<td>[ –0.086 : 0.141 ]</td>
</tr>
<tr>
<td>Republican</td>
<td>0.076</td>
<td>[ –0.031 : 0.087 ]</td>
</tr>
<tr>
<td>Committee Relationship</td>
<td>-0.181</td>
<td>[ –0.302 : -0.168 ]</td>
</tr>
<tr>
<td>Confirmation Preparation</td>
<td>-0.316</td>
<td>[ –0.598 : -0.286 ]</td>
</tr>
<tr>
<td>Hours/Week</td>
<td>0.447</td>
<td>[ 0.351 : 0.457 ]</td>
</tr>
<tr>
<td>President Orientation</td>
<td>-0.338</td>
<td>[ –0.621 : -0.309 ]</td>
</tr>
</tbody>
</table>

*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?

▸ 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 $m$???

▸ 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

$$(Y_1, \ldots, Y_n) \sim f(y_1, \ldots, y_n \mid \theta, \psi_1, \ldots, \psi_n) = \prod_{i} f(y_i \mid \theta, \psi_i)$$

$\psi_1, \ldots, \psi_n$ iid from $G \sim \mathcal{DP}$

$\mathcal{DP}$ is the Dirichlet Process

$\triangleright$ Base measure $\phi_0$ and precision parameter $m$

$\triangleright$ The vector $\theta$ contains all model parameters

Blackwell and MacQueen (1973) proved

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

Where $\delta$ denotes the Dirac delta function.
Some Distributional Structure

  - 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, $\{B_1, \ldots, B_K\}$,
  $$ (G(B_1), \ldots, G(B_K)) \sim \mathcal{D}(mG_0(B_1), \ldots, mG_0(B_K)), $$
A Mixed Dirichlet Process Random Effects Model

Likelihood Function

The likelihood function is integrated over the random effects
\[
L(\theta \mid \mathbf{y}) = \int f(y_1, \ldots, y_n \mid \theta, \psi_1, \ldots, \psi_n) \pi(\psi_1, \ldots, \psi_n) \, d\psi_1 \cdots d\psi_n
\]

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

The partition \( C \) defines the subclusters
\( \mathbf{y}_{(j)} \) is the vector of \( y_i \)s in subcluster \( j \)
\( \psi_j \) is the common parameter for that subcluster
A Mixed Dirichlet Process Random Effects Model
Matrix Representation of Partitions

- Start with the model
  \[ Y|\psi \sim N(X\beta + \psi, \sigma^2 I), \text{ where } \psi_i \sim \mathcal{DP}(m, N(0, \tau^2)), \quad i = 1, \ldots, n \]

- With Likelihood Function
  \[ L(\theta | y) = \frac{\Gamma(m)}{\Gamma(m+n)} \sum_{k=1}^{n} m^k \left[ \sum_{C:|C|=k} \prod_{j=1}^{k} \Gamma(n_j) \int f(y_{(j)} | \theta, \psi_j) \phi_0(\psi_j) \, d\psi_j \right], \]

- 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

▶ \( \psi = A\eta, \eta \sim N_k(0, \sigma^2 I) \)

\[
Y | A, \eta \sim N(X\beta + A\eta, \sigma^2 I), \quad \eta \sim N_k(0, \tau^2 I),
\]

▶ **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
\[ \pi(\theta, A \mid y) = \frac{m^k f(y \mid \theta, A)\pi(\theta)}{\int_{\Theta} \sum_A m^k f(y \mid \theta, A)\pi(\theta) \, d\theta}. \]

Model

Random effects

Model parameters \( \theta \)

→ sampling is straightforward

Dirichlet Process parameters

\( A \) : the subclusters
\( m \) : the precision parameter
MCMC Sampling Scheme
Model Parameters and Dirichlet Process Parameters

For $t = 1, \ldots, T$, at iteration $t$

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

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

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

\[
q^{(t+1)} \sim \text{Dirichlet}(n_1^{(t)} + 1, \ldots, n_k^{(t)} + 1, 1, \ldots, 1)
\]

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

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

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

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
- \(
\text{Var } h(Y) \text{ is smaller for any square-integrable function } h.
\)

(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

90% Intervals for Coefficients

Politics
ReadPap
PtyThink
IDString
TaxLess
DeathPen
Lords
ScengBen
ScoPref1
RSex
Rage
RSocCla2
Tenure1
PresB
IndPar

Probability of Voting
Conservative ↑ with:

▷ Interest in politics (Politics)
▷ Read newspapers (ReadPap)
▷ Supports fewer taxes (TaxLess)
▷ Return death penalty (DeathPen)

▷ Some Other Surprising Results .....
Estimation in Dirichlet Process Random Effects Models: Scottish Election Data [16]

Scottish Election Data - Credible Interval Comparison

90% Intervals for Coefficients
Dirichlet = Black, Normal = Blue

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 \( \psi = A\eta \),

\[ Y_{ij} \mid \mu, \eta, \sigma^2, A \sim \mathcal{N} (\mu \mathbf{1} + A\eta, \sigma^2 \mathbf{I}) \quad \eta \mid \sigma^2 \sim \mathcal{N}_k (0, c\sigma^2 \mathbf{I}_k) \]

\[ \mu \mid \sigma^2 \sim \mathcal{N} (0, v\sigma^2) \quad \sigma^2 \sim \mathcal{IG} (a, b), \]

- and the hyperparameters are assumed known.
Investigating the Intervals
Why are they shorter?

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

► We can show that

| The mean from the Dirichlet Process model is smaller than | The mean from the normal model |

► For all $y$ not containing a within-subcluster contrast

► Implications

► The set of $y$ 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

\[ Y = X\beta + Z\psi + \epsilon \]

\[ \triangleright \psi \sim \mathcal{DP}(m, N(0, \tau^2)) \quad \triangleright \epsilon \sim N(0, \sigma^2 I) \]

► Conditional on \( A \), \( \psi = A\eta \), \( \eta \sim N(0, \tau^2 I) \), and

\[ Y = X\beta + ZA\eta + \epsilon \]

► With Mean \( EY = E[E(Y|A)] = X\beta \)

► And Variance

\[ V = \text{Var}(Y) = E[\text{Var}(Y|A)] + \text{Var}[E(Y|A)] = E[\text{Var}(Y|A)] \]
Gauss-Markov Theorem
First Application

- Straightforward Application of theorem
  - Zyskind and Martin (1969); Harville (1976)

- BLUE
  \[ \tilde{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y \]

- BLUP
  \[ \tilde{\psi} = CV^{-1}(Y - X\tilde{\beta}), \]
  - \( C = \text{Cov}(Y, \psi) \)
  - \( V = \text{Var}(Y) \)

- Neat Theory
  - What is \( C \)?
  - What is \( V \)?
Using the Gauss-Markov Theorem
Calculating the Variance

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

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

▷ with

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

▷ \( r_1, r_2, \ldots, r_k \) 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

\[ Y_{ij} = x_i' \beta + \psi_i + \varepsilon_{ij}, \quad 1 \leq i \leq r, \quad 1 \leq j \leq t, \]

which is the previous model with \( Z = B \) where

\[
B = \begin{bmatrix}
1_t & 0 & \cdots & 0 \\
0 & 1_t & \cdots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & 1_t \\
\end{bmatrix}_{n \times r},
\]

Resulting in

\[
d = Cor(Y_{i,j}, Y_{i',j'}) = \tau^2 \sum_A P(A) a_i'a_j
\]
Covariance Matrix

A Special Case

For the model

\[ Y = X\beta + B\psi + \varepsilon \]

The covariance matrix is

\[
V = \begin{bmatrix}
\sigma^2 I + \tau^2 J & dJ & dJ & \cdots & dJ \\
dJ & \sigma^2 I + \tau^2 J & dJ & \cdots & dJ \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
dJ & dJ & \cdots & dJ & \sigma^2 I + \tau^2 J
\end{bmatrix},
\]

where \( I \) is the \( t \times t \) identity matrix, \( J \) 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.

- 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
  \[ Y = X\beta + B\psi + \epsilon \]

- The OLS Estimator of \( \beta \) is
  \[ \hat{\beta} = (X'X)^{-1}X'Y \]

- When is OLS=BLUE?
  - This is “Fun with Matrix Algebra”
  - Relationship between \( X, B, \) and \( V \)
  - Zyskind (1967); Puntanen and Styan (1989)
    \[ HV = VH \text{ where } H = X(X'X)^{-1}X'. \]
  - Alternative eigenvector/eigenvalue conditions
OLS=BLUE

Some Conditions

For the model

\[ Y = X\beta + B\psi + \varepsilon \]

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

In particular, for the one-way random effects model

\[ Y = 1\mu + B\psi + \varepsilon, \]

we have

\[ \hat{\beta} = (X'X)^{-1}X'Y = (X'V^{-1}X)^{-1}X'V^{-1}Y = \bar{Y}. \]
Distribution of the BLUE $\bar{Y}$

Oneway Model

Here we look at

$$Y = 1\mu + B\psi + \epsilon,$$

Some results generalize (in paper)

The BLUE $\bar{Y}$ has density

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

$$f(\bar{y}|A) = N(1\mu, \sigma^2 I + \frac{r^2}{\sigma^2}BAA'B')$$

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

$m$ is the precision parameter
Properties of $f_m(y)$
Oneway Model

Unimodal

$m \to 0$, $\bar{Y} \sim N(\mu, \frac{1}{n}\sigma^2 + \tau^2)$
  ▶ One Cluster

$m \to \infty$, $\bar{Y} \sim N(\mu, \frac{1}{n}(\sigma^2 + \tau^2t))$
  ▶ $n$ Clusters
  ▶ Classical oneway model

$F_0(y)$ < $F_m(y)$ < $F_\infty(y)$
Fattest Tails  Thinnest Tails
Distribution of the BLUE $\bar{Y}$
Example Cutoff Points

- 95% Confidence Bounds

- $Y_{ij} = \mu + \psi_i + \varepsilon_{ij}, \ 1 \leq i \leq 6, \ 1 \leq j \leq 6, \ , \sigma^2 = \tau^2 = 1$

<table>
<thead>
<tr>
<th>$m$</th>
<th>0</th>
<th>.1</th>
<th>.5</th>
<th>1</th>
<th>2</th>
<th>5</th>
<th>20</th>
<th>$\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m$</td>
<td>1.987</td>
<td>1.917</td>
<td>1.706</td>
<td>1.566</td>
<td>1.355</td>
<td>1.145</td>
<td>0.952</td>
<td>0.864</td>
</tr>
</tbody>
</table>

- Conservative Confidence Bounds

- Can also estimate $\sigma^2$ and $\tau^2$
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

- 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

Improvements to the estimation procedure and the MCMC

Beyond the Linear Model
Conclusions
Classical Approach

Point Estimation

- Covariance Matrix
  - Calculable
  - Interpretation of precision parameter

- Estimates
  - OLS and BLUE reasonable

Next
- Variance Comparisons?
- Coverage of Bayes Intervals?
Thank You for Your Attention

casella@ufl.edu
Findings So Far

DPP on RE can uncover latent clustering.

DPP on RE can produce lower SE for regression parameters on average.

Estimation of the precision parameter; improved Gibbs sampler.

Slice sampling worse than KS mixture representation or MH algorithm.

Logistic model, uncovering latent information with difficult data.

OLS, BLUE, and comparisons with Bayes estimates