

# Characterizing the Variance Improvement in Linear Dirichlet Random Effects Models

Minjung Kyung\*  
University of Florida

Jeff Gill†  
Washington University

George Casella‡  
University of Florida

August 11, 2009

## Abstract

An alternative to the classical mixed model with normal random effects is to use a Dirichlet process to model the random effects. Such models have proven useful in practice, and we have observed a noticeable variance reduction, in the estimation of the fixed effects, when the Dirichlet process is used instead of the normal. In this paper we formalize this notion, and give a theoretical justification for the expected variance reduction. We show that for almost all data vectors, the posterior variance from the Dirichlet random effects model is smaller than that from the normal random effects model.

---

\*Postdoctoral Associate, Department of Statistics, University Florida, Gainesville, FL 32611. Supported by National Science Foundation Grants DMS-0631632 and SES-0631588. Email: [kyung@stat.ufl.edu](mailto:kyung@stat.ufl.edu).

†Professor, Center for Applied Statistics, Washington University, One Brookings Dr., Seigle Hall LL-085, St. Louis, MO. Supported by National Science Foundation Grants DMS-0631632 and SES-0631588. Email: [jgill@wustl.edu](mailto:jgill@wustl.edu).

‡Distinguished Professor, Department of Statistics, University of Florida, Gainesville, FL 32611. Supported by National Science Foundation Grants DMS-0631632 and SES-0631588. Email: [casella@stat.ufl.edu](mailto:casella@stat.ufl.edu).

# 1 Introduction

The popular general linear mixed model has the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\eta} + \varepsilon, \tag{1}$$

where the response  $\mathbf{Y}$  is modeled as a linear function of the fixed effect  $\boldsymbol{\beta}$  and the random effects  $\boldsymbol{\eta}$ , with known design or observation matrices  $\mathbf{X}$  and  $\mathbf{Z}$ . It is typical to model both  $\varepsilon$  and  $\boldsymbol{\eta}$  with independent normal distributions. This setup can be extended to a generalized linear mixed model by specifying a suitable link function for some categorical outcome variable, and obviously provides a more flexible specification. Details of these models with various link functions, covering both statistical inferences and computational methods, can be found in the recent texts by McCulloch and Searle (2001) and Jiang (2007).

Variations of these models were used by Burr and Doss (2005), Dorazio *et al.* (2008) and Gill and Casella (2009), where the distributional assumption on  $\boldsymbol{\eta}$  is changed to a Dirichlet process. It was typically found that the richer Dirichlet model resulted in lower posterior variances on the fixed effects. Indeed, Gill and Casella (2009) and Kyung *et al.* (2009) show some examples with striking improvement in variance estimates when moving from normal random effects to Dirichlet random effects. This evidence is anecdotal, based on observing variance estimates from various published data analyses. In this paper we investigate some of the underlying theory that could explain this phenomenon.

## 1.1 Background

Dirichlet process mixture models were introduced by Ferguson (1973), who defined the process and investigated basic properties. Antoniak (1974) proved the posterior distribution is a mixture of Dirichlet processes, and Blackwell and MacQueen (1973) showed that the marginal distribution of the Dirichlet process is equal to the distribution of the  $n^{\text{th}}$  step of a Polya urn process. In particular, they demonstrated that for the Dirichlet process, if a new observation is obtained, it either has the same value of a previously drawn observations, or it has a new value drawn from a distribution  $G_0$ , the base measure. The frequency of new components from  $G_0$  is controlled by  $m$ , the precision parameter. Other work that characterizes the properties of the Dirichlet process includes Korwar and Hollander (1973), who characterize the joint distribution and look at nonparametric empirical Bayes estimation of the distribution function based on Dirichlet process priors, and Sethuraman (1994), who shows that the Dirichlet measure is a distribution on the space of all probability measures and it gives probability one to the subset of discrete probability measures. In terms of estimation, the results of Lo (1984) and Liu (1996) allow us to write the likelihood function in a form suitable for estimation of parameters.

Much work has been done in developing estimation strategies, particularly those based on Markov chain Monte Carlo (MCMC) algorithms. In our previous work (Kyung *et al.* 2009), where

we proposed a new MCMC algorithm for a linear mixed model with a Dirichlet process random effect term, we noticed that when fitting mixed models to survey data from a recent Scottish election, by every standard measure of fit, the generalized linear mixed model with a Dirichlet process random effect term outperformed a simple Bayesian probit model with diffuse uniform prior distributions on the parameters and normal random effects. This is important, since the latter model is part of the standard Bayesian toolkit, particularly in the social sciences. When the lengths of credible intervals were compared, we found that the Dirichlet model resulted in uniformly shorter intervals than those of a normal random effects model. Thus, Kyung *et al.* argued that the richer random effects model is able to remove more extraneous variability, resulting in tighter credible intervals. However, this is an anecdotal observation, based on the results from the Scottish data analysis and a few others.

## 1.2 Overview

In this paper, we compare the marginal posterior distribution of the variances for the Dirichlet random effects model to those from a normal random effects model, to theoretically verify the anecdotal observations. We are able to show that for almost any typical data vector, the posterior variance from the Dirichlet model is smaller than that from the normal. In Section 2 we describe the Dirichlet random effects model and the case that we consider here. Section 3 compares posterior variances, and develops a matrix theorem that shows how the Dirichlet posterior variance is smaller than that of the normal. Finally, Section 4 has a short discussion.

## 2 Dirichlet Random Effects Models

In this section we give some details about the likelihood function in a general Dirichlet random effects model, and show how those results help us to obtain a simpler representation of the linear Dirichlet random effects model

### 2.1 A General Dirichlet Random Effects Model

A general random effects Dirichlet model can be written

$$\begin{aligned} (Y_1, \dots, Y_n) &\sim f(y_1, \dots, y_n \mid \boldsymbol{\theta}, \psi_1, \dots, \psi_n) = \prod_i f(y_i \mid \boldsymbol{\theta}, \psi_i) \\ \psi_i &\sim \mathcal{DP}(m, \phi_0), \quad i = 1, \dots, n, \end{aligned} \tag{2}$$

where the random variable  $Y_i$  has density  $f(y_i \mid \boldsymbol{\theta}, \psi_i)$ ,  $\mathcal{DP}$  is the Dirichlet Process with base measure  $\phi_0$  and concentration parameter  $m$ . The vector  $\boldsymbol{\theta}$  contains all of the model parameters. Blackwell and MacQueen (1973) proved that for  $\psi_1, \dots, \psi_n$  iid from  $G \sim \mathcal{DP}(m, \phi_0)$ , the joint distribution

of  $\boldsymbol{\psi}$  is a product of successive conditional distributions of the form:

$$\psi_i | \psi_1, \dots, \psi_{i-1}, m \sim \frac{m}{i-1+m} \phi_0(\psi_i) + \frac{1}{i-1+m} \sum_{l=1}^{i-1} \delta(\psi_l = \psi_i) \quad (3)$$

where  $\delta$  denotes the Dirac delta function. Applying this formula, the results of Lo (1984, Lemma 2) and Liu (1996, Theorem 1), we can write the likelihood as

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

where  $C$  defines the subclusters,  $\mathbf{y}_{(j)}$  is the vector of  $y_i$ s that are in subcluster  $j$ , and  $\psi_j$  is the common parameter for that subcluster. There are  $\mathcal{S}_{n,k}$  different subclusters  $C$ , the Stirling Number of the Second Kind. A subcluster  $C$  is a partition of the sample of size  $n$  into  $k$  groups,  $k = 1, \dots, n$ , and since the grouping is done nonparametrically rather than on substantive criteria, we call these “subclusters” to distinguish these from substantively determined clusters that may exist in the data. That is, it is likely that any real underlying clusters would be broken up into multiple subclusters by the nonparametric fit since there is little penalty for over-separation of these subclusters. Thus, the subclustering process assigns different normal parameters across groups and the same parameters within groups: cases are iid only if they are assigned to the same subcluster.

Each subcluster  $C$  can be associated with an  $n \times k$  matrix  $\mathbf{A}$  defined by

$$A = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$$

where  $a_i$  is a  $1 \times k$  vector corresponding to  $Y_i$ . The vector  $a_i$  has all zeros except for a 1 in the position corresponding the group to which  $Y_i$  is assigned. Note that the column sums of  $\mathbf{A}$  are  $(n_1, n_2, \dots, n_k)$ , the number of observations in the groups, and there are  $\mathcal{S}_{n,k}$  such matrices. Specifically, if the subcluster  $C$  is partitioned into groups  $\{S_1, \dots, S_k\}$ , then if  $i \in S_j$ ,  $\psi_i = \eta_j$  and the random effect can be rewritten as

$$\boldsymbol{\psi} = A\boldsymbol{\eta}, \quad (4)$$

where  $\boldsymbol{\eta} = (\eta_1, \dots, \eta_k)$  and  $\eta_j \stackrel{iid}{\sim} \phi_0$  for  $j = 1, \dots, k$ . We then can write the likelihood function as

$$L(\boldsymbol{\theta} | \mathbf{y}) = \frac{\Gamma(m)}{\Gamma(m+n)} \sum_{k=1}^n m^k \sum_{A \in \mathcal{A}_k} \prod_{j=1}^k \Gamma(n_j) \int f(\mathbf{y} | \boldsymbol{\theta}, A\boldsymbol{\eta}) \phi_0(\boldsymbol{\eta}) d\boldsymbol{\eta}, \quad (5)$$

where  $\mathcal{A}_k$  is the set of all  $n \times k$  matrices  $\mathbf{A}$  and  $\eta_j \sim \phi_0$ , independent. Note that if the integral in (5) can be done analytically, as will happen when using a normal base measure in model (2), we have effectively eliminated the random effects from the likelihood, replacing them with the  $\mathbf{A}$  matrices, which serve to group the observations.

## 2.2 A Linear Dirichlet Random Effects Model

We now focus on the simpler case of linear mixed models and, for ease of comparison and to minimize the algebraic load, we consider a special case of (1), the oneway mixed effects model where

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

where  $\mu$  is the fixed effect (intercept) and  $\psi_i$  are the subject specific random effects that the  $i$ th case shares with other cases assigned to the same subcluster. We further assume that  $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$  and the  $\psi_i$  are independent draws from a Dirichlet process with base measure  $\mathcal{N}(0, c\sigma^2)$ . It then follows from the development in Section 2.1 that, conditional on the subcluster matrix  $\mathbf{A}$ , the vector of observations has distribution  $\mathbf{Y}|\mathbf{A} \sim \mathcal{N}(\mu\mathbf{1} + \mathbf{A}\boldsymbol{\eta}, \sigma^2\mathbf{I})$ , where  $\boldsymbol{\eta}_{k \times 1}$  is normally distributed. The complete specification of the model is

$$\begin{aligned} \mathbf{Y}|\mu, \boldsymbol{\eta}, \sigma^2, \mathbf{A} &\sim \mathcal{N}(\mu\mathbf{1} + \mathbf{A}\boldsymbol{\eta}, \sigma^2\mathbf{I}) & \boldsymbol{\eta}|\sigma^2 &\sim \mathcal{N}_n(\mathbf{0}, c\sigma^2\mathbf{I}_K) \\ \mu|\sigma^2 &\sim \mathcal{N}(0, v\sigma^2) & \sigma^2 &\sim \mathcal{IG}(a, b), \end{aligned} \quad (6)$$

By marginalizing the random effects from the joint distribution of response and random effects, we have

$$\mathbf{Y}|\mu, \sigma^2, \mathbf{A} \sim \mathcal{N}(\mu\mathbf{1}, \sigma^2(\mathbf{I} + c\mathbf{A}\mathbf{A}')), \quad \mu|\sigma^2 \sim \mathcal{N}(0, v\sigma^2), \quad \text{and } \sigma^2 \sim \mathcal{IG}(a, b).$$

The joint posterior distribution is given by

$$\pi(\mu, \sigma^2|\mathbf{Y}, \mathbf{A}) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n+1}{2}+a+1} \exp\left\{-\frac{b}{\sigma^2} - \frac{1}{2v\sigma^2}\mu^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mu\mathbf{1})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mu\mathbf{1})\right\},$$

where  $\boldsymbol{\Sigma} = [\mathbf{I} - \mathbf{A}(\frac{1}{c}\mathbf{I} + \mathbf{A}'\mathbf{A})^{-1}\mathbf{A}']^{-1} = \mathbf{I} + c\mathbf{A}\mathbf{A}'$ . Straightforward but tedious manipulations allow us to write the joint posterior as

$$\pi(\mu, \sigma^2|\mathbf{Y}, \mathbf{A}) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n+1}{2}+a+1} \exp\left\{-\frac{N_D}{2\sigma^2}(\mu - \delta_D(\mathbf{y}))^2\right\} \exp\left(-\frac{b}{\sigma^2} - \frac{1}{2\sigma^2}\mathbf{y}'\mathbf{B}_D^{-1}\mathbf{y}\right),$$

where

$$N_D = \sum_{k=1}^K \frac{n_k}{1 + cn_k} + \frac{1}{v} \quad \text{and} \quad \delta_D(\mathbf{y}) = \frac{1}{N_D} \sum_{k=1}^K \frac{n_k}{1 + cn_k} \bar{y}_k,$$

and  $\mathbf{B}_D = \mathbf{I} + c\mathbf{A}\mathbf{A}' + v\mathbf{1}\mathbf{1}'$ . This yields the full conditional distributions

$$\begin{aligned} \mu|\sigma^2, \mathbf{Y}, \mathbf{A} &\sim \mathcal{N}\left(\delta_D(\mathbf{y}), \frac{\sigma^2}{N_D}\right) \\ \sigma^2|\mu, \mathbf{Y}, \mathbf{A} &\sim \text{IG}\left(\frac{n+1}{2} + a, b + \frac{N_D}{2}(\mu - \delta_D(\mathbf{y}))^2 + \frac{1}{2}\mathbf{y}'\mathbf{B}_D^{-1}\mathbf{y}\right), \end{aligned}$$

and by respectively integrating out  $\mu$  and  $\sigma^2$ , we now obtain the marginal posterior distributions

$$\begin{aligned} \pi(\sigma^2|\mathbf{Y}, \mathbf{A}) &\propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+a+1} \exp\left(-\frac{b}{\sigma^2} - \frac{1}{2\sigma^2}\mathbf{y}'\mathbf{B}_D^{-1}\mathbf{y}\right) \\ \pi(\mu|\mathbf{Y}, \mathbf{A}) &\propto \left[b + \frac{N_D}{2}(\mu - \delta_D(\mathbf{y}))^2 + \frac{1}{2}\mathbf{y}'\mathbf{B}_D^{-1}\mathbf{y}\right]^{-\left(\frac{n+1}{2}+a\right)}. \end{aligned} \quad (7)$$

We note that the distribution of  $\mu$  is a transformed Student's  $t$ , while for  $\sigma^2$  we have

$$\sigma^2 | \mathbf{Y}, \mathbf{A} \sim \text{IG} \left( \frac{n}{2} + a, b + \frac{1}{2} \mathbf{y}' \mathbf{B}_D^{-1} \mathbf{y} \right),$$

leading to straightforward simulation.

Thus, the posterior variance  $\sigma^2$  in the linear Dirichlet mixed model has a mean that is proportional to  $\mathbf{y}' \mathbf{B}_D^{-1} \mathbf{y}$ , and it is this quantity that we focus on. In fact, both the posterior variance of  $\mu$  and the posterior mean of  $\sigma^2$  in a linear Dirichlet mixed model have the form:

$$c_d \times \left( b + \frac{1}{2} \mathbf{y}' [\mathbf{I} + c \mathbf{A} \mathbf{A}' + v \mathbf{1} \mathbf{1}']^{-1} \mathbf{y} \right), \quad (8)$$

where  $c_d > 0$  is a constant.

### 3 Comparing Posterior Variances

We compare the posterior variances of  $\mu$  for linear mixed model with Dirichlet random effects to that with normal random effects. We first describe an eigenvalue inequality that guarantees the Dirichlet variances are smaller, then we prove a matrix theorem that shows when the inequality holds. We verify that the Dirichlet model satisfies the conditions of the theorem, and indicate how the results can be generalized.

#### 3.1 Eigenvalues

From (6), we obtain the normal random effects model as a special case by setting  $K = n$  and  $\mathbf{A} = \mathbf{I}$ . Thus, under the normal model the variance has posterior distribution

$$\sigma^2 | \mathbf{Y} \sim \text{IG} \left( \frac{n}{2} + a, b + \frac{1}{2} \mathbf{y}' \mathbf{B}_N^{-1} \mathbf{y} \right),$$

with  $\mathbf{B}_N = (1 + c) \mathbf{I} + v \mathbf{1} \mathbf{1}'$ .

We now see if the mean of the posterior distribution of  $\sigma^2$ , using the Dirichlet, is smaller than the corresponding mean for the normal model, that is, we want to show that

$$\frac{\mathbf{y}' \mathbf{B}_N^{-1} \mathbf{y}}{\mathbf{y}' \mathbf{B}_D^{-1} \mathbf{y}} = \frac{\mathbf{y}' [(c + 1) \mathbf{I} + v \mathbf{1} \mathbf{1}']^{-1} \mathbf{y}}{\mathbf{y}' [\mathbf{I} + c \mathbf{A} \mathbf{A}' + v \mathbf{1} \mathbf{1}']^{-1} \mathbf{y}} \geq 1,$$

which is equivalent to showing

$$\lambda_{\min} (\mathbf{B}_N^{-1} \mathbf{B}_D) = \lambda_{\min} \left( [(c + 1) \mathbf{I} + v \mathbf{1} \mathbf{1}']^{-1} [\mathbf{I} + c \mathbf{A} \mathbf{A}' + v \mathbf{1} \mathbf{1}'] \right) \geq 1,$$

where  $\lambda_{\min}(\cdot)$  denotes the smallest characteristic root of a matrix. First, note that

$$[(c + 1) \mathbf{I} + v \mathbf{1} \mathbf{1}']^{-1} = a \mathbf{I} - b \mathbf{J},$$

where  $\mathbf{J}$  is an  $n \times n$  matrix of 1s and

$$a = \frac{1}{c+1} \quad b = \frac{v}{(c+1)(c+1+nv)} = a \frac{v}{(c+1+nv)}. \quad (9)$$

Thus,

$$\begin{aligned} \mathbf{B}_N^{-1} \mathbf{B}_D &= a\mathbf{I} + \{(a-nb)v-b\} \mathbf{J} + ac\mathbf{A}\mathbf{A}' - bc\mathbf{A}\mathbf{A}'\mathbf{J} \\ &= a(\mathbf{I} + c\mathbf{A}\mathbf{A}') - bc\mathbf{J}(\mathbf{A}\mathbf{A}' - \mathbf{I}) \end{aligned} \quad (10)$$

because  $(a-nb)v-b = \frac{cv}{(c+1)(c+1+nv)} = bc$ .

Next we describe all of the eigenvectors and eigenvalues of  $\mathbf{B}_N^{-1} \mathbf{B}_D$ . Without loss of generality we assume that the  $\mathbf{A}$  matrix is arranged as

$$\mathbf{A} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_k} \end{pmatrix},$$

and we can now classify the eigenvalues of  $\mathbf{B}_N^{-1} \mathbf{B}_D$  into two groups, as follows:

1. There are  $n-k$  eigenvectors that correspond to contrasts within the groups. One set of these can be constructed with pairwise differences, as the following example shows. Suppose  $n=9, k=3$  and  $n_1=4, n_2=3, n_3=2$ . The following  $n-k=6$  vectors are eigenvectors of  $\mathbf{B}_N^{-1} \mathbf{B}_D$ :

	$n_1$	$n_2$	$n_3$
1)	$\{(1, -1, 0, 0)\}$	$(0, 0, 0)$	$(0, 0)$
2)	$\{(1, 0, -1, 0)\}$	$(0, 0, 0)$	$(0, 0)$
3)	$\{(1, 0, 0, -1)\}$	$(0, 0, 0)$	$(0, 0)$
4)	$\{(0, 0, 0, 0)\}$	$(1, -1, 0)$	$(0, 0)$
5)	$\{(0, 0, 0, 0)\}$	$(1, 0, -1)$	$(0, 0)$
6)	$\{(0, 0, 0, 0)\}$	$(0, 0, 0)$	$(1, -1)$

An eigenvector,  $x$ , of this form satisfies  $\mathbf{A}x = \mathbf{J}x = 0$ , and thus all of these eigenvectors have eigenvalue equal to  $a = 1/(1+c)$ .

2. The remaining  $k$  eigenvectors are of the form

$$L = \begin{pmatrix} \omega_1 \mathbf{1}_{n_1} \\ \omega_2 \mathbf{1}_{n_2} \\ \vdots \\ \omega_k \mathbf{1}_{n_k} \end{pmatrix},$$

for constants  $\omega_1, \dots, \omega_k$  satisfying  $\sum_{j=1}^k n_j \omega_j^2 = 1$ .

So we see that if the data vector  $\mathbf{y}$  consists solely of a contrast within one of the subclusters, the variance of the normal model will be smaller. However, for cases other than this the variance inequality will go the other way, as the following development shows. Direct matrix multiplication shows that for vectors of the form of  $L$  we have

$$L' \mathbf{B}_N^{-1} \mathbf{B}_D L = a \sum_{j=1}^k n_j (1 + cn_j) \omega_j^2 - bc \left[ \sum_{j=1}^k n_j (n_j - 1) \omega_j \right] \left[ \sum_{j=1}^k n_j \omega_j \right] = L' M L,$$

where  $\mathbf{D}(a_j)$  is a diagonal matrix with diagonal elements  $(a_1, \dots, a_k)$  and

$$M = a \mathbf{D}(n_j[1 + cn_j]) - bc [n_1(n_1 - 1) \cdots n_k(n_k - 1)]' (n_1 \cdots n_k).$$

Subject to the constraint  $\sum_{j=1}^k n_j \omega_j^2 = 1$ , the minimum of this quadratic form is the smallest root of the matrix  $M \mathbf{D}(1/n_j)$ . Next, some straightforward manipulations allow us to write

$$M \mathbf{D}(1/n_j) = a \mathbf{D}(1 + cn_j) - bc \mathbf{D}(n_j) \mathbf{D}(n_j - 1) \mathbf{1} \mathbf{1}'. \quad (11)$$

In the next section we develop a matrix result that will characterize the eigenvalues of this matrix.

### 3.2 A Matrix Theorem

Searle (1982, page 116) shows that for a diagonal matrix  $\mathbf{D}$  with nonzero diagonal elements, the determinant of  $\mathbf{D} + \mathbf{1} \mathbf{1}'$  is given by  $|\mathbf{D} + \mathbf{1} \mathbf{1}'| = (\prod d_i) (1 + \sum(1/d_i))$ , which is equal to the product of the eigenvalues. The more relevant version of this equation is  $|\mathbf{D} - \mathbf{1} \mathbf{1}'| = (\prod d_i) (1 - \sum(1/d_i))$ . However, Searle does not give the eigenvalues of either matrix. With some minor conditions on  $d_j$  we can exhibit the eigenvalues.

**Theorem 1** *Let  $\mathbf{D}$  be a  $k \times k$  diagonal matrix with elements  $d_i$  satisfying (i)  $d_i > 1$  for all  $i$  and (ii)  $\sum_i (d_i - 1)^{-1} < 1$ . Then the eigenvalues of the matrix  $\mathbf{D} - \mathbf{1} \mathbf{1}'$  are given by*

$$\lambda_j = d_j \left( 1 - \sum_i (1/d_i) \right)^{r_j}, \quad j = 1, \dots, k, \quad \text{where } \sum_j r_j = 1. \quad (12)$$

The  $r_j$ s are solutions to the equations

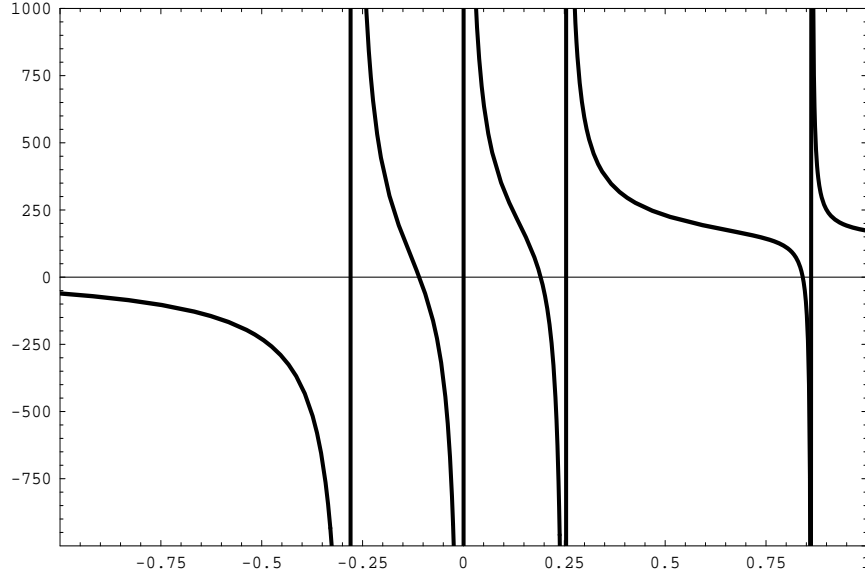
$$\sum_i \frac{1}{d_i - d_j (1 - \sum_i (1/d_i))^{r_j}} = 1, \quad j = 1, \dots, k. \quad (13)$$

Moreover,  $\lambda_j \geq 1$  for  $j = 1, \dots, k$ .

**Proof:** If the  $\lambda_j$  are of the form in (12), the defining eigenvalue equation, for a fixed  $j$ , is

$$(\mathbf{D} - \mathbf{1} \mathbf{1}') x = d_j \left( 1 - \sum_i (1/d_i) \right)^{r_j} x \Rightarrow x_i = \frac{1}{d_i - d_j (1 - \sum_i (1/d_i))^{r_j}},$$

Figure 1: For  $\mathbf{n} = \{10, 7, 5, 2, 1\}$ , a graph of the left side of (13) as a function of  $r_j$ , with  $j = 2$ .



which is satisfied for  $r_j$  satisfying (13). Note that condition (ii) insures that  $\sum_i(1/d_i) < 1$ . If these equations have solutions, these are the eigenvalues, and the determinant formula guarantees that  $\sum_j r_j = 1$ . Moreover, suppose that  $\lambda_j < 1$  for some  $j$ . Then, for that  $j$ , we have  $d_i - d_j(1 - \sum_i(1/d_i))^{r_j} > d_i - 1$ , so the left side of (13) is less than  $\sum_i(d_i - 1)^{-1} < 1$ , and equality cannot be attained.

It only remains to show that there exist  $(r_1, \dots, r_k)$  that solve the equations in (13). In fact there are many solutions, characterized by arguments similar to the following. Assume that  $d_1 \leq d_2 \leq \dots \leq d_k$ . For fixed  $j$ , the function  $[1 - \sum_i(1/d_i)]^{r_j}$  increases to 1 as  $r_j$  decreases to 0 and, at 0, the left side of (13) is  $+\infty$ . Let  $r_j^*$  satisfy  $d_j[1 - \sum_i(1/d_i)]^{r_j^*} = d_{j-1}$ , then as  $r_j : r_j^* \rightarrow 1$ , the left side of (13) goes from  $+\infty \rightarrow -\infty$ , and the equation has a solution.  $\square$

As an example, Figure 1 is a graph of the left side of (13) as a function of  $r_j$ , showing the multiplicity of solutions.

The following corollary covers a more general form of the matrix, which is directly applicable to our matrix (11)

**Corollary 1** *Let  $\mathbf{D}$  and  $\mathbf{H}$  be a  $k \times k$  diagonal matrices with diagonal elements  $d_i$  and  $h_i$  satisfying (i)  $d_i > 1$  and  $h_i > 0$  for all  $i$ , and (ii)  $\sum_i h_i(d_i - 1)^{-1} < 1$ . Then the eigenvalues of the matrix  $\mathbf{D} - \mathbf{H}\mathbf{1}\mathbf{1}'$  are given by*

$$\lambda_j = d_j \left( 1 - \sum_i (h_i/d_i) \right)^{r_j}, \quad j = 1, \dots, k, \quad \text{where } \sum_j r_j = 1. \quad (14)$$

The  $r_j$ s are solutions to the equations

$$\sum_i \frac{h_i}{d_i - d_j (1 - \sum_i (h_i/d_i))^{r_j}} = 1. \quad (15)$$

Moreover,  $\lambda_j \geq 1$  for  $j = 1, \dots, k$ .

**Proof:** First note that

$$|\mathbf{D} - \mathbf{H}\mathbf{1}\mathbf{1}'| = \left( \prod d_i \right) \left( 1 - \sum (h_i/d_i) \right), \quad (16)$$

which suggests the form of the eigenvalues. The conditions on  $d_i$  and  $h_i$  insure the solutions for the  $r_j$ , and that  $\lambda_j > 1$ .  $\square$

### 3.3 Variance Comparison

Theorem 1 and Corollary 1 characterize the eigenvalues of the matrix (11), and we now can state the variance result.

**Theorem 2** *The mean of the posterior distribution of the variance from the Dirichlet random effects model, given in (7), is smaller than that of the normal random effects model for all  $\mathbf{y}$  not containing a within subcluster contrast.*

**Proof:** The development in Section 3.1 shows that the theorem will be proved if we show that all of the eigenvalues of the matrix (11) are greater than or equal to one. We apply Corollary 1 with

$$d_i = a(1 + cn_j) \text{ and } h_j = bcn_j(n_j - 1).$$

It is clear that all  $d_i$  are positive, and thus we only need show that  $\sum_i h_i(d_i - 1)^{-1} < 1$ . Recalling the definitions of  $a$  and  $b$  from (9), we have for  $c > 0$  and  $v > 0$ ,

$$\begin{aligned} \sum_j \frac{h_j}{d_j - 1} &= \sum_j \frac{bcn_j(n_j - 1)}{a(1 + cn_j) - 1} = \sum_j \frac{cvn_j(n_j - 1)}{(1 + c + vn)(1 + cn_j - 1 - c1)} \\ &= \sum_j \frac{vn_j}{(1 + c + vn)} < \sum_j \frac{n_j}{n} \leq 1, \end{aligned} \quad (17)$$

insuring that all eigenvalues of  $M\mathbf{D}(1/n_j)$  are at least 1. Finally note that the minimum eigenvalue 1 is attained if some  $n_j = 1$ , which is evident from the form of the matrix (11).  $\square$

As an example, for  $n = 25$  and  $k = 5$ , we generated all of the partitions of  $n$  into  $k$  subsets. There are 192 such sets, and the eigenvalues are distributed as follows with the associated minimum:

$\min_j n_j$	Frequency	$\lambda_{\min}$
1	108	1
2	54	1.0466 – 1.3450
3	23	1.0452 – 1.0506
4	6	1.0501 – 1.0516
5	1	1.0520.

### 3.4 Generalization

Here we outline how these results might be generalized beyond the model (6), replacing  $\mu\mathbf{1}$  with  $\mathbf{X}\boldsymbol{\beta}$ , where  $\mathbf{X}$  is a known design matrix. This yields

$$\begin{aligned} \mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\eta}, \sigma^2, \mathbf{A} &\sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{A}\boldsymbol{\eta}, \sigma^2\mathbf{I}) & \boldsymbol{\eta}|\sigma^2 &\sim \mathcal{N}_n(\mathbf{0}, c\sigma^2\mathbf{I}_K) \\ \boldsymbol{\beta}|\sigma^2 &\sim \mathcal{N}(\mathbf{0}, v\sigma^2\mathbf{I}) & \sigma^2 &\sim \text{IG}(a, b). \end{aligned} \quad (18)$$

The matrix algebra is now more involved, making it more difficult to describe the eigenvalues of the relevant matrix. Analogous to (10), we now have

$$\begin{aligned} \mathbf{B}_N^{-1}\mathbf{B}_D &= [(c+1)\mathbf{I} + v\mathbf{X}\mathbf{X}']^{-1}[\mathbf{I} + c\mathbf{A}\mathbf{A}' + v\mathbf{X}\mathbf{X}'] \\ &= a(\mathbf{I} - \mathbf{X}(\frac{1}{av}\mathbf{I} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{I} + c\mathbf{A}\mathbf{A}' + v\mathbf{X}\mathbf{X}'). \end{aligned}$$

Now consider a vector  $w$  with  $\mathbf{X}w = 0$ , so  $w$  is not in the column space of  $\mathbf{X}$ . Multiplication then shows that  $w$  is an eigenvector of  $\mathbf{B}_N^{-1}\mathbf{B}_D$  only if it is an eigenvector of  $a(\mathbf{I} + c\mathbf{A}\mathbf{A}')$ . This puts us back in the case covered in Section 3.1, and either  $w$  contains within cluster contrasts, or the eigenvalues are greater than 1. If  $w$  is in the column space of  $\mathbf{X}$ , then  $w = \mathbf{X}t$ , and we can write  $w$  as a linear combination of the eigenvectors of  $\mathbf{X}'\mathbf{X}$ . Suppose, for simplicity, that  $t$  is an eigenvector of  $\mathbf{X}'\mathbf{X}$  with eigenvalue  $\lambda$ . Then  $\mathbf{X}'\mathbf{X}t = \lambda t$ ,  $\mathbf{X}(\frac{1}{av}\mathbf{I} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}t = \frac{v\lambda}{c+1+v\lambda}\mathbf{X}t$ , and

$$\mathbf{B}_N^{-1}\mathbf{B}_D w = \mathbf{B}_N^{-1}\mathbf{B}_D \mathbf{X}t = \frac{1}{c+1+v\lambda} [(1+v\lambda)\mathbf{I} + c\mathbf{A}\mathbf{A}'] \mathbf{X}t,$$

and now we argue as before. For  $w = \mathbf{X}t$  to be an eigenvector of  $\mathbf{B}_N^{-1}\mathbf{B}_D$  either it has a within subcluster contrast, or the eigenvalue is greater than one.

This argument is simplified in that the general vector  $w$  would be decomposed into a part in the column space of  $\mathbf{X}$ , and an orthogonal part, and the part in the column space of  $\mathbf{X}$  would then be represented by a linear combination of eigenvectors. So a full description of the eigenvalues in the general case is somewhat involved, but follows the same pattern that we see in the simpler case.

## 4 Discussion

We have derived a sufficient condition on the data vector  $\mathbf{y}$  to insure that the posterior variance from the Dirichlet random effects model is smaller than that from the normal random effects model. Although the condition is formally unverifiable (since we do not observe  $\mathbf{A}$ ), in practice this is not the case. The Dirichlet posterior variance might only be bigger if the  $\mathbf{y}$  vector has a within-subcluster contrast, and in most cases we will not be able to find any subset of the  $\mathbf{y}$  vector that sums to zero. Moreover, as pointed out by the referee, under the model (6), the set of  $\mathbf{y}$  containing a within-subcluster contrast has measure zero, so the Dirichlet posterior variance is almost surely smaller than that of the normal random effects model.

We note that our results hold for Dirichlet priors on the random effects, and not for a *Mixture of Dirichlet Processes (MDP)*. In the latter case the Dirichlet process is the error distribution, with possibly additional hyperparameters for the base measure. Our model (2) is a *Dirichlet Process Mixture (DPM)*, which has a latent variable modeled with a Dirichlet process prior. (See Ghosal *et al.* 1999 and Ghosal 2009 for details.)

The results here give a theoretical justification to the belief that the richer Dirichlet random effects model is able to remove more extraneous variability, resulting in tighter credible intervals. This result has been observed in data examples, and now we understand that we can almost always expect shorter intervals when using the Dirichlet model.

## References

- Antoniak, C. E. (1974) "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems" *Annals of Statistics* **2**, 1157-1174.
- Blackwell, D. and MacQueen, J. B. (1973) "Ferguson Distributions Via Pólya Urn Schemes." *Annals of Statistics* **1**, 353-355.
- Burr, D. and Doss, H. (2005). "A Bayesian semi-parametric model for random effects meta-analysis." *Journal of the American Statistical Association* **100**, 242-251
- Dorazio, R. M., Mukherjee, B., Zhang, L., Ghosh, M., Jelks, H. L. and Jordan, F. (2008). "Modeling Unobserved Sources of Heterogeneity in Animal Abundance Using a Dirichlet Process Prior." *Biometrics* **64**, 635-644.
- Ferguson, T. S. (1973) "A Bayesian analysis of Some Nonparametric Problems." *Annals of Statistics* **1**, 209-230.
- Ghosal, S. (2009). Dirichlet process, related priors and posterior asymptotics. To appear in *Bayesian Nonparametrics in Practice*. Hjort, N. L., et al., eds. Cambridge University Press.
- Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1999). Consistent semiparametric Bayesian inference about a location parameter. *Journal of Statistical Planning and Inference* **77**, 181-193.
- Gill, J. and Casella, G. (2009) "Nonparametric Priors for Ordinal Bayesian Social Science Models: Specification and Estimation." *Journal of the American Statistical Association* **104**, 453-464.
- Jiang, J. (2007) *Linear and Generalized Linear Mixed Models and Their Applications*, Springer.
- Korwar, R. M. and Hollander, M. (1973) "Contributions to the Theory of Dirichlet Processes." *Annals of Probability* **1**, 705-711
- Kyung, M, Gill, J., and Casella G. (2009). "Estimation in Dirichlet Random Effects Models." To appear in the *Annals of Statistics*. Available at <http://www.stat.ufl.edu/~casella/Papers/GLMDM3-Linear.pdf>
- Kyung, M, Gill, J., and Casella G. (2009). "Sampling Schemes for Generalized Linear Dirichlet Random Effects Models." Technical Report, University of Florida, Department of Statistics. Available at [http://www.stat.ufl.edu/~casella/Papers/sampling\\_schemes-1A.pdf](http://www.stat.ufl.edu/~casella/Papers/sampling_schemes-1A.pdf)
- Liu, J. S. (1996) "Nonparametric Hierarchical Bayes Via Sequential Imputations." *Annals of Statistics* **24**, 911-930
- Lo, A. Y. (1984) "On A Class of Bayesian Nonparametric Estimates: I. Density Estimates." *Annals of Statistics* **12**, 351-357

McCulloch, C. E. and Searle, S. R. (2001) *Generalized, Linear, and Mixed Models*, Wiley, New York.

Sethuraman, J. (1994). "A Constructive Definition of Dirichlet Priors." *Statistica Sinica* **4**, 639-650.

Searle, S. R. (1982). *Matrix Algebra Useful for Statistics*. New York: John Wiley.