LINEAR MIXED MODEL ESTIMATION WITH DIRICHLET PROCESS RANDOM EFFECTS

By

CHEN LI

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2012

To my parents and my sister

ACKNOWLEDGMENTS

I would like to sincerely thank my advisor Dr. George Casella for his guidance, patience and help. I feel very lucky to get to know him and learn under him. I learn a lot from him, not only about the knowledge but also about his work ethic and the attitude to the life.

I would like to thank everyone on my supervisory committee: Dr. Malay Ghosh, Dr. Linda Young and Dr. Volker Mai for their guidance and encouragement. Their suggestions and help make a big impact on this dissertation.

I would like to thank the faculty at the Department of Statistics. They taught me a lot, both in and out of the classroom. I am very lucky to be a graduate student at the Department of Statistics.

I would like to thank all my friends both in USA and in China for their friendship and support.

I thank my parents for their support and confidence in me. Without their support and encouragement, I would not have the courage and ability to pursue my dreams. I also want to thank my sister Yan, my brother-in-law Jian, my brother Yang and my nephew Ziyang for their consistent help and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

LINEAR MIXED MODEL ESTIMATION WITH DIRICHLET PROCESS RANDOM EFFECTS

By

Chen Li

August 2012

Chair: George Casella
Major: Statistics

The linear mixed model is very popular, and has proven useful in many areas of applications. (See, for example, McCulloch and Searle (2001), Demidenko (2004), and Jiang (2007).)

Usually people assume that the random effect is normally distributed. However, as this distribution is not observable, it is possible that the distribution of the random effect is non-normal (Burr and Doss (2005), Gill and Casella (2009), Kyung et al. (2009, 2010)). We assume that the random effect follows a Dirichlet process, as discussed in Burr and Doss (2005), Gill and Casella (2009), Kyung et al. (2009, 2010).

In this dissertation, we first consider the Dirichlet process as a model for classical random effects, and investigate their effect on frequentist estimation in the linear mixed model. We discuss the relationship between the BLUE (Best Linear Unbiased Estimator) and OLS (Ordinary Least Squares) in Dirichlet process mixed models, and also give conditions under which the BLUE coincides with the OLS estimator in the Dirichlet process mixed model.

In addition, we investigate the model from the Bayesian view, and discuss the properties of estimators under different model assumptions, compare the estimators under the frequentist model and different Bayesian models, and investigate minimaxity. Furthermore, we apply the linear mixed model with Dirichlet Process random effects to a real data set and get satisfactory results.

CHAPTER 1
INTRODUCTION

The linear mixed model is very popular, and has proven useful in many areas of applications. (See, for example, McCulloch and Searle (2001), Demidenko (2004), and Jiang (2007).) It is typically written in the form

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

where $\mathbf{Y}$ is an $n \times 1$ vector of responses, $\mathbf{X}$ is an $n \times p$ known design matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector of coefficients, $\mathbf{Z}$ is another known $n \times r$ matrix, multiplying the $r \times 1$ vector $\boldsymbol{\psi}$, a vector of random effects, and $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 \mathbf{I}_n)$ is the error.

It is typical to assume that $\boldsymbol{\psi}$ is normally distributed. However, as this distribution is not observable, it is possible that the distribution of the random effect is non-normal (Burr and Doss (2005), Gill and Casella (2009), Kyung et al. (2009, 2010)). It has now become popular to change the distributional assumption on $\boldsymbol{\psi}$ to a Dirichlet process, as discussed in Burr and Doss (2005), Gill and Casella (2009), Kyung et al. (2009, 2010).

The first use of Dirichlet process as prior distributions was by Ferguson (1973); see also Antoniak (1974), who investigated the basic properties. At about the same time, Blackwell and MacQueen (1973) proved that the marginal distribution of the Dirichlet process was the same as the distribution of the nth step of a Polya urn process. There was not a great deal of work done in this topic in the following years, perhaps due to the difficulty of computing with Dirichlet process priors. The theory was advanced by the work of Korwar and Hollander (1973), Lo (1984), and Sethuraman (1994). However, not until the 1990s and, in particular, the Gibbs sampler, did work in this area take off. There we have contributions from Liu (1996), who furthered the theory, and work by Escobar and West (1995), MacEachern and Muller (1998), and Neal (2000), and others, who used Gibbs sampling to do Bayesian computations. More recently, Kyung et al. (2009) investigated a variance reduction property of the Dirichlet process prior, and Kyung et al.

(2010) provided a new Gibbs sampler for the linear Dirichlet mixed model and discussed estimation of the precision parameter of the Dirichlet process.

Since the 1990s, the Bayesian approach has seen the most use of models with Dirichlet process priors. Here, however, we want to consider the Dirichlet process as a model for classical random effects, and to investigate their effect on frequentist estimation in the linear mixed model.

Many papers discuss the MLE of a mixture of normal densities. For example, Young and Coraluppi (1969) developed a stochastic approximation algorithm to estimate the mixtures of normal density with unknown means and unknown variance. Day (1969) provided a method of estimating a mixture of two normal distribution with the same unknown covariance matrix. Peters and Walker (1978) discussed an iterative procedure to get the MLE for a mixture of normal distributions. Xu and Jordan (1996) discussed the EM algorithm for the finite Gaussian mixtures and discussed the advantages and disadvantages of EM for the Gaussian mixture models.

However, we cannot use the methods mentioned in the above papers to get the MLE in Dirichlet process mixed models. The above papers considered the density $\sum \alpha_i f_i(x|\theta_i)$, where θ_i is a parameter; the proportion α_i is also parameter, independent of θ_i, with $\sum \alpha_i = 1$. However, in the Dirichlet process mixed model, the density considered is $\sum_{\mathbf{A}} P(\mathbf{A}) f_i(x|\mathbf{A})$, where the proportion $P(\mathbf{A})$ depends on the matrix $\mathbf{A}$ and $f_i(x|\mathbf{A})$ also depends on the matrix $\mathbf{A}$. Here $\mathbf{A}$ is a $r \times k$ matrix. The $r \times k$ matrix $\mathbf{A}$ is a binary matrix; each row is all zeros except for one entry, which is a 1, which depicts the cluster to which that observation is assigned. Of course, both k and $\mathbf{A}$ are unknown. We will discuss more details about the matrix $\mathbf{A}$ in next chapter. The weights $P(\mathbf{A})$ are correlated with the corresponding components $f_i(x|\mathbf{A})$. Thus, the methods and results in these papers cannot be used here directly. We do not discuss the MLE here. We will consider other methods to estimate the fixed effects—the best linear unbiased estimator

(BLUE) and ordinary least squares (OLS) for the fixed effects and MINQUE/sample covariance matrix method for the variance components σ^2 and τ^2 here.

The Gauss-Markov Theorem, which finds the BLUE, is given by Zyskind and Martin (1969) for the linear model, and by Harville (1976) for the linear mixed model, where he also obtained the best linear unbiased predictor (BLUP) of the random effects. Robinson (1991) discussed BLUP and the estimation of random effects, and Afshartous and Wolf (2007) focused on the inference of random effects in multilevel and mixed effects models. Huang and Lu (2001) extended Gauss-Markov theorem to include nonparametric mixed-effects models. Many papers have discussed the relationship between OLS and BLUE, with the first results obtained by Zyskind (1967). Puntanen and Styan (1989) discussed this relationship in a historical perspective.

By the Gauss-Markov Theorem, we can write the BLUE for the fixed effects $\boldsymbol{\beta}$ in a closed form. We give the formula of the corresponding variance-covariance matrix, which helps us get the covariance matrix directly. We are concerned with finding the best linear unbiased estimator (BLUE), and seeing when this coincides with the ordinary least squares (OLS) estimator. We provide conditions, called "Eigenvector Conditions" and "Matrix Conditions" respectively, under which there is the equality between the OLS and BLUE. By these theorems, we can just use OLS as the BLUE under many cases, which avoids the difficulties and computational efferts of estimating the variance components σ^2, τ^2 and precision parameter m. In addition, we find that the covariance is directly related to the precision parameter of the Dirichlet process, giving a new interpretation of this parameter. The monotonicity property of the correlation is also investigated. Furthermore, we provide a method to construct confidence intervals.

Another problem in the Dirichlet process mixed model is to estimate the parameters σ^2 and τ^2. In the Dirichlet process mixed model, the distribution of responses is a mixture of normal densities, not a single normal distribution, which might lead to some difficulty when we try to use some methods (for example, maximum likelihood) to

estimate the parameters. We will discuss three methods (MINQUE, MINQE, and sample covariance matrix) to find the estimators for σ^2 and τ^2 and show a simulation study. These three methods do not need the response follows a normal distribution. The simulation study shows that the estimators from the sample covariance matrix are very satisfactory. In addition, we can also get satisfactory estimation of covariance by using the sample covariance matrix method.

In the situation when the variance components are unknown, Kackar and Harville (1981) discussed the construction of estimators with unknown variance components and show that the estimated BLUE and BLUP remain unbiased when the estimators of the standard errors are even and translation-invariant. Other works include Kackar and Harville (1984), who gave a general approximation of mean squared errors when using estimated variances, and Das et al. (2004), who discussed mean squared errors of empirical predictors in general cases when using the ML or REML to estimate the errors. We will show that the estimators for σ^2 and τ^2 by the sample covariance matrix satisfy the even and translation-invariant conditions. So the estimator of $\boldsymbol{\beta}$ (or $\boldsymbol{\psi}$, or their linear combinations) with estimators of σ^2 and τ^2 from the sample covariance matrix is still unbiased. On the other hand, by Das et al. (2004) we know that the estimation by MINQUE also satisfies the even and translation-invariant conditions. Then the estimators of $\boldsymbol{\beta}$ (or $\boldsymbol{\psi}$, or their linear combinations) with estimators of σ^2 and τ^2 from MINQUE are also unbiased. All the results mentioned above will be shown in detail in the Chapter 2.

We have discussed the classical estimation under the Dirichlet process mixed model above. We will also compare the performance of the Dirichlet model with the performance of the classical normal model through some data analysis. We will consider both simulated data sets and a real data set. First, we will consider some simulation studies. Then we will move to apply the Dirichlet model to a real data set. We use both the Dirichlet model and the classical normal model to fit the simulated data and the real

data set, and compare the corresponding results. The results show that the Dirichlet process mixed model is robust and tends to give better results. All the numerical analysis results are listed in the Chapter 3.

The way we used to get the above results is from the frequentist viewpoint. Another way to discuss the Dirichlet process mixed model is from the Bayesian viewpoint. We always put priors on $\boldsymbol{\beta}$ when using Bayesian methods. Different priors and different random effects might lead to different estimators, different MSE and different Bayes risks. We can assume that the random effects follow a normal distribution. We can also assume that the random effects follow the Dirichlet process. We can put a normal distribution prior on $\boldsymbol{\beta}$. We can also put the flat prior on $\boldsymbol{\beta}$. We are interested in the answer to the question: which prior/model is better. The Chapter 4 consider this question. In order to compare the priors and models, we will first give the fours models. We can get the corresponding Bayesian estimators and show the corresponding MSE and Bayes risks of these Bayesian estimators and discuss which model is better. More details in the oneway model are also discussed.

Under the classical normal mixed model, we know the minimax estimators of the fixed effects in some special cases. We want to know if there are still some minimax estimators of the fixed effects under the Dirichlet process mixed model. We will discuss the minimaxity and admissibility of the estimators, and to show the admissibility of confidence intervals under the squared error loss. We will show that $\overline{\mathbf{Y}}$ is minimax in the Dirichlet process oneway model. This result also holds for the multivariate case. The Chapter 5 will discuss these properties.

The dissertation is organized as follows. In Chapter 2 we will derive the BLUE and the BLUP, examine the BLUE-OLS relationship, and look at interval estimation. In Section 2.7 we will give the some methods to estimate the covariance components σ^2 and τ^2 and provide a simulation study to compare the methods. In Chapter 3 we will show the performance of the Dirichlet process mixed model by fitting the simulated data

sets and a real data sets. In Chapter 4 we will discuss the Dirichlet process mixed model from the Bayesian viewpoint. We will compare the models with different priors on β and different random effects to see which one is better. In Chapter 5 we will investigate the minimaxity and admissibility under the Dirichlet process mixed model in some special cases. At last, we will give a conclusion. There is a technical appendix at the end.

CHAPTER 2
POINT ESTIMATION AND INTERVAL ESTIMATION

Here we consider estimation of β in (1–1), where we assume that the random effects follow a Dirichlet process. The Gauss-Markov Theorem (Zyskind and Martin (1969); Harville (1976)) is applicable in this case, and can be used to find the BLUE of β. In Section 2.1, we give the BLUE of $\boldsymbol{\beta}$ and BLUP of $\boldsymbol{\psi}$ and, In Sections 2.2-2.3 we investigate conditions under which OLS is BLUE. In Section 2.4, we give some examples to show the equality between the OLS and the BLUE. In Section 2.5, we give a method to construct confidence intervals. Section 2.6 shows properties under the Dirichlet process oneway model. Section 2.7 discusses the methods to estimate the variance components σ^2 and τ^2.

Consider model (1–1), but now we allow the vector $\boldsymbol{\psi}$ to follow a Dirichlet process with a normal base measure and precision parameter m, $\psi_i \sim \mathcal{DP}(m, N(0, \tau^2))$. Blackwell and MacQueen (1973) showed that if $\psi_1, \psi_2, \ldots$ *are* i.i.d. from $G \sim \mathcal{DP}(m, \phi_0)$, the joint distribution of $\boldsymbol{\psi}$ is a product of the form

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

As discussed in Kyung et al. (2010), this expression tells us that there might be clusters because the value of ψ_i can be equal to one of the previous values with positive probability. The implication of this representation is that the random effects from the Dirichlet process can have common values, and this led Kyung et al. (2010) to use a conditional representation of (1–1) of the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{A}\boldsymbol{\eta} + \boldsymbol{\varepsilon}, \tag{2–1}$$

where $\boldsymbol{\psi} = \mathbf{A}\boldsymbol{\eta}$, $\mathbf{A}$ is a $r \times k$ matrix, $\boldsymbol{\eta} \sim N_k(0, \tau^2\mathbf{I}_k)$, and $\mathbf{I}_k$ is the $k \times k$ identity matrix. The $r \times k$ matrix $\mathbf{A}$ is a binary matrix; each row is all zeros except for one entry, which is a 1, which depicts the cluster to which that observation is assigned. Both k and $\mathbf{A}$ are

unknown, but we do know that if $\mathbf{A}$ has column sums $\{r_1, r_2, ..., r_k\}$, then the marginal distribution of $\mathbf{A}$ (Kyung et al. (2010)), under the $\mathcal{DP}$ random effects model is

$$P(\mathbf{A}) = \pi(r_1, r_2, ..., r_k) = \frac{\Gamma(m)}{\Gamma(m+r)} m^k \prod_{j=1}^{k} \Gamma(r_j). \tag{2–2}$$

If $\mathbf{A}$ is known then (2–1) is a standard normal random effects linear mixed model, and we have

$$\mathrm{E}(\mathbf{Y}|\mathbf{A}) = X\boldsymbol{\beta},\ \mathrm{Var}(\mathbf{Y}|\mathbf{A}) = \sigma^2\mathbf{I}_\mathrm{n} + \tau^2\mathbf{ZAA}'\mathbf{Z}'.$$

When $\mathbf{A}$ is unknown, it still remains that $\mathrm{E}(\mathbf{Y}|\mathbf{A}) = \mathbf{X}\boldsymbol{\beta}$, but now we have

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

as the second term on the right side is zero. It then follows that

$$\mathbf{V} = \mathrm{Var}(\mathbf{Y}) = \sigma^2\mathbf{I}_\mathrm{n} + \tau^2 \sum_A P(\mathbf{A})\mathbf{ZAA}'\mathbf{Z}' = \sigma^2\mathbf{I}_\mathrm{n} + \mathbf{ZWZ}', \tag{2–3}$$

where $\mathbf{W} = \tau^2 \sum_A P(\mathbf{A})\mathbf{AA}' = [w_{ij}]_{r\times r} = \tau^2 E(a_i' a_j) = \mathbf{I}(i \neq j)d\tau^2 + \mathbf{I}(i = j)\tau^2$, and by Appendix C

$$d = \sum_{i=1}^{r-1} im \frac{\Gamma(m+r-1-i)\Gamma(i)}{\Gamma(m+r)} \tag{2–4}$$

and $\mathbf{I}(\cdot)$ is the indicator function.

This is, $\mathbf{W} = \begin{bmatrix} \tau^2 & d\tau^2 & \cdots & d\tau^2 \\ d\tau^2 & \tau^2 & \cdots & d\tau^2 \\ \vdots & \vdots & \vdots & \vdots \\ d\tau^2 & \cdots & d\tau^2 & \tau^2 \end{bmatrix}$.

Let $\mathbf{V} = [v_{ij}]_{n\times n}$ and $\mathbf{Z} = [z_{ij}]$. For $i \neq j$, $v_{ij} = \sum_k \sum_l z_{ik} w_{kl} z_{jl}$, which might depend on d and $\mathbf{Z}$. Thus the covariance of Y_i and Y_j might depend on $\mathbf{Z}$, τ^2, r and m.

Example 1. We consider a model of the form

$$Y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \psi_i + \varepsilon_{ij},\ \ 1 \leq i \leq r,\ 1 \leq j \leq t, \tag{2–5}$$

which is similar to (1–1), except here we might consider ψ_i to be a subject-specific random effect. If we let $\mathbf{Y} = [Y_{11}, ..., Y_{1t}, ..., Y_{r1}, ..., Y_{rt}]'$, $\mathbf{1}_t = [1, ..., 1]'_{t\times 1}$, and

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

where $n = rt$, then model (2–5) can be written

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{BA}\boldsymbol{\eta} + \boldsymbol{\varepsilon}, \tag{2–6}$$

so $\mathbf{Y}|\mathbf{A} \sim N(\mathbf{1}\mu, \sigma^2\mathbf{I}_n + \tau^2\mathbf{BAA'B'}])$. The BLUE of β is given in (2–9), and has variance $(\mathbf{X'V}^{-1}\mathbf{X})^{-1}$, but now we can evaluate $\mathbf{V}$ by Eq.(2–3) obtaining

$$\mathbf{V} = \begin{bmatrix} \sigma^2\mathbf{I} + \tau^2\mathbf{J} & d\tau^2\mathbf{J} & d\tau^2\mathbf{J} & \cdots & d\tau^2\mathbf{J} \\ d\tau^2\mathbf{J} & \sigma^2\mathbf{I} + \tau^2\mathbf{J} & d\tau^2\mathbf{J} & \cdots & d\tau^2\mathbf{J} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ d\tau^2\mathbf{J} & d\tau^2\mathbf{J} & \cdots & d\tau^2\mathbf{J} & \sigma^2\mathbf{I} + \tau^2\mathbf{J} \end{bmatrix}, \tag{2–7}$$

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

If $i \neq i'$, the correlation is

$$Corr(Y_{i,j}, Y_{i',j'}) = \frac{d\tau^2}{\sigma^2 + \tau^2} = \frac{\tau^2 \sum_{i=1}^{r-1} im \frac{\Gamma(m+r-1-i)\Gamma(i)}{\Gamma(m+r)}}{\sigma^2 + \tau^2}. \tag{2–8}$$

This last expression is quite interesting, as it relates the precision parameter m to the correlation in the observations, a relationship that was not apparent before. Although we are not completely sure of the behavior of this function, we expected that the correlation would be a decreasing function of m. This would make sense, as a bigger value of m implies more clusters in the process, leading to smaller correlations. This is not the case, however, as Figure 2-1 shows. We can establish that d is decreasing when m is either small or large, but the middle behavior is not clear. What we can establish

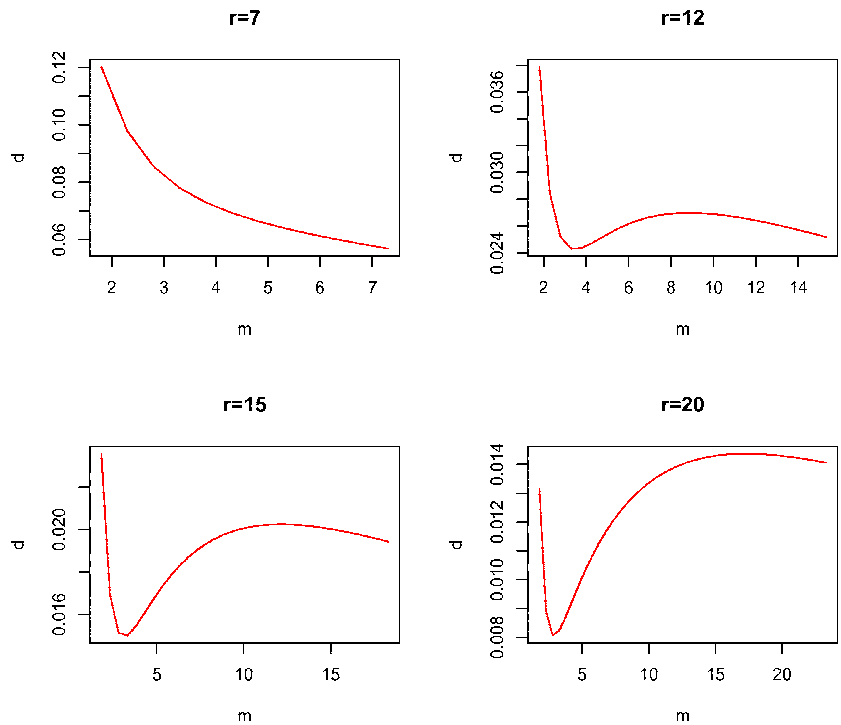

Figure 2-1. The relationship between d and m, with $r = 7, 12, 15, 20$.

about the behavior of d is summarized in the following theorem, whose proof is given in Appendix A

Theorem 2. *Let d be the same as before. Then d is a decreasing in m on $m \geq \sqrt{(r-2)(r-1)}$ or $0 \leq m \leq 2$.*

Proof. See Appendix A. □

2.1 Gauss-Markov Theorem

We can now apply the Gauss-Markov Theorem, as in Harville (1976), to obtain the BLUE of $\boldsymbol{\beta}$ and the BLUP of $\boldsymbol{\psi}$:

$$\widetilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}, \ \widetilde{\boldsymbol{\psi}} = \mathbf{C}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}), \tag{2–9}$$

where $\mathbf{C} = \mathrm{Cov}(\mathbf{Y}, \boldsymbol{\psi}) = \tau^2 \sum_{\mathrm{A}} P(\mathbf{A})\mathbf{ZAA}' = \tau^2\mathbf{ZW}$. It also follows from Harville (1976) that for predicting $\mathbf{w} = \mathbf{L}'\boldsymbol{\beta} + \boldsymbol{\psi}$, for some known matrix $\mathbf{L}'$, such that $\mathbf{L}'\boldsymbol{\beta}$ is estimable, the BLUP of $\mathbf{w}$ is $\widetilde{\mathbf{w}} = \mathbf{L}'\widetilde{\boldsymbol{\beta}} + \mathbf{CV}^{-1}(\mathbf{Y} - \mathbf{X}\widetilde{\boldsymbol{\beta}})$. To use (2–9) to calculate the BLUP requires

either knowledge of $\mathbf{V}$, or the verification that the BLUE is equal to the OLS estimator, and we need to know $\mathbf{C}$ to use the BLUP. Unfortunately, we have neither in the general case.

There are a number of conditions under which these estimators are equal (e.g. Puntanen and Styan (1989); Zyskind (1967)), all looking at the relationship between $\mathrm{Var}(\mathbf{Y})$ and $\mathbf{X}$. For example, when $\mathrm{Var}(\mathbf{Y})$ is nonsingular, one necessary and sufficient condition is that $\mathbf{H}\mathrm{Var}(\mathbf{Y}) = \mathrm{Var}(\mathbf{Y})\mathbf{H}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$, which also implies that $\mathbf{H}\mathrm{Var}(\mathbf{Y})$ is symmetric.

From Zyskind (1967) we know that another necessary and sufficient condition for OLS being BLUE is that a subset of $r_X (r_X = rank(\mathbf{X}))$ eigenvectors of $\mathrm{Var}(\mathbf{Y})$ exists forming a basis of the column space of $\mathbf{X}$.

Since $\mathbf{W} = d\mathbf{J}_r + \tau^2(1-d)\mathbf{I}_r$, where $\mathbf{J}_r$ is a $r \times r$ matrix of ones, we can rewrite the matrix $\mathbf{V}$ as

$$\mathbf{V} = \sigma^2\mathbf{I}_\mathrm{n} + \mathbf{ZWZ}' = \sigma^2\mathbf{I}_\mathrm{n} + d\tau^2\mathbf{ZJ}_\mathrm{r}\mathbf{Z}' + \tau^2(1-d)\mathbf{ZZ}', \tag{2–10}$$

where the matrices $\mathbf{ZJ}_\mathrm{r}\mathbf{Z}'$ and $\mathbf{ZZ}'$ are free of the parameters m, σ and τ. By working with these matrices we will be able to deduce conditions of equality of OLS and BLUE that are free of unknown parameters.

2.2 Equality of OLS and BLUE: Eigenvector Conditions

We first derive conditions on the eigenvectors of $\mathbf{ZZ}'$ and $\mathbf{ZJZ}'$ that imply the equality of the OLS and BLUE. These conditions are not easy to verify, and may not be very useful in practice. However, they do help with the understanding of the structure of the problem, and give necessary and sufficient conditions in a special case.

Let $\mathbf{g}_1 = [s, .., s]^T$ and $\mathbf{g}_2 = [-\sum_{i=1}^{r-1} l_i, l_1, ..., l_{r-1}]^T$, where s and l_i, $i = 1, ..., r-1$ are arbitrary real numbers.

Since $\mathbf{W} = \begin{bmatrix} \tau^2 & d & \cdots & d \\ d & \tau^2 & \cdots & d \\ \vdots & \vdots & \vdots & \vdots \\ d & \cdots & d & \tau^2 \end{bmatrix}$, we know there are two distinct nonzero eigenvalues $\lambda_1 = (r-1)d\tau^2 + \tau^2$ (algebraic multiplicity is 1) and $\lambda_2 = \tau^2 - d\tau^2$ (algebraic multiplicity is r-1), whose corresponding eigenvectors of $\mathbf{W}$ are $\mathbf{g}_1$ and $\mathbf{g}_2$ respectively.

Let $\mathbf{E}_1 = \{\mathbf{g}_1\} \bigcup \{\mathbf{g}_2\}$,

$\mathbf{E}_2 = \{\mathbf{Z}\mathbf{g}_1 : \mathbf{g}_1 \neq \mathbf{0}\}$,

$\mathbf{E}_3 = \{\mathbf{Z}\mathbf{g}_2 : \mathbf{g}_2 \neq \mathbf{0}\}$,

and $\mathbf{E}_4 = \{\mathbf{g} : \mathbf{Z}'\mathbf{g} = \mathbf{0}, \mathbf{g} \neq \mathbf{0}\}$.

We assume $\mathbf{Z}$ is with full column rank, i.e. $rank(\mathbf{Z}) = r$. However, we assume that $\mathbf{Z}'\mathbf{Z}\mathbf{g}_i = \text{a constant} \times \mathbf{g}_i$, $i = 1, 2$. Note that $\mathbf{Z}'\mathbf{Z} = c\mathbf{I}_r$ is a special case of this assumption.

By the form of Eq.(2–3), we know that a vector $\mathbf{g}$ is an eigenvector of $\mathbf{V}$ if and only if it is an eigenvector of $\mathbf{Z}\mathbf{W}\mathbf{Z}'$. The following theorems and corollaries list the eigenvectors of $\mathbf{Z}\mathbf{W}\mathbf{Z}'$, i.e., list the eigenvectors of $\mathbf{V}$. If we know all the eigenvectors of $\mathbf{V}$, we can get a necessary and sufficient condition to guarantee the OLS being the BLUE.

Theorem 3. *Consider the linear mixed model (1–1). Assume that* $\mathbf{Z}$ *satisfies the above assumptions. The OLS is the BLUE if and only if there are* $r_X (r_X = rank(X))$ *elements in the set* $\bigcup_{j=2}^{4} \mathbf{E}_j$ *forming a basis of the column space of* $\mathbf{X}$.

Proof. See Appendix B □

2.3 Equality of OLS and BLUE: Matrix Conditions

It is hard to list forms of all the eigenvectors of $\mathbf{V}$ for a general $\mathbf{Z}$, since we do not know the form of the matrix $\mathbf{Z}$. Also it is hard to check if $\mathbf{H}\mathrm{Var}(\mathbf{Y})$ is symmetric, since $\mathrm{Var}(Y)$ depends on the unknown parameters σ, τ, m. However, we can give some sufficient condition to guarantee the OLS being the BLUE.

Theorem 4. *Consider the model (2–1). Let* $\mathbf{H}$ *is the same as before. We have the following conclusions:*

If $\mathbf{HZJ_rZ'}$ *and* $\mathbf{HZZ'}$ *are two symmetric matrices, then the OLS is the BLUE.*

If $\mathbf{HZJ_rZ'}$ *is symmetric and* $\mathbf{HZZ'}$ *is not symmetric (or if* $\mathbf{HZJ_rZ'}$ *is not symmetric and* $\mathbf{HZZ'}$ *is not symmetric), then the OLS is not BLUE.*

Proof. From the covariance matrix expression (2–10), the conclusions are clear.

If $\mathbf{HZJ_rZ'}$ and $\mathbf{HZZ'}$ are two symmetric matrices, then $\mathbf{HV}$ is symmetric. Thus, the OLS is the BLUE.

If $\mathbf{HZJ_rZ'}$ is symmetric and $\mathbf{HZZ'}$ is not symmetric (or if $\mathbf{HZJ_rZ'}$ is not symmetric and $\mathbf{HZZ'}$ is not symmetric), then $\mathbf{HV}$ is not symmetric. Thus, the OLS is not the BLUE. □

This theorem give us a sufficient condition for the equality of the OLS and the BLUE. The theorem also give us a sufficient condition that the OLS is not the BLUE. However, when both $\mathbf{HZJ_rZ'}$ and $\mathbf{HZZ'}$ are not symmetric, there is no conclusion for the relationship between the OLS and the BLUE.

Corollary 5. *If* $C(\mathbf{Z}) \subseteq C(\mathbf{X})$, *i.e., the column space of* $\mathbf{Z}$ *is contained in the column space of* $\mathbf{X}$, *then* $\mathbf{ZZ'H}$ *and* $\mathbf{ZJZ'H}$ *are symmetric, where* $\mathbf{H}$ *is same as before. Thus, the OLS is the BLUE now.*

Proof. Since $C(\mathbf{Z}) \subseteq C(\mathbf{X})$, there exist a matrix $\mathbf{Q}$ such that $\mathbf{Z} = \mathbf{XQ}$.

Then we have

$$\mathbf{ZZ'H} = \mathbf{XQQ'X'X}(\mathbf{X'X})^{-1}\mathbf{X'} = \mathbf{XQQ'X'},$$

which is symmetric.

$$\mathbf{ZJZ'H} = \mathbf{XQJQ'X'X'X}(\mathbf{X'X})^{-1}\mathbf{X'} = \mathbf{XQJQ'X'},$$

which is also symmetric.

By the discussion above, the OLS is the BLUE. □

2.4 Some Examples

Example 6. Consider a special case of the Example 1: the randomized complete block design: $Y_{ij} = \mu + \alpha_i + \psi_j + \varepsilon_{ij}, 1 \leq i \leq a, 1 \leq j \leq b$, where ψ_j is the effect for being in block j. Assume $\psi_j \sim \mathcal{DP}(m, N(0, \tau^2))$.

Then the model can be written as

$$\mathbf{T} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\psi} + \boldsymbol{\varepsilon},$$

where $\mathbf{X} = \mathbf{B}$, $\mathbf{Z}^T = [\mathbf{I}_b, ..., \mathbf{I}_b]_{b,n}$, and $\boldsymbol{\beta} = [\beta_1, ..., \beta_a]^T = [\mu + \alpha_1, ..., \mu + \alpha_a]^T$.

We can use the theorems discussed above or use the results in Example 1 to check if the OLS is the BLUE.

By straightforward calculation, we have $\mathbf{Z}^T\mathbf{H} = \mathbf{Z}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \frac{1}{b}[\mathbf{1}_b, ..., \mathbf{1}_b]_{b,n}$, where $\mathbf{1}_b$ is a $b \times 1$ vector whose every element is 1.

In addition $\mathbf{Z}\mathbf{Z}'\mathbf{H} = \mathbf{J}_n$, where $\mathbf{J}_n$ is a $n \times n$ matrix whose every element is 1.

Similarly, $\mathbf{Z}\mathbf{J}\mathbf{Z}'\mathbf{H} = b\mathbf{J}_n$.

Thus, by the previous discussion we know that the OLS is the BLUE now.

Example 7. Consider a model:

$$Y_{ijk} = \mathbf{x}_i'\boldsymbol{\beta} + \alpha_i + \gamma_j + \varepsilon_{ijk}, \quad 1 \leq i \leq a, \ 1 \leq j \leq b, 1 \leq k \leq n_{ij}, \tag{2–11}$$

where α_i, γ_j are random effects.

Without loss of generality, assume $a = b = n_{ij} = 2$. Thus,

$$\mathbf{Z}^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

We can use the theorems to see if the OLS is the BLUE.

For example, assume $\mathbf{X}^T = \begin{bmatrix} 1 & -1 & 1 & -1 & 0 & 0 & 2 & -2 \\ 1 & -1 & 0 & 0 & 1 & -1 & 2 & -2 \\ 0 & 0 & -1 & -1 & 1 & -1 & 2 & -2 \end{bmatrix}$.

Then $\mathbf{HZZ}'$ and $\mathbf{HZJZ}'$ are symmetric, where $\mathbf{H}$ is the same as before. Then by the Theorem 4, we know that the OLS is BLUE now.

However, for some other $\mathbf{X}$s, the OLS might not be the BLUE.

For example, if $\mathbf{X}^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 & 1 & -1 & 1 & -1 \\ 0 & 0 & 1 & 1 & 1 & -1 & 1 & -1 \end{bmatrix}$.

For this $\mathbf{X}$, $\mathbf{HZJZ}'$ is symmetric and $\mathbf{H}_1\mathbf{ZZ}'$ is not symmetric. Thus, the OLS is not the BLUE now by the previous discussion.

Example 8. Consider $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\psi} + \boldsymbol{\varepsilon}$.

Assume

$$\mathbf{Z}^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \end{bmatrix}.$$

The $\mathbf{Z}$ matrix satisfies the condition $\mathbf{Z}'\mathbf{Z} = c\mathbf{I}$. We can apply the theorem to check if the OLS is the BLUE.

For example, assume $\mathbf{X}^T = \begin{bmatrix} 1 & -1 & 1 & -1 & 0 & 0 & 2 & -2 \\ 1 & 1 & -1 & -1 & 3 & 3 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$.

By regular algebra calculation we find that the elements in $\bigcup_{j=2}^{4} \mathbf{E}_j$ do form a basis of the column space of $\mathbf{X}$. Thus the OLS is the BLUE.

However, if $\mathbf{X}^T = \begin{bmatrix} 1 & -1 & 1 & -1 & 0 & 0 & 2 & -2 \\ 1 & 1 & -1 & -1 & 3 & 3 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$.

By regular algebra calculation, we know that the elements in $\bigcup_{j=2}^{4} \mathbf{E}_j$ do not form a basis of the column space of $\mathbf{X}$. Thus the OLS is not the BLUE now.

Example 9. Consider a balanced ANOVA model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\psi} + \boldsymbol{\varepsilon}$ when $rank(\mathbf{X}) = length(\boldsymbol{\psi})$. In this case, $\mathbf{X}$ and $\mathbf{B}$ have the same column space. Thus, we can just consider the model $\mathbf{Y} = \mathbf{B}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\psi} + \boldsymbol{\varepsilon}$. Since each column of $\mathbf{B}$ can be written as a linear combination of the eigenvectors ω_1 and ω_2, by the discussion of the Example 1, we have that the OLS is the BLUE in the model $\mathbf{Y} = \mathbf{B}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\psi} + \boldsymbol{\varepsilon}$. In another word, the OLS is the BLUE in the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\psi} + \boldsymbol{\varepsilon}$.

2.5 Interval Estimation

In this section we show how to put confidence intervals on the fixed effects β_i in the general case of model (2–1). Let $\mathbf{G} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$, so the BLUE for $\boldsymbol{\beta}$ is $\widetilde{\boldsymbol{\beta}} = \mathbf{GY}$. If we define $\mathbf{e}_1 = [1, 0, ..., 0]'$, $\mathbf{e}_2 = [0, 1, 0, ..., 0]'$, ..., $\mathbf{e}_p = [0, ..., 0, 1]'$, then the estimate for β_i is $\widetilde{\beta}_i = \mathbf{e}_i^T\mathbf{GY},\ i = 1, 2, ..., p$.

We want to find the $b_i,\ (i = 1, 2, ..., p)$ such that, $P(\widetilde{\beta}_i \leq b_i) = \alpha$, for $0 < \alpha < 1$, and we start with

$$\widetilde{\beta}_i|\mathbf{A} \sim N(\beta_i, \mathbf{e}_i'\mathbf{GV}_\mathsf{A}\mathbf{G}'\mathbf{e}_i), \quad \mathbf{V}_\mathsf{A} = \tau^2\mathbf{ZAA}'\mathbf{Z}' + \sigma^2\mathbf{I}_\mathsf{n},$$

so $\alpha = P(\widetilde{\beta}_i \leq b_i) = \sum_\mathsf{A} P(\mathbf{A})\Phi(\frac{b_i - \beta_i}{\mathbf{e}_i'\mathbf{GV}_\mathsf{A}\mathbf{G}'\mathbf{e}_i})$.

It turns out that we can get easily computable upper and lower bounds on $\mathbf{e}_i'\mathbf{GV}_\mathsf{A}\mathbf{G}'\mathbf{e}_i$, which would allow us to either approximate b_i, or use a bisection.

It is straightforward to check that the matrix $[(n-1)\mathbf{I} + \mathbf{J}] - \mathbf{AA}'$ is always nonnegative definite for every $\mathbf{A}$, and thus by the expression of matrix $\mathbf{V}_A$, we have

$$\sigma^2\mathbf{e}_i'\mathbf{GI}_\mathsf{n}\mathbf{G}'\mathbf{e}_i \leq \mathbf{e}_i'\mathbf{GV}_\mathsf{A}\mathbf{G}'\mathbf{e}_i \leq \mathbf{e}_i'\mathbf{G}(\tau^2\mathbf{Z}[(\mathbf{n}-\mathbf{1})\mathbf{I} + \mathbf{J}]\mathbf{Z}' + \sigma^2\mathbf{I}_\mathsf{n})\mathbf{G}'\mathbf{e}_i.$$

This inequality gives us a lower bond and a upper bond for $\mathbf{e}_i'\mathbf{GV}_\mathsf{A}\mathbf{G}'\mathbf{e}_i$, which can help form the bounding normal distributions.

Now let Z_α^0 and Z_α^1 be the upper α cutoff points from the bounding normal distributions, so we have $Z_\alpha^0 \leq b_i \leq Z_\alpha^1$.

Table 2-1. Estimated cutoff points of Example 10, with $\alpha = 0.95$, $\sigma^2 = \tau^2 = 1$, and $m = 3$. "Iterations" is the number of steps to convergence, Z_α^0 and Z_α^1 are the bounding normal cutoff points, and b_1 is the cutoff point.

	Iterations	z_α^0	z_α^1	b_1	$P(\widetilde{\beta}_1 \leq b_1)$
$r = 4, t = 5$	8	1.531	6.873	2.2614	0.9496
$r = 5, t = 5$	7	1.340	6.768	2.1460	0.9508
$r = 6, t = 6$	8	1.124	6.851	1.9521	0.9508
$r = 7, t = 6$	7	1.013	6.692	1.8559	0.9497
$r = 8, t = 8$	8	0.861	7.032	1.7050	0.9499

Now we can use these endpoints for a conservative confidence interval, or, in some cases, we can calculate (exactly or by Monte Carlo) the cdf of $\tilde{\boldsymbol{\beta}}$. We give a small example.

Example 10. Consider the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\psi} + \boldsymbol{\varepsilon}$, where $\mathbf{X} = [x_1, x_2, x_3]$, $x_1 = [1, ..., 1]'$, $x_2 = [1, 2, ..., n]'$, $x_3 = [1^2, 2^2, ..., n^2]'$. For $\alpha = 0.95$, $\sigma^2 = \tau^2 = 1$, and $m = 3$, we find b_1, such that $\alpha = P(\widetilde{\boldsymbol{\beta}}_1 \leq b_1)$. Details are in the Table 2-1.

We see that the lower bound tends to be closer to the exact cutoff, but this is a function of the choice of m. In general we can use the conservative upper bound, or use an approximation such as $(Z_\alpha^0 + Z_\alpha^1)/2$.

2.6 The Oneway Model

In this section we only consider the oneway model. In this special case we can investigate further properties of the estimator $\overline{\mathbf{Y}}$, such as unimodality, symmetry, and the effect of the precision parameter m.

The oneway model is

$$Y_{ij} = \mathbf{1}\mu + \psi_i + \varepsilon_{ij}, \quad 1 \leq i \leq r, \; 1 \leq j \leq t,$$

i.e.,

$$\overline{\mathbf{Y}}|\mathbf{A} \sim N(\mu, \sigma_A^2), \quad \sigma_A^2 = \frac{1}{n^2}(n\sigma^2 + \tau^2 t^2 \sum_\ell r_\ell^2) \tag{2–12}$$

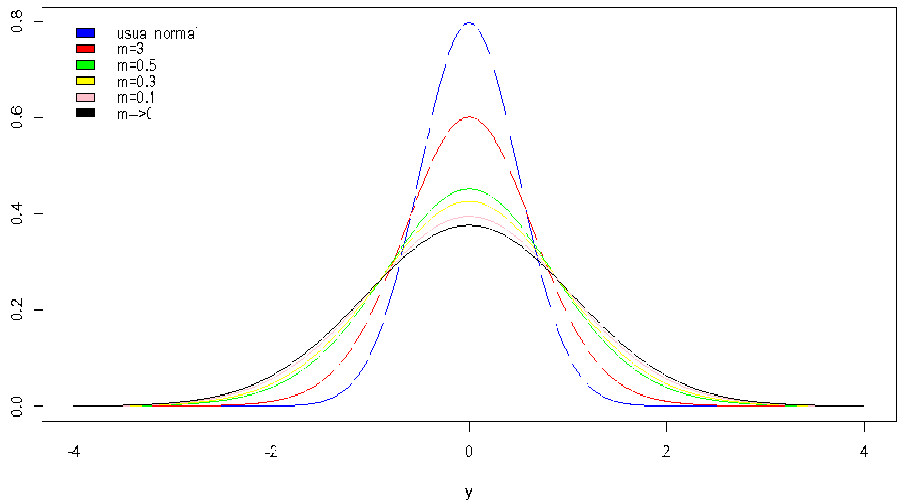

Figure 2-2. Densities of $\overline{Y}$ corresponding to different values of m with $\sigma = \tau = 1$.

where we recall that the r_ℓ are the column sums of the matrix **A**. We denote the density of $\overline{\mathbf{Y}}$ by $f_m(y)$, which can be represented as

$$f_m(y) = \sum_{\mathrm{A}} f(y|\mathbf{A})P(\mathbf{A}), \tag{2–13}$$

where $f(y|\mathbf{A})$ is the normal density with mean μ and variance σ_A^2, and $P(\mathbf{A})$ is the marginal probability of the matrix **A**, as in (2–2). The subscript m is the precision parameter, which appears in the probability of **A**.

Figure 2-2 is a plot of the pdf $f_m(y)$ with $n = 8$, for different m with $\sigma = \tau = 1$. The figure shows that the density of $\overline{Y}$ is symmetric and unimodal. It is also apparent that, in the tails, the densities with smaller m are above those with bigger m.

2.6.1 Variance Comparisons

By the previous result, we know that $\overline{\mathbf{Y}}$ is the BLUE under the Dirichlet process oneway model Here we want to compare the variances of the the BLUE $\overline{\mathbf{Y}}$ under the Dirichlet process oneway model and the classical oneway model, with normal random effects. We will see that $\mathrm{Var}(\overline{\mathbf{Y}})$ under the Dirichlet model is larger than that under the normal model.

The oneway model has the matrix form

$$\mathbf{Y} = \mathbf{1}\mu + \mathbf{BA}\boldsymbol{\eta} + \varepsilon, \tag{2–14}$$

where $\varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, $\boldsymbol{\psi} = \mathbf{A}\boldsymbol{\eta}$, and $\boldsymbol{\eta}_{k\times 1} \sim N(0, \tau^2\mathbf{I}_k)$, and

$$\mathrm{Var}(\overline{\mathbf{Y}}|\mathbf{A}) = \frac{1}{n^2}\sigma^2\mathbf{1}'(\mathbf{I} + \frac{\tau^2}{\sigma^2}\mathbf{BAA}'\mathbf{B}')\mathbf{1}.$$

Recall that the column sums of $\mathbf{A}$ are $(r_1, r_2, \ldots, r_k)$, and denote this vector by $\mathbf{r}$. It is straightforward to verify that $\mathbf{1}'\mathbf{B} = t\mathbf{1}'$, and then $\mathbf{1}'\mathbf{BA} = t\mathbf{r}'$. Recalling that $\sum_j r_j = r$ and $n = rt$, we have

$$\mathbf{1}'(\mathbf{I} + \frac{\tau^2}{\sigma^2}\mathbf{BAA}'\mathbf{B}')\mathbf{1} = n + \frac{\tau^2}{\sigma^2}t^2\mathbf{r}'\mathbf{r} = n + \frac{\tau^2}{\sigma^2}t^2\sum_{j=1}^{k} r_j^2.$$

Thus, the conditional variance under the Dirichlet model is

$$\mathrm{Var}(\overline{\mathbf{Y}}|\mathbf{A}) = \frac{1}{n^2}\left(n\sigma^2 + \tau^2t^2\sum_{j=1}^{k} r_j^2\right). \tag{2–15}$$

Now $\sum_{j=1}^{k} r_j^2 > (\sum_{j=1}^{k} r_j)^2/k$, and since k is necessarily no greater than r, we have $\sum_{j=1}^{k} r_j^2 \geq (\sum_{j=1}^{k} r_j)^2/r$ and

$$\mathrm{Var}(\overline{\mathbf{Y}}|\mathbf{A}) \geq \frac{\sigma^2}{n^2}\left(n + \frac{\tau^2}{\sigma^2}t^2\frac{(\sum_{j=1}^{k} r_j)^2}{r}\right) = \frac{\sigma^2}{n^2}\left(n + \frac{\tau^2}{\sigma^2}t^2 r\right) = \mathrm{Var}(\overline{\mathbf{Y}}|\mathbf{I}),$$

where $\mathrm{Var}(\overline{\mathbf{Y}}|\mathbf{I})$ is just the corresponding variance under the classical oneway model.

Thus, every conditional variance of $\overline{\mathbf{Y}}$ under the Dirichlet model is bigger than the variance in the normal model, so the unconditional variance of $\overline{\mathbf{Y}}$ under the Dirichlet model is also bigger than that under the normal model.

2.6.2 Unimodality and Symmetry

For every $\mathbf{A}$ and every real number y,

$$f(\mu + y|\mathbf{A}) = f(\mu - y|\mathbf{A}), \quad P(\mathbf{A}) \geq 0.$$

Thus, $f_m(\mu + y) = f_m(\mu - y)$, that is, the marginal density is symmetric about the point μ. Also, it is easy to show that

1. If $\mu \geq y_1 \geq y_2$, $f(\mu|\mathbf{A}) \geq f(y_1|\mathbf{A}) \geq f(y_2|\mathbf{A}), \Longrightarrow f_m(\mu|\mathbf{A}) \geq f_m(y_1|\mathbf{A}) \geq f_m(y_2|\mathbf{A})$;
2. If $\mu \leq y_1 \leq y_2$, $f(\mu|\mathbf{A}) \geq f(y_1|\mathbf{A}) \geq f(y_2|\mathbf{A}), \Longrightarrow f_m(\mu|\mathbf{A}) \geq f_m(y_1|\mathbf{A}) \geq f_m(y_2|\mathbf{A})$,

and thus the marginal density is unimodal around the point μ.

2.6.3 Limiting Values of m

Now we look at the limiting cases: $m = 0$ and $m \to \infty$. We will show that $f_m(y)$ remains a proper density when $m = 0$ and $m \to \infty$.

Theorem 11. *When $m = 0$ or $m \to \infty$, the marginal densities $\pi(r_1, r_2, ..., r_k)$ in (2–2) degenerate to a single point. Specifically.*

$$\lim_{m\to\infty} \pi(r_1, r_2, ..., r_k) = \begin{cases} 1, & k = 1, \\ 0, & k = 2, ..., r, \end{cases}$$

and

$$\lim_{m\to 0} \pi(r_1, r_2, ..., r_k) = \begin{cases} 1, & k = r, \\ 0, & k = 1, ..., r-1. \end{cases}$$

It then follows from (2–13) that

$$\begin{aligned} \overline{\mathbf{Y}}|m = 0 &\sim N(\mu, \frac{1}{n}\sigma^2 + \tau^2), \\ \overline{\mathbf{Y}}|m = \infty &\sim N(\mu, \frac{1}{n}(\sigma^2 + \tau^2 t)). \end{aligned}$$

Proof. From (2–2) we can write

$$\pi(r_1, r_2, ..., r_k) = \frac{m^{k-1}}{(m+r-1)(m+r-2)\cdots(m+1)} \prod_{j=1}^{k} \Gamma(r_j).$$

The denominator $(m + r - 1)(m + r - 2)\cdots(m + 1)$ is a polynomial of degree $(r - 1)$, and goes to $(r - 1)!$ when $m \to 0$. Thus $\pi(r_1, r_2, ..., r_k) \to 0$ unless $k = 1$. When $m \to \infty$, $\pi(r_1, r_2, ..., r_k)$ again goes to zero unless $k = r$, making the numerator a polynomial of degree $r - 1$. The densities of $\bar{Y}$ follow from substituting into (2–13). □

When $m = 0$ all of the observations are in the same cluster, the $\mathbf{A}$ matrix degenerates to $(1, 1, \dots, 1)'$. At $m = \infty$, each observation is in it own cluster, $\mathbf{A} = \mathbf{I}$, and the distribution of $\bar{\mathbf{Y}}$ is that of the classic normal random effects model.

2.6.4 Relationship Among Densities of $\overline{\mathbf{Y}}$ with Different m

In this section, we compare the tails of the densities of $\overline{\mathbf{Y}}$ with different parameter m and show that the tails of densities with smaller m are always above the tails of densities with larger m. Recall (2–12), and note that

$$\sigma_\infty^2 = \frac{1}{n}(\sigma^2 + \tau^2 t) \leq \sigma_A^2 = \frac{1}{n^2}(n\sigma^2 + \tau^2 t^2 \sum_\ell r_\ell^2) \leq \frac{1}{n}\sigma^2 + \tau^2 = \sigma_0^2, \tag{2–16}$$

and so σ_0^2 is the largest variance. We can then establish the following theorem.

Theorem 12. *If $m_1 < m_2$, then*

$$\lim_{y\to\infty} \frac{f_{m_2}(y)}{f_{m_1}(y)} < 1. \tag{2–17}$$

Proof. From (2–13), If $y \to \infty$ and $m_1 < m_2$,

$$\begin{aligned}
\frac{f_{m_2}(y)}{f_{m_1}(y)} &= \frac{\sum_A P(\mathbf{A}) \frac{1}{\sqrt{2\pi\sigma_A^{2(m_2)}}} \exp(-\frac{(y-\mu)^2}{2\sigma_A^{2(m_2)}})}{\sum_A P(\mathbf{A}) \frac{1}{\sqrt{2\pi\sigma_A^{2(m_1)}}} \exp(-\frac{(y-\mu)^2}{2\sigma_A^{2(m_1)}})} \\
&= \frac{\sum_A P(\mathbf{A}) \frac{1}{\sqrt{2\pi\sigma_A^{2(m_2)}}} \exp(-\frac{(y-\mu)^2}{2}(\frac{1}{\sigma_A^{2(m_2)}} - \frac{1}{\sigma_0^2}))}{\sum_A P(\mathbf{A}) \frac{1}{\sqrt{2\pi\sigma_A^{2(m_1)}}} \exp(-\frac{(y-\mu)^2}{2}(\frac{1}{\sigma_A^{2(m_1)}} - \frac{1}{\sigma_0^2}))}.
\end{aligned}$$

Since $\sigma_0^2 \geq \sigma_A^2$, the exponential term goes to zero as $y \to \infty$ unless $\sigma_0^2 = \sigma_A^2$. This only happens when $\mathbf{A} = \mathbf{A}_0 = (1, 1, \dots, 1)'$, and thus

$$\lim_{y\to\infty} \frac{f_{m_2}(y)}{f_{m_1}(y)} = \frac{P_{m_2}(\mathbf{A} = \mathbf{A}_0)}{P_{m_1}(\mathbf{A} = \mathbf{A}_0)} = \frac{(m_1 + r - 1)(m_1 + r - 2)\cdots(m_1 + 1)}{(m_2 + r - 1)(m_2 + r - 2)\cdots(m_2 + 1)} < 1.$$

□

Therefore, when y is large enough, the tails of densities with smaller m are always above the tails of densities with larger m. As an application of this theorem, we can compare the densities when $0 < m < \infty$ with the densities in the limiting cases – when

$m = 0$, and $m = \infty$. In fact, for sufficiently large y we have

$$f_\infty(y) \leq f_m(y) \leq f_0(y),$$

and the tails of any density are always between the tails of densities with $m = 0$ and $m = \infty$.

This gives us a method to find the cutoff points in the Dirichlet process oneway model. Since we have bounding cutoff points, we could use the cutoff corresponding to $m = 0$ as a conservative bound. Alternatively, we could use a bisection method if we had some idea of the value of m. We see in Table 2-2 that there is a relatively wide range of cutoff values, and that the conservative cutoff could be quite large.

Table 2-2. Estimated cutoff points ($\alpha = 0.975$) under the Dirichlet model $Y_{ij} = \mu + \psi_i + \varepsilon_{ij},\ \ 1 \leq i \leq 6,\ 1 \leq j \leq 6$, for different values of m. $\sigma^2 = \tau^2 = 1$.

m	0	.1	.5	1	20	∞
Estimated Cutoff	1.987	1.917	1.706	1.566	0.952	0.864

2.7 The Estimation of σ^2, τ^2 and Covariance

Another problem in the Dirichlet process mixed model(or Dirichlet process oneway model) is to estimate the parameters σ^2 and τ^2. In the Dirichlet process mixed model, the distribution of the responses is a mixture of normal densities, not a single normal distribution, which might lead to some difficulty when we try to use some methods (for example, maximum likelihood method) to estimate the parameters.

The following sections will discuss three methods (MINQUE, MINQE, and sample covariance matrix) to find the estimators for σ^2 and τ^2 and show a simulation study. These three methods do not need that the responses follow a normal distribution. The simulation study shows that the estimation from the sample covariance matrix are very satisfactory. In addition, we can also get the estimation of covariance from the sample covariance matrix method.

2.7.1 MINQUE for σ^2 and τ^2

As discussed in Searle et al. (2006), Rao (1979), Brown (1976), Rao (1977) and Chaubey (1980), the minimum norm quadratic unbiased estimation (MINQUE) does not require the normality assumption.

The Dirichlet mixed model $\mathbf{Y}|\mathbf{A} \sim \mathbf{N}(\mathbf{X}\beta, \sigma^2\mathbf{I}_n + \tau^2\mathbf{ZAA}'\mathbf{Z}')$ can be written as

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \tag{2–18}$$

where $\mathrm{Var}(\varepsilon) = \sigma^2\mathbf{I}_n + \tau^2 \sum \mathbf{P}(\mathbf{A})\mathbf{ZAA}'\mathbf{Z}' = \sigma^2\mathbf{I}_n + d\mathbf{ZJ}_r\mathbf{Z}' + (1-d)\tau^2\mathbf{ZZ}'$. Denote $\mathbf{T}_1 = \mathbf{I}_n$, $\mathbf{T}_2 = \mathbf{ZJ}_r\mathbf{Z}'$, and $\mathbf{T}_3 = \mathbf{ZZ}'$. Note that $\tau^2 > d$.

Let $\boldsymbol{\theta} = (\sigma^2, \tau^2, \tau^2 - d)$, $\mathbf{S} = (S_{ij})$, $S_{ij} = tr(\mathbf{QT}_i\mathbf{QT}_j)$, $\mathbf{Q} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. By Rao (1977), Chaubey (1980) and Mathew (1984), for given $\mathbf{p}$, if $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$ satisfies

$$\mathbf{S}\boldsymbol{\lambda} = \mathbf{p},$$

a MINQUE of $\mathbf{p}'\boldsymbol{\theta}$ is $\mathbf{Y}'(\sum_i \lambda_i\mathbf{QT}_i\mathbf{Q})\mathbf{Y}$.

By letting $p = c(1, 0, 0)$, $p = (0, 1, 0)$ and $p = (0, 1, -1)$, we can get estimators of σ^2, τ^2 and $d\tau^2$.

Let $\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}..., \mathbf{Y}^{(N)}$ be N vectors, independently and identically distributed as $\mathbf{Y}$ in the model (2–18). Then the model (2–18) becomes

$$\widetilde{\mathbf{Y}} = \widetilde{\mathbf{X}}\beta + \widetilde{\varepsilon}, \tag{2–19}$$

where $\widetilde{\mathbf{Y}} = [\mathbf{Y}^{(1)\mathrm{T}}, \mathbf{Y}^{(2)\mathrm{T}}..., \mathbf{Y}^{(N)\mathrm{T}}]^{\mathrm{T}}$, $\widehat{\mathbf{X}} = [\mathbf{X}', \mathbf{X}', ..., \mathbf{X}']^{\mathrm{T}}$ and $\widehat{\varepsilon} = [\varepsilon', ..., \varepsilon']^T$.

Let $\widehat{\theta}_i$ be the MINQUE of the variance components for the model (2–19). $\widehat{\theta}_1 = \widehat{\sigma}^2$ and $\widehat{\theta}_2 = \widehat{\tau}^2$. By the corollary 1 in Brown (1976), we know that $N^{\frac{1}{2}}(\widehat{\theta}_i - \theta_i)$ has limiting normal distribution. Thus, the $\widehat{\sigma}^2$ and $\widehat{\tau}^2$ mentioned above have limiting normal distributions.

However, the MINQUE can also be negative in some cases. Mathew (1984) and Pukelsheim (1981) (in their Theorem 1 and 2) discussed some conditions to make

the MINQUE nonnegative definite. It is easy to show when we estimate σ^2 or τ^2 by MINQUE, the Dirichlet process mixed model does not satisfy the condition of Theorem 1 and 2 in Pukelsheim (1981). Thus, the estimations by MINQUE might be negative in our model. When a MINQUE is negative, we can replace it with other positive estimator, such as MINQE(minimum norm quadratic estimator).

2.7.2 MINQE for σ^2 and τ^2

There is another estimator called MINQE(minimum norm quadratic estimator). The MINQE is discussed in many papers, such as Rao and Chaubey (1978), Rao (1977), Rao (1979) and Brown (1976).

Define $(\alpha_1^2, \alpha_2^2, \alpha_3^2)$ to be a prior values for $(\sigma^2, \tau^2, \tau^2 - d\tau^2)$. Let $c_i = \frac{\alpha_i^4 p_i}{n}$, $i = 1, 2, 3$; $\mathbf{V}_\alpha = \alpha_1 \mathbf{T}_1 + \alpha_2 \mathbf{T}_2 + \alpha_3 \mathbf{T}_3$; $\mathbf{P}_\alpha = \mathbf{X}(\mathbf{X}'\mathbf{V}_\alpha^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}_\alpha^{-1}$; $\mathbf{R}_\alpha = \mathbf{I} - \mathbf{P}_\alpha$.

In MINQE, we can use the following estimator to estimate $\mathbf{p}'\theta$:

$$\mathbf{Y}^\top \sum_{\mathrm{i}=1}^{3} \mathbf{c}_\mathrm{i} \mathbf{R}_\alpha^\top \mathbf{V}_\alpha^{-1} \mathbf{T}_\mathrm{i} \mathbf{V}_\alpha^{-1} \mathbf{R}_\alpha \mathbf{Y}.$$

Thus, the estimators of σ^2 and τ^2 is $\widehat{\sigma}^2 = \frac{\alpha_1^4}{\mathrm{n}} \mathbf{Y}^\top \mathbf{R}_\alpha^\top \mathbf{V}_\alpha^{-1} \mathbf{T}_1 \mathbf{V}_\alpha^{-1} \mathbf{R}_\alpha \mathbf{Y}$;

$\widehat{\tau}^2 = \frac{\alpha_2^4}{\mathrm{n}} \mathbf{Y}^\top \mathbf{R}_\alpha^\top \mathbf{V}_\alpha^{-1} \mathbf{T}_2 \mathbf{V}_\alpha^{-1} \mathbf{R}_\alpha \mathbf{Y}$. $\widehat{\mathbf{d}} = \mathbf{1} - \frac{1}{\widehat{\tau}^2} \frac{\alpha_3^4}{\mathrm{n}} \mathbf{Y}^\top \mathbf{R}_\alpha^\top \mathbf{V}_\alpha^{-1} \mathbf{T}_3 \mathbf{V}_\alpha^{-1} \mathbf{R}_\alpha \mathbf{Y}$.

$\widehat{\sigma}^2$ and $\widehat{\tau}^2$ are both positive.

2.7.3 Estimations of σ^2, τ^2 and Covariance by the Sample Covariance Matrix

In this part, we only consider a model of the form

$$Y_{ij} = \mathbf{x}_i' \boldsymbol{\beta} + \psi_i + \varepsilon_{ij}, \quad 1 \le i \le r, \ 1 \le j \le t, \tag{2–20}$$

which has the covariance matrix (as shown before)

$$\mathbf{V} = \begin{bmatrix} \sigma^2 \mathbf{I} + \tau^2 \mathbf{J} & d\tau^2 \mathbf{J} & d\tau^2 \mathbf{J} & \cdots & d\tau^2 \mathbf{J} \\ d\tau^2 \mathbf{J} & \sigma^2 \mathbf{I} + \tau^2 \mathbf{J} & d\tau^2 \mathbf{J} & \cdots & d\tau^2 \mathbf{J} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ d\tau^2 \mathbf{J} & d\tau^2 \mathbf{J} & \cdots & d\tau^2 \mathbf{J} & \sigma^2 \mathbf{I} + \tau^2 \mathbf{J} \end{bmatrix}, \tag{2–21}$$

where $\mathbf{I}$ is the $t \times t$ identity matrix, $\mathbf{J}$ is a $t \times t$ matrix of ones. We will give another method to estimate the σ^2, τ^2 and d and discuss some properties. Given a sample consisting of h independent observations $\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)} \ldots, \mathbf{Y}^{(h)}$ of n-dimensional random variables, an unbiased estimator of the covariance matrix $\mathbf{Var}(\mathbf{Y}) = \mathbf{E}(\mathbf{Y} - \mathbf{E}(\mathbf{Y}))(\mathbf{Y} - \mathbf{E}(\mathbf{Y}))^{\top}$ is the sample covariance matrix

$$\widehat{\mathbf{V}}(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)} \ldots, \mathbf{Y}^{(h)}) = \frac{\mathbf{1}}{\mathbf{h-1}} \sum_{i=1}^{h} (\mathbf{Y}^{(i)} - \overline{\mathbf{Y}})(\mathbf{Y}^{(i)} - \overline{\mathbf{Y}})^{\top} \tag{2–22}$$

In fact, with the estimated $\widehat{\mathbf{V}}$ we can get estimators of σ^2, τ^2 and d.

The variance-covariance matrix has the same structure as Eq.(2–21). We can use the estimated block matrix $\sigma^2\mathbf{I} + \tau^2\mathbf{J}$ on diagonal position to estimate σ^2 and τ^2. The average of the diagonal elements in the block matrices $\sigma^2\mathbf{I} + \tau^2\mathbf{J}$ can be considered as estimators of $\sigma^2 + \tau^2$. The average of the off-diagonal elements in the block matrices $\sigma^2\mathbf{I} + \tau^2\mathbf{J}$ can be considered as estimators of τ. The difference of these two averages can be use to estimate σ^2. In addition, we can use the average of elements on the block matrices $d\tau^2\mathbf{J}$ to estimate the d.

2.7.4 Further Discussion

As discussed in Robinson (1991), Harville (1977), Kackar and Harville (1984), Kackar and Harville (1981), and Das et al. (2004), when σ^2 and τ^2 are unknown, we can still use an estimator of $\boldsymbol{\beta}$ (or $\boldsymbol{\psi}$, or their linear combinations) with σ^2 and τ^2 replaced by their corresponding estimators in the expression of BLUE.

Kackar and Harville (1981) showed that the estimated $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\psi}}$ are still unbiased when the estimators of σ^2 and τ^2 are even and translation-invariant, i.e. when $\widehat{\sigma}^2(y) = \widehat{\sigma}^2(-y), \widehat{\tau}^2(y) = \widehat{\tau}^2(-y), \widehat{\sigma}^2(y + X\beta) = \widehat{\sigma}^2(y), \widehat{\tau}^2(y + X\beta) = \widehat{\tau}^2(y)$.

It is clear that $\widehat{\mathbf{V}}(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)} \ldots, \mathbf{Y}^{(h)})$ in Eq.(2–22) satisfies the even condition. Since, $\widehat{\mathbf{V}}(\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(h)}) = \widehat{\mathbf{V}}(\mathbf{Y}^{(1)} + \mathbf{X}\beta, \mathbf{Y}^{(2)} + \mathbf{X}\beta \ldots, \mathbf{Y}^{(h)} + \mathbf{X}\beta)$, for every β, it also satisfies the translation-invariant condition. As discussed in the previous section, the estimation of σ^2 and τ^2 by the sample covariance matrix can be written in the form

$\sum_{i,j} H_i \widehat{\mathbf{V}}(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)} ..., \mathbf{Y}^{(h)}) G_j$, where H_i, G_j are matrix free of σ^2 and τ^2. Thus, the estimators for σ^2 and τ^2, by the sample covariance matrix, also satisfy the even and translation-invariant conditions. The estimator of $\boldsymbol{\beta}$ (or $\boldsymbol{\psi}$, or their linear combinations) is still unbiased.

On the other hand, by Das et al. (2004) we know the estimators from MINQUE also satisfy the even and translation-invariant conditions. The estimators of $\boldsymbol{\beta}$ (or $\boldsymbol{\psi}$, or their linear combinations) is also unbiased.

2.7.5 Simulation Study

Example 13. In this example, consider the Dirichlet process oneway model:

$$Y_{ij} = \mu + \psi_i + \varepsilon_{ij}, \quad 1 \leq i \leq 7, \; 1 \leq j \leq 7, \tag{2–23}$$

i.e. $\mathbf{Y} = \mathbf{1}\mu + \mathbf{BA}\boldsymbol{\eta} + \boldsymbol{\varepsilon}$.

We want to compare the performance of the methods to estimate σ^2 and τ^2.

We first assume that the true $\sigma^2 = \tau^2 = 1, 10$. We simulate 1000 data sets for $\sigma^2 = \tau^2 = 1, 10$ respectively. For given σ and τ, we use the method discussed in Kyung et al. (2010) to generate the matrix $\mathbf{A}$. At $(t+1)$th step, we use the following expression to generate the matrix $\mathbf{A}$: $\mathbf{q}^{(t+1)} = (q_1^{(t+1)}, ..., q_r^{(t+1)}) \sim \mathrm{Dirichlet}(r_1^{(t)} + 1, ..., r_k^{(t)} + 1, 1, ..., 1)$; every row of $\mathbf{A}$ a_i, $a_i \sim \mathrm{Multinomial}(1, \mathbf{q}^{(t+1)})$. Thus, we can generate the matrix $\mathbf{A}$. Since for given $\mathbf{A}$, $\mathbf{Y} \sim \mathbf{N}(\mathbf{1}\mu, \sigma^2 \mathbf{I}_n + \tau^2 \mathbf{BAA}'\mathbf{B}')$, we can generate $\mathbf{Y}$. Here assume $\mu = 0$. So we can generate the corresponding data sets for different σ and τ.

We use four methods (MINQUE, MINQE, ANOVA, and sample covariance matrix) to estimate σ^2 and τ^2. When using MINQE, we use the the prior values $(1, 1)$ for (σ^2, τ^2) for all data sets. For every method, we calculate the mean of the 1000 corresponding estimators and the mean squared errors (MSE). The results are listed in Table 2-3 and 2-4.

In these tables, the estimators using sample variance always give smallest MSE for $\hat{\sigma}^2$ and $\hat{\tau}^2$, no matter whether the true σ^2 and τ^2 are big or small. The mean squared

errors for MINQUE and MINQE are almost the same, although the true σ^2 and τ^2 might be far away from the prior value $(1, 1)$. However, we also find, on average, the estimators of τ^2 and σ^2 by MINQUE, MINQE and the sample covariance matrix are almost the same as the true values. However, the estimators by MINQUE and MINQE have smaller bias, but have larger variance.

The estimators using the sample covariance matrix have small bias and smaller variance. The MSE of the estimators using the sample covariance matrix is much smaller than others. The ANOVA estimators are not satisfactory. In Table 2-5, we calculate the cutoff points and corresponding coverage probability by using the results in Table 2-3. Obviously, the methods of by the sample covariance matrix give the best results.

Table 2-3. Estimated σ^2 and τ^2 under the Dirichlet process oneway model $Y_{ij} = \mu + \psi_i + \varepsilon_{ij}, 1 \leq i \leq 7, 1 \leq j \leq 7$. $\sigma^2 = \tau^2 = 1$.

method	mean of $\hat{\sigma}^2$	mean of $\hat{\tau}^2$	MSE of $\hat{\tau}^2$	MSE of $\hat{\tau}^2$
MINQUE	1.003	1.007	0.08	1.15
MINQE	1.003	1.007	0.08	1.15
ANOVA	1.74	0.14	1.27	1.33
Sample covariance matrix	1.003	1.004	0.005	0.003

Table 2-4. Estimated σ^2 and τ^2 under the Dirichlet process oneway model $Y_{ij} = \mu + \psi_i + \varepsilon_{ij}, 1 \leq i \leq 7, 1 \leq j \leq 7$. $\sigma^2 = \tau^2 = 10$.

method	mean of $\hat{\sigma}^2$	mean of $\hat{\tau}^2$	MSE of $\hat{\tau}^2$	MSE of $\hat{\tau}^2$
MINQUE	9.97	10.10	8.00	143.43
MINQE	9.97	10.09	7.99	143.49
ANOVA	17.46	-1.20	146.63	135.66
Sample covariance matrix	10.48	9.97	0.69	0.86

Table 2-5. Estimated cutoff points for density of **Y** with the estimators of σ^2 and τ^2 in Table 2-3 under the Dirichlet model $Y_{ij} = \mu + \psi_i + \varepsilon_{ij}$, $1 \leq i \leq 7, 1 \leq j \leq 7$. $m = 3, \mu = 0, \alpha = 0.95$. True $\sigma^2 = \tau^2 = 1$.

method	estimated cutoff	P($\overline{Y}$ <estimated cutoff)
MINQUE	1.04	0.955
MINQE	1.04	0.955
ANOVA	0.57	0.822
Sample covariance matrix	1.04	0.955

CHAPTER 3
SIMULATION STUDIES AND APPLICATION

We have discussed the classical estimation under the Dirichlet model in the previous sections. In this part, we will compare the performances of the Dirichlet process mixed model with the classical normal mixed model. We use both the Dirichlet model and the classical normal model to fit some simulated data sets and a real data set, and compare the corresponding results.

3.1 Simulation Studies

First, we will do some simulation studies to investigate the performance of the Dirichlet linear mixed model. We will generate the data files from the two models: the linear mixed model with Dirichlet process random effects and the linear mixed model with normal random effects. Then we use both the Dirichlet model and the normal model to fit the simulated data sets and compare the results of the Dirichlet linear mixed model with the results of the classical normal linear mixed model.

3.1.1 Data Generation and Estimations of Parameters

We generate the data from two models.

Data Origin 1. The data are generated from the classical normal mixed model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\psi} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\psi} \sim N(\mathbf{0}, \tau^2\mathbf{I}_r)$ and $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}_n\sigma^2)$. $r = 5$ and $n = 25$.

Data Origin 2. The data are generated from the Dirichlet mixed model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\psi} + \boldsymbol{\varepsilon}$, where $\psi_j \sim \mathcal{DP}(m, N(0, \tau^2))$ and $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}_n\sigma^2)$. For given σ and τ, we use the method discussed in Kyung et al. (2010) to generate the matrix $\mathbf{A}$: $\mathbf{q}^{(t+1)} = (q_1^{(t+1)}, ..., q_r^{(t+1)}) \sim \mathrm{Dirichlet}(r_1^{(t)} + 1, ..., r_k^{(t)} + 1, 1, ..., 1)$; every row of $\mathbf{A}$ a_i, $a_i \sim \mathrm{Multinomial}(1, \mathbf{q}^{(t+1)})$. Thus, we can generate the matrix $\mathbf{A}$. Since for given $\mathbf{A}$, $\mathbf{Y} \sim \mathbf{N}(\mathbf{1}\mu, \sigma^2\mathbf{I}_n + \tau^2\mathbf{BAA}'\mathbf{B}')$, we can generate $\mathbf{Y}$.

Let the true $\boldsymbol{\beta} = [1, 0, 1, 1, 1]^T$, $m = 1$.

And $\mathbf{X} = \begin{bmatrix} \mathbf{F} \\ \mathbf{I}_5 \\ \vdots \\ \mathbf{I}_5 \end{bmatrix}$, where $\mathbf{F} = \begin{bmatrix} 1 & & & \\ & 2 & & \\ & & \ddots & \\ & & & 5 \end{bmatrix}$.

We generate 1000 data sets for $\sigma^2 = \tau^2 = 1, 5, 10$ respectively from both origins.

We use both the classical normal mixed model and the Dirichlet mixed model to fit the generated data sets. We calculate the corresponding estimators for $\boldsymbol{\beta}$, corresponding MSE and the estimations for τ^2 and σ^2.

Based on the results we in previous sections, we know that the BLUE of $\boldsymbol{\beta}$ and the BLUP $\boldsymbol{\psi}$ are

$$\widetilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y},\ \widetilde{\boldsymbol{\psi}} = \mathbf{C}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}), \tag{3–1}$$

where $\mathbf{V}$ is the covariance matrix. The variance of $\widetilde{\boldsymbol{\beta}}$ is

$$\mathrm{Var}(\widetilde{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}.$$

In addition, we use the sample covariance matrix to estimate σ^2, τ^2 and d.

Since the covariance matrix of $\mathbf{Y}$ in classical normal mixed model is a special case (corresponding to $d = 0$ as shown before) of the covariance matrix in Dirichlet mixed model, we can get the same estimators of τ^2 and σ^2 by using the sample covariance matrix method for both models due to the special structure of the covariance matrix.

On the other hand, we can test some hypotheses, based on the estimation of $\boldsymbol{\beta}$ and its corresponding variance matrix. For example we can test: $\mathbf{L}\boldsymbol{\beta} = 0$, where $\mathbf{L}$ is a a single row.

Let

$$T = \frac{\mathbf{L}\widetilde{\boldsymbol{\beta}}}{\sqrt{\mathbf{L}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{L}'}}. \tag{3–2}$$

Although we do not know the exact distribution of T now, we can use the generated 1000 data sets to find the estimated distribution of T and find the corresponding estimated p value for the observation $\mathbf{y}$.

3.1.2 Simulation Results

Let $\hat{\boldsymbol{\beta}}_D$ be the estimated BLUE by using the Dirichlet model and $\hat{\boldsymbol{\beta}}_N$ be the estimated BLUE by using the normal mixed model.

Using the generated data sets, we can estimate the covariance matrix by the sample covariance matrix. Let $\hat{\mathbf{V}}_D$ and $\hat{\mathbf{V}}_N$, where $\hat{\mathbf{V}}_D$ is the estimated covariance matrix by using the Dirichlet mixed model and $\hat{\mathbf{V}}_N$ is the estimated covariance matrix by using the classical normal mixed model. We can also get the estimated CDFs of T in both the Dirichlet mixed model and the classical normal mixed model.

The estimation results are listed in Table 3-1 and 3-3. d is the true covariance and $\hat{d}$ is estimation of d in the Dirichlet mixed model.

From the Table 3-1and 3-3, we find that the MSE($\hat{\boldsymbol{\beta}}_D$) are almost same for the same τ and σ, no matter which distribution the true data are from. However, the changes of MSE($\hat{\boldsymbol{\beta}}_N$) is bigger when the true model of the data changes from the normal model to the Dirichlet model. The estimated $\hat{d}$ is close to the true d by using the Dirichlet mixed model.

Table 3-1. The simulations results with $\sigma^2 = \tau^2 = 1$. $m = 1$.

Data Origin	$\hat{\sigma}^2$	$\hat{\tau}^2$	MSE($\hat{\boldsymbol{\beta}}_D$)	MSE($\hat{\boldsymbol{\beta}}_N$)	d	$\hat{d}$
Normal Model	1.00	1.02	1.004	0.832	0	0.019
Dirichlet Model	1.00	0.93	1.001	1.107	0.333	0.326

Table 3-2. The simulations results with $\sigma^2 = \tau^2 = 5$. $m = 1$.

Data Origin	$\hat{\sigma}^2$	$\hat{\tau}^2$	MSE($\hat{\boldsymbol{\beta}}_D$)	MSE($\hat{\boldsymbol{\beta}}_N$)	d	$\hat{d}$
Normal Model	5.03	5.03	5.219	4.358	0	0.067
Dirichlet Model	5.04	4.95	4.862	5.421	1.667	1.474

Table 3-3. The simulations results with $\sigma^2 = \tau^2 = 10. m = 1$.

Data Origin	$\hat{\sigma}^2$	$\hat{\tau}^2$	MSE($\hat{\beta}_D$)	MSE($\hat{\beta}_N$)	d	$\hat{d}$
Normal Model	10.04	9.75	9.946	7.931	0	0.107
Dirichlet Model	9.83	10.88	9.844	11.322	3.333	3.904

3.2 Application to a Real Data Set

We will use the Dirichlet mixed model to analyze one real data set. The data is from Professor Volker Mai's lab. They want to investigate the contribution of gut microbiota to colorectal cancer. A unmatched pairs case-control study was performed in human volunteers undergoing a screening colonoscopy. Fecal samples and multiple colon biopsy samples from the individuals was collected. The group did 16S rRNA sequence analysis to obtain the data. Our goal is to test if there is difference between case and control for each bacteria (one-sided and two-sided tests).

The data file contains the counts for bacteria. For each bacteria, there are count numbers for 30 cases and 30 controls. There are 6321 rows in the data. In the data set, the rows correspond to the bacteria and the columns correspond to the count numbers for cases and controls. We want to test if there is difference between the cases and controls for each bacteria.

3.2.1 The Model and Estimation

For each bacteria, we use the following model to fit the data:

$$\widetilde{Y}_{ij} = \mu + \alpha_i + \gamma_{ij} + \varepsilon_{ij},\ i = 1, 2,\ j = 1, ..., 30, \tag{3–3}$$

where $\widetilde{Y}_{ij}$ is either $arcsin(\sqrt{p_{ij}})$ or $\log(Y_{ij} + \delta)$, p_{ij} is the column-wise proportion and δ is small real number. In other words, we consider two kinds of transformation of the response for the model in this part. $i = 1$ means "case" and $i = 2$ means "control". $j = 1, ..., 30$ means the index of subjects (volunteers).

We cannot use the sample covariance matrix to estimate σ, τ, and d in this real data set, since we do not have a sample. We only have one observation for each

bacteria. However, we can use the minimum norm quadratic unbiased estimation (MINQUE) method to estimate σ, τ, and d for this observation.

Assume T_0 is corresponding to the T statistic for this observation.

We do not know the exact distribution of T, but we can estimate its distribution non-parametrically by using the permutation method. We randomly permute the treatment labels and recalculate the $\hat{\sigma}$, $\hat{\tau}$, $\hat{d}$ and T statistic.

Repeating this procedure for B times, we can obtain a set of T statistics $T_1, ..., T_B$. Then we can estimate the p value.

For example, if we want to test $H_0 : \alpha_1 = \alpha_2;\ H_A : \alpha_1 > \alpha_2$, then the p value for this bacteria is:

$$p = \frac{\#\{T_i \geq T_0, i = 1, ..., B\}}{B}.$$

3.2.2 Simulation Results for the Models in the Section 3.2.1

In order to compare the performance of the models, we will first do some simulation studies, since we do not know the true results for the models in the Section 3.2.1. We generate data sets from the negative binomial distributions with several parameters setups. In this section, we just show the numerical results for different data setups by using different methods. The discussion and conclusion will be shown in next section.

Table 3-4 show all the details of the parameters under all these data setups. We generate the data from negative binomial(μ, r), i.e., mean=μ, variance=$\mu + \frac{\mu^2}{r}$. In every data setup, we generate 10 datasets. Each dataset has 800 rows for each data setup. μ^{100}_{case} means that we generate the data with $\mu = \mu^{100}_{case}$ for the cases in the first 100 rows. $\mu^{100}_{control}$ means that we generate the data with $\mu = \mu^{100}_{control}$ for the controls in the first 100 rows. μ^{other} means that we generate the data with $\mu = \mu^{other}$ for the cases and controls in the rest rows.

Data Setup 1. Generate the data from negative binomial(μ, r), i.e., mean=μ, variance=$\mu + \frac{\mu^2}{r}$.

Generate 10 data sets. Each data set has 800 rows. Take $r = 0.1$.

Table 3-4. The Data Setups.

	r	μ^{100}_{case}	$\mu^{100}_{control}$	μ^{other}
Data Setup 1	0.1	2	0.3	0.3
Data Setup 2	0.02	1	0.3	0.3
Data Setup 3	0.005	2	0.3	0.3
Data Setup 4	0.02	5	0.3	0.3
Data Setup 5	0.1	10	0.3	0.3
Data Setup 6	0.02	10	0.3	0.3

Let $\mu = 2$, for the cases in the first 100 rows; $\mu = 0.3$, for the controls in the first 100 rows;

Let $\mu = 0.3$, for the cases and controls in the rest rows.

By doing this, we generate a data set with about 86% zeros. Max value is 103 and min value is 0. The original real data has about 94% zeros. The maximum value is 1573 and the minimum value is 0.

Use the normal model, the Dirichlet model and the LRT to find the p values.

First use the usual transformation: $arcsin(\sqrt{\frac{Y_{ij}}{n_{ij}}})$.

On average, if the significance level is 0.1, the Dirichlet model gives 99 (58) bacteria significant; the normal model gives 104 (58) bacteria significant. 310 bacteria (97) are significant by LRT.

One “estimate” of FDR is 41/99=0.41 in the Dirichlet model; the “estimate” of FDR is 46/104=0.44 in the normal model; the “estimate” of FDR is 41/310=0.69 when using LRT.

The number in ( ) is the number of significant bacteria from the first 100 bacteria.

On average, if the significance level is 0.05; the Dirichlet model gives 62 (42) bacteria significant; the normal model gives 63 (41) bacteria significant. 161 (91) bacteria are significant by LRT.

One “estimate” of FDR is 20/62=0.32 in the Dirichlet model; the “estimate” of FDR is 22/63=0.35 in the normal model; the “estimate” of FDR is 70/161=0.43 when using LRT.

Then use the usual transformation: $\log(Y_{ij})$.

On average, if the significance level is 0.1, the Dirichlet model gives 128 (61) bacteria significant; the normal model gives 104 (58) bacteria significant.

One "estimate" of FDR is 67/128=0.52 in the Dirichlet model; the "estimate" of FDR is 46/104=0.44 in the normal model.

On average, if the significance level 0.05; the Dirichlet model gives 87 (26) bacteria significant; the normal model gives 95 (26) bacteria significant.

One "estimate" of FDR is 61/87=0.70 in the Dirichlet model; the "estimate" of FDR is 59/95=0.62 in the normal model.

Data Setup 2. Generate the data from negative binomial(μ, r), i.e., mean=μ, variance=$\mu + \frac{\mu^2}{r}$.

Generate 10 data sets. Each data set has 800 rows. Take $r = 0.02$.

Let $\mu = 1$, for the cases in the first 100 rows; $\mu = 0.3$, for the controls in the first 100 rows;

Let $\mu = 0.3$, for the cases and controls in the rest rows.

Use the normal model, the Dirichlet model and the LRT to find the p values.

First use the usual transformation: $arcsin(\sqrt{\frac{Y_{ij}}{n_{ij}}})$.

On average, if the significance level is 0.1, the Dirichlet model gives 78 (24) bacteria significant; the normal model gives 96 (28) bacteria significant. 145 bacteria (48) are significant by LRT.

One "estimate" of FDR is 54/78=0.69 in the Dirichlet model; the "estimate" of FDR is 68/96=0.70 in the normal model; the "estimate" of FDR is 97/145=0.67 when using LRT.

On average, if the significance level 0.05; the Dirichlet model gives 41 (13) bacteria significant; the normal model gives 32 (12) bacteria significant. 61 (29) bacteria are significant by LRT.

One "estimate" of FDR is 28/41=0.68 in the Dirichlet model; the "estimate" of FDR is 20/32=0.63 in the normal model; the "estimate" of FDR is 32/61=0.53 when using LRT.

Then use the usual transformation: $\log(Y_{ij})$.

On average, if the significance level is 0.1, the Dirichlet model gives 87 (26) bacteria significant; the normal model gives 95 (26) bacteria significant.

One “estimate” of FDR is 61/87=0.70 in the Dirichlet model; the “estimate” of FDR is 69/95=0.72 in the normal model.

On average, if the significance level 0.05; the Dirichlet model gives 44 (13) bacteria significant; the normal model gives 31 (11) bacteria significant.

One “estimate” of FDR is 28/44=0.64 in the Dirichlet model; the “estimate” of FDR is 20/31=0.64 in the normal model.

Data Setup 3. Generate the data from negative binomial(μ, r), i.e., mean=μ, variance=$\mu + \frac{\mu^2}{r}$.

Generate 10 data sets. Each data set has 800 rows. Take $r = 0.005$. With this small r, we increase the variance of the negative binomial distribution, which help us to generate some ‘extreme’ numbers.

Let $\mu = 2$, for the cases in the first 100 rows; $\mu = 0.3$, for the controls in the first 100 rows;

Let $\mu = 0.3$, for the cases and controls in the rest rows.

Use the normal model, the Dirichlet model and the LRT to find the p values.

First use the usual transformation: $arcsin(\sqrt{\frac{Y_{ij}}{n_{ij}}})$.

On average, if the significance level is 0.1, the Dirichlet model gives 52 (8) bacteria significant; the normal model gives 44 (7) bacteria significant. 12 bacteria (3) are significant by LRT.

One “estimate” of FDR is 44/52=0.84 in the Dirichlet model; the “estimate” of FDR is 37/44=0.84 in the normal model; the “estimate” of FDR is 9/12=0.75 when using LRT.

On average, if the significance level 0.05; the Dirichlet model gives 23 (3) bacteria significant; the normal model gives 3 (0) bacteria significant. 2 (1) bacteria are significant by LRT.

One “estimate” of FDR is 20/23=0.87 in the Dirichlet model; the “estimate” of FDR is 3/3=1.0 in the normal model; the “estimate” of FDR is 1/2=0.50 when using LRT.

Then use the usual transformation: $\log(Y_{ij})$.

On average, if the significance level is 0.1, the Dirichlet model gives 47 (8) bacteria significant; the normal model gives 44 (8) bacteria significant.

One “estimate” of FDR is 39/47=0.83 in the Dirichlet model; the “estimate” of FDR is 36/44=0.82 in the normal model.

On average, if the significance level 0.05; the Dirichlet model gives 23 (4) bacteria significant; the normal model gives 4 (0) bacteria significant.

One “estimate” of FDR is 19/23=0.83 in the Dirichlet model; the “estimate” of FDR is 1 in the normal model.

Data Setup 4. Generate the data from negative binomial(μ, r), i.e., mean=μ, variance=$\mu + \frac{\mu^2}{r}$.

Generate 10 data sets. Each data set has 800 rows. Take $r = 0.02$.

Let $\mu = 5$, for the cases in the first 100 rows; $\mu = 0.3$, for the controls in the first 100 rows;

Let $\mu = 0.3$, for the cases and controls in the rest rows.

Use the normal model, the Dirichlet model and the LRT to find the p values.

First use the usual transformation: $arcsin(\sqrt{\frac{Y_{ij}}{n_{ij}}})$.

On average, if the significance level is 0.1, the Dirichlet model gives 80 (27) bacteria significant; the normal model gives 93 (30) bacteria significant. 163 bacteria (61) are significant by LRT.

One “estimate” of FDR is 53/80=0.66 in the Dirichlet model; the “estimate” of FDR is 63/93=0.67 in the normal model; the “estimate” of FDR is 102/163=0.63 when using LRT.

On average, if the significance level 0.05; the Dirichlet model gives 44 (15) bacteria significant; the normal model gives 35 (13) bacteria significant. 87 (41) bacteria are significant by LRT.

One “estimate” of FDR is 29/44=0.66 in the Dirichlet model; the “estimate” of FDR is 22/35=0.63 in the normal model; the “estimate” of FDR is 46/87=0.53 when using LRT.

Then use the usual transformation: $\log(Y_{ij})$.

On average, if the significance level is 0.1, the Dirichlet model gives 98 (32) bacteria significant; the normal model gives 90 (30) bacteria significant.

One “estimate” of FDR is 66/98=0.67 in the Dirichlet model; the “estimate” of FDR is 60/90=0.67 in the normal model. On average, if the significance level 0.05; the Dirichlet model gives 50 (19) bacteria significant; the normal model gives 35 (13) bacteria significant.

One “estimate” of FDR is 31/50=0.62 in the Dirichlet model; the “estimate” of FDR is 22/35=0.63 in the normal model.

Data Setup 5. Generate the data from negative binomial(μ, r), i.e., mean=μ, variance=$\mu + \frac{\mu^2}{r}$.

Generate 10 data sets. Each data set has 800 rows. Take $r = 0.1$.

Let $\mu = 10$, for the cases in the first 100 rows; $\mu = 0.3$, for the controls in the first 100 rows;

Let $\mu = 0.3$, for the cases and controls in the rest rows.

Use the normal model, the Dirichlet model and the LRT to find the p values.

First use the usual transformation: $arcsin(\sqrt{\frac{Y_{ij}}{n_{ij}}})$.

On average, if the significance level is 0.1, the Dirichlet model gives 94 (79) bacteria significant; the normal model gives 97 (80) bacteria significant.

One “estimate” of FDR is 14/94=0.15 in the Dirichlet model; the “estimate” of FDR is 17/97=0.18 in the normal model.

On average, if the significance level 0.05; the Dirichlet model gives 74 (65) bacteria significant; the normal model gives 73 (64) bacteria significant. 165 (100) bacteria are significant by LRT.

One “estimate” of FDR is 9/74=0.12 in the Dirichlet model; the “estimate” of FDR is 9/73=0.12 in the normal model; the “estimate” of FDR is 65/165=0.39 when using LRT.

Then use the usual transformation: $\log(Y_{ij})$.

On average, if the significance level is 0.1, the Dirichlet model gives 146 (82) bacteria significant; the normal model gives 97 (75) bacteria significant.

One “estimate” of FDR is 64/146=0.44 in the Dirichlet model; the “estimate” of FDR is 22/97=0.23 in the normal model.

On average, if the significance level 0.05; the Dirichlet model gives 102 (71) bacteria significant; the normal model gives 70 (59) bacteria significant.

One “estimate” of FDR is 31/102=0.30 in the Dirichlet model; the “estimate” of FDR is 11/70=0.16 in the normal model.

Data Setup 6. Generate the data from negative binomial(μ, r), i.e., mean=μ, variance=$\mu + \frac{\mu^2}{r}$.

Generate 10 data sets. Each data set has 800 rows. Take $r = 0.02$.

Let $\mu = 10$, for the cases in the first 100 rows; $\mu = 0.3$, for the controls in the first 100 rows;

Let $\mu = 0.3$, for the cases and controls in the rest rows.

Use the normal model, the Dirichlet model and the LRT to find the p values.

First use the usual transformation: $arcsin(\sqrt{\frac{Y_{ij}}{n_{ij}}})$.

On average, if the significance level is 0.1, the Dirichlet model gives 77 (28) bacteria significant; the normal model gives 96 (31) bacteria significant; the LRT gives 117 (52) bacteria significant.

One “estimate” of FDR is 49/77=0.64 in the Dirichlet model; the “estimate” of FDR is 65/96=0.68 in the normal model; the “estimate” of FDR is 65/117=0.56 when using LRT.

On average, if the significance level 0.05; the Dirichlet model gives 44 (16) bacteria significant; the normal model gives 31 (13) bacteria significant. 90 (52) bacteria are significant by LRT.

One “estimate” of FDR is 28/44=0.64 in the Dirichlet model; the “estimate” of FDR is 18/31=0.58 in the normal model; the “estimate” of FDR is 38/90=0.42 when using LRT.

Then use the usual transformation: $\log(Y_{ij})$.

On average, if the significance level is 0.1, the Dirichlet model gives 104 (37) bacteria significant; the normal model gives 96 (32) bacteria significant.

One “estimate” of FDR is 67/104=0.64 in the Dirichlet model; the “estimate” of FDR is 64/96=0.67 in the normal model.

On average, if the significance level 0.05; the Dirichlet model gives 55 (23) bacteria significant; the normal model gives 32 (14) bacteria significant.

One “estimate” of FDR is 32/55=0.58 in the Dirichlet model; the “estimate” of FDR is 18/32=0.56 in the normal model.

3.2.3 Discussion of the Simulation Studies Results

From the simulation studies above, we find that there is no big difference between the performances under the classical linear mixed model and the performance under the Dirichlet process mixed model. The performance of the transformation $arcsin(\sqrt{\frac{Y_{ij}}{n_{ij}}})$ is very slightly better than that of the transformation $\log(Y_{ij})$.

However, the generated data sets are not similar as the original one, i.e., the generated data sets by the negative binomial distributions are not as extreme as the original real data. The original real data have about 94% zeros, with the maximum value 1573 and the minimum value 0. We can generate the corresponding zeros in the simulated data sets, but the maximum values in the simulated data sets are much smaller.

3.2.4 Results using the Real Data Set

Now we move to use the model (3–3) to analyze the original data. First, we use $arcsin(\sqrt{\frac{Y_{ij}}{n_{ij}}})$ as the response in the model (3–3).

The Dirichlet mixed model with the transformation $arcsin(\sqrt{\frac{Y_{ij}}{n_{ij}}})$ tends to give less bacteria significant. If the significance level $\alpha = 0.1$, the Dirichlet model gives 474 bacteria significant; the normal model gives 1005 bacteria significant.If the significance level $\alpha = 0.05$, the Dirichlet model gives 205 bacteria significant; the normal model gives 488 bacteria significant.

CHAPTER 4
BAYESIAN ESTIMATION UNDER THE DIRICHLET MIXED MODEL

In the previous sections, we discussed classical estimation under the Dirichlet process mixed model. We also provided general conditions under which the OLS estimator is the BLUE. We investigated the ways to estimate the variance components σ^2 and τ^2. All the analysis for the Dirichlet model are from the frequentist viewpoint.

Another way to discuss the Dirichlet process model is from the Bayesian way. Since the 1990s, the Bayesian approach has seen the most use in the Dirichlet mixed models. For example, Antoniak (1974) showed that the posterior distribution could be described as being a mixture of Dirichlet processes. Escobar and West (1995) illustrated Bayesian density estimation by using the mixture of Dirichlet process. Kyung et al. (2009) investigated a variance reduction property of the Dirichlet process prior, and Kyung et al. (2010) provided a new Gibbs sampler for the linear Dirichlet mixed model and discussed estimation of the precision parameter of the Dirichlet process. We move to Bayesian estimation in this section.

In fact, we always put priors on $\boldsymbol{\beta}$ when using Bayesian methods. Different priors and different random effects might lead to different estimators, different MSE and different Bayes risks. We can assume that the random effects follow a normal distribution. We can also assume that the random effects follow the Dirichlet process. We can put a normal distribution prior on $\boldsymbol{\beta}$. We can also put the flat prior on $\boldsymbol{\beta}$. We will investigate which prior/model is better.

In this part, we will first give the fours models which we want to discuss and the corresponding Bayesian estimators. We will show the corresponding MSE and Bayes risks of these Bayesian estimators and discuss which model is better. Furthermore, we will investigate in more detail the Dirichlet process oneway model and compare the posterior variance, variance and MSE.

4.1 Bayesian Estimators under Four Models

The BLUE and OLS estimators have been discussed in detail in previous sections. In this part, we move to Bayesian estimation. Firstly, we will give four models with different priors on β and different random effects. Then, we will give the corresponding Bayesian estimators for β under these four models.

4.1.1 Four Models and Corresponding Bayesian Estimators

We consider four models with different random effects and different priors on β :

Model 1 (Dirichlet+Normal): the mixed model with the Dirichlet process random effects and the normal distribution prior on β (as in Kyung et al. (2009)):

$$\mathbf{Y}|\beta, \mathbf{A}, \sigma^2 \sim \mathbf{N}(\mathbf{X}\beta, \sigma^2(\mathbf{I} + c\mathbf{ZAA}'\mathbf{Z}')); \beta \sim \mathbf{N}(\mathbf{0}, \nu\sigma^2\mathbf{I}_n); \sigma^2 \sim \mathcal{IG}(a, b). \tag{4–1}$$

Model 2 (Normal+Normal): the mixed model with the normal random effects and the normal distribution prior on β :

$$\mathbf{Y}|\beta, \sigma^2 \sim \mathbf{N}(\mathbf{X}\beta, \sigma^2(\mathbf{I} + c\mathbf{ZZ}')); \beta \sim \mathbf{N}(\mathbf{0}, \nu\sigma^2\mathbf{I}_n); \sigma^2 \sim \mathcal{IG}(a, b). \tag{4–2}$$

Model 3 (Dirichlet+flat prior): the mixed model with the Dirichlet process random effects and the flat prior on β :

$$\mathbf{Y}|\beta, \mathbf{A}, \sigma^2 \sim \mathbf{N}(\mathbf{X}\beta, \sigma^2(\mathbf{I} + c\mathbf{ZAA}'\mathbf{Z}')); \beta \sim \textit{flat prior}; \sigma^2 \sim \mathcal{IG}(a, b). \tag{4–3}$$

Model 4 (Normal+flat prior): the mixed model with the normal random effects and the flat prior on β :

$$\mathbf{Y}|\beta, \sigma^2 \sim \mathbf{N}(\mathbf{X}\beta, \sigma^2(\mathbf{I} + c\mathbf{ZZ}')); \beta \sim \textit{flat prior}; \sigma^2 \sim \mathcal{IG}(a, b). \tag{4–4}$$

With these four models, we can get the corresponding the Bayesian estimators. First consider Model 1.

Let $\Sigma_A = \mathbf{I} + \mathbf{cZAA}'\mathbf{Z}'$. Then we can obtain conditional density

$$\pi(\beta, \sigma^2|\mathbf{A}, \mathbf{Y}) \propto (\frac{1}{\sigma^2})^{\frac{n+1}{2}+a+1} \exp\{-\frac{b}{\sigma^2} - \frac{1}{2\nu\sigma^2}\beta^T\beta - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\beta)^T\Sigma_A^{-1}(\mathbf{Y} - \mathbf{X}\beta)\}. \tag{4–5}$$

By integrating out the σ^2 in (4–5), we have

$$\pi(\beta|\mathbf{A}, \mathbf{Y}) = C(\mathbf{A}, \mathbf{Y}, n)\frac{\Gamma(\frac{n+1}{2} + a)}{(b + \frac{1}{2\nu}\beta^T\beta + \frac{1}{2}(\mathbf{Y} - \mathbf{X}\beta)^T\Sigma_A^{-1}(\mathbf{Y} - \mathbf{X}\beta))^{\frac{n+1}{2}+a}}, \tag{4–6}$$

where $C(\mathbf{A}, \mathbf{Y}, n) = \frac{b^a}{2\pi\Gamma(a)}\nu^{-n/2}|\Sigma|^{-1/2}$.

By summing up all the possible $\mathbf{A}$ matrices, we can have the density for $\beta|\mathbf{Y}$:

$$\pi(\beta|\mathbf{Y}) = \sum P(\mathbf{A})C(\mathbf{A}, \mathbf{Y}, n)\frac{\Gamma(\frac{n+1}{2}+a)}{(b+\frac{1}{2\nu}\beta^T\beta+\frac{1}{2}(\mathbf{Y}-\mathbf{X}\beta)^T\Sigma_A^{-1}(\mathbf{Y}-\mathbf{X}\beta))^{\frac{n+1}{2}+a}}. \tag{4–7}$$

Let $\mathbf{H}_\mathrm{A} = \frac{\mathrm{X}^\top\Sigma_\mathrm{A}^{-1}\mathrm{X}}{2} + \frac{\mathrm{I_p}}{2\nu}$, $\mathbf{S}_\mathrm{A} = \frac{1}{2}\mathbf{Y}^\top\mathbf{\Sigma}_\mathrm{A}^{-1}\mathbf{X}$ and $\mathbf{D}_\mathrm{A} = \mathbf{H}_\mathrm{A}^{-1}\mathbf{S}_\mathrm{A}'$. Then Eq.(4–7) can be rewritten as

$$\pi(\beta|\mathbf{Y}) = \sum P(\mathbf{A})\frac{C(\mathbf{A}, \mathbf{Y}, n)\Gamma(\frac{n+1}{2}+a)}{(b+\frac{1}{2}\mathbf{Y}^T\Sigma_A^{-1}\mathbf{Y}-\mathbf{D}_\mathrm{A}^\top\mathbf{H}_\mathrm{A}\mathbf{D}_\mathrm{A}+(\beta-\mathbf{D}_\mathrm{A})^\top\mathbf{H}_\mathrm{A}(\beta-\mathbf{D}_\mathrm{A}))^{\frac{n+1}{2}+a}}. \tag{4–8}$$

With the expression of (4–8), we can obtain the corresponding posterior mean and posterior variance.

The Bayesian estimator for β, $\hat{\beta}^B$, is the mean of $\pi(\beta|\mathbf{Y})$ is (4–8), i.e.,

$$\hat{\beta}^B = \mathrm{E}(\beta|\mathbf{Y}) = \sum P(\mathbf{A})\mathbf{D}_\mathrm{A} = \sum \mathbf{P}(\mathbf{A})(\mathbf{X}^\top\mathbf{\Sigma}_\mathrm{A}^{-1}\mathbf{X}+\frac{\mathbf{I}_\mathrm{p}}{\nu})^{-1}\mathbf{X}^\top\mathbf{\Sigma}_\mathrm{A}^{-1}\mathbf{Y}. \tag{4–9}$$

It is clear that the Bayesian estimator under Model 1 is biased.

By similar calculation, we can obtain the Bayesian estimator for β under Model 2, $\hat{\beta}^I$,

$$\hat{\beta}^I = (\mathbf{X}^\top\mathbf{\Sigma}_\mathrm{I}^{-1}\mathbf{X}+\frac{\mathbf{I}_\mathrm{p}}{\nu})^{-1}\mathbf{X}^\top\mathbf{\Sigma}_\mathrm{I}^{-1}\mathbf{Y}, \tag{4–10}$$

where $\Sigma_I = \mathbf{I} + \mathbf{cBB}'$. Again, it is biased.

For Model 3, the Bayesian estimator for β is

$$\hat{\beta}^{BF} = \sum P(\mathbf{A})(\mathbf{X}^\top\mathbf{\Sigma}_\mathrm{A}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{\Sigma}_\mathrm{A}^{-1}\mathbf{Y}. \tag{4–11}$$

For Model 4, the Bayesian estimator for β is

$$\hat{\beta}^{IF} = (\mathbf{X}^\top\mathbf{\Sigma}_\mathrm{I}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{\Sigma}_\mathrm{I}^{-1}\mathbf{Y}. \tag{4–12}$$

It is clear that under Models 3 and 4 the estimators $\hat{\beta}^{BF}$ and $\hat{\beta}^{IF}$ are unbiased. Under Model 1 and 2 the estimators $\hat{\beta}^B$ and $\hat{\beta}^I$ are biased. With these estimators, we

can calculate the corresponding MSE and Bayes risks and compare the models in following sections.

4.1.2 More General Cases

In Model 1 and Model 2, the prior on $\boldsymbol{\beta}$ is $\mathbf{N}(\mathbf{0}, \nu\sigma^2\mathbf{I}_n)$, which has mean $\mathbf{0}$. Now we consider a general case: the mean is not a zero vector. Assume $\boldsymbol{\beta}_0$ is a known vector. Modify the prior on $\boldsymbol{\beta}$ in Model 1 and Model 2 to $\mathbf{N}(\boldsymbol{\beta}_0, \nu\sigma^2\mathbf{I}_n)$. Thus, Model 1 and Model 2 change to:

Model 1*: the mixed model with the Dirichlet process random effects and the normal distribution prior on $\boldsymbol{\beta}$:

$$\mathbf{Y}|\boldsymbol{\beta}, \mathbf{A}, \sigma^2 \sim \mathbf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2(\mathbf{I} + c\mathbf{Z}\mathbf{A}\mathbf{A}'\mathbf{Z}')); \boldsymbol{\beta} \sim \mathbf{N}(\boldsymbol{\beta}_0, \nu\sigma^2\mathbf{I}_n); \sigma^2 \sim \mathcal{IG}(a, b). \tag{4–13}$$

Model 2*: the mixed model with the normal random effects and the normal distribution prior on $\boldsymbol{\beta}$:

$$\mathbf{Y}|\boldsymbol{\beta}, \sigma^2 \sim \mathbf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2(\mathbf{I} + c\mathbf{Z}\mathbf{Z}')); \boldsymbol{\beta} \sim \mathbf{N}(\boldsymbol{\beta}_0, \nu\sigma^2\mathbf{I}_n); \sigma^2 \sim \mathcal{IG}(a, b). \tag{4–14}$$

By a similar calculation, the four Bayesian estimators for Model 1*-2* are as follows.

$$\hat{\boldsymbol{\beta}}^{*B} = \sum P(\mathbf{A})(\mathbf{X}^\top\boldsymbol{\Sigma}_A^{-1}\mathbf{X} + \frac{\mathbf{I}_p}{\nu})^{-1}(\mathbf{X}^\top\boldsymbol{\Sigma}_A^{-1}\mathbf{Y} + \frac{\boldsymbol{\beta}_0}{\nu}). \tag{4–15}$$

$$\hat{\boldsymbol{\beta}}^{*I} = (\mathbf{X}^\top\boldsymbol{\Sigma}_I^{-1}\mathbf{X} + \frac{\mathbf{I}_p}{\nu})^{-1}(\mathbf{X}^\top\boldsymbol{\Sigma}_I^{-1}\mathbf{Y} + \frac{\boldsymbol{\beta}_0}{\nu}). \tag{4–16}$$

4.2 The Oneway Model

In the previous section, we have obtained expressions of the Bayesian estimators under four models. In this part, we want to investigate more details about the Bayesian version of the oneway model. We will consider the four models (Model 1-Model 4) discussed in previous section. First, we will rewrite the Bayesian estimators under the four models in a more concise way and show the relationship between the Bayesian estimators and the BLUE. And we will compare their variance, biases and means squared errors. In the last part of this section we will discuss the choice of the parameter ν.

4.2.1 Estimators

As shown in (4–9), under Model 1, the Bayesian estimator for μ is

$$\hat{\mu}^B = \sum P(\mathbf{A}) \frac{\mathbf{1}^\top \mathbf{\Sigma}_{\mathsf{A}}^{-1}}{\mathbf{1}^\top \mathbf{\Sigma}_{\mathsf{A}}^{-1} \mathbf{1} + \frac{1}{\nu}} \mathbf{Y}. \tag{4–17}$$

Here, $\sum P(\mathbf{A}) \frac{1^\top \Sigma_{\mathsf{A}}^{-1}}{1^\top \Sigma_{\mathsf{A}}^{-1} 1 + \frac{1}{\nu}}$ is a vector.

Denote $\mathbf{h}^T = (h_1, ..., h_n) = \sum P(\mathbf{A}) \frac{1^\top \Sigma_{\mathsf{A}}^{-1}}{1^\top \Sigma_{\mathsf{A}}^{-1} 1 + \frac{1}{\nu}}$. So, $\hat{\mu}^B = \mathbf{h}^\top \mathbf{Y}$.

Similarly, under Model 2, the Bayesian estimator for μ is

$$\hat{\mu}^I = \frac{\mathbf{1}^\top \mathbf{\Sigma}_{\mathsf{I}}^{-1}}{\mathbf{1}^\top \mathbf{\Sigma}_{\mathsf{I}}^{-1} \mathbf{1} + \frac{1}{\nu}} \mathbf{Y}, \tag{4–18}$$

where, $\frac{1^\top \Sigma_{\mathsf{I}}^{-1}}{1^\top \Sigma_{\mathsf{I}}^{-1} 1 + \frac{1}{\nu}}$ is a vector.

Denote $\mathbf{e}^T = (e_1, ..., e_n) = \frac{1^\top \Sigma_{\mathsf{I}}^{-1}}{1^\top \Sigma_{\mathsf{I}}^{-1} 1 + \frac{1}{\nu}}$. So, $\hat{\mu}^I = \mathbf{e}^\top \mathbf{Y}$. Let $l_I = \sum e_i$.

The following theorem shows the relationship between the Bayesian estimators $\hat{\mu}^B, \hat{\mu}^I$ and the BLUE.

Theorem 14. *Let $l = \sum h_i$ and $l_I = \sum e_i$, where $0 < l < 1$ and $0 < l_I < 1$, l and l_I are free of the matrix $\mathbf{A}$ and depend only on ν, σ^2, τ^2 and m.*

Then $\mathbf{h}^T = (\frac{l}{n}, ..., \frac{l}{n})$ and $\mathbf{e}^T = (\frac{l_I}{n}, ..., \frac{l_I}{n})$. In addition, $l_I \geqslant l$, and,

$$\hat{\mu}^B = \mathbf{h}^\top \mathbf{Y} = l\overline{\mathbf{Y}}, \ \hat{\mu}^I = \mathbf{e}^\top \mathbf{Y} = l_I \overline{\mathbf{Y}}$$

where $\overline{\mathbf{Y}}$ is the BLUE for μ under the Dirichlet process oneway model.

Moreover,

$$\hat{\mu}^{BF} = \hat{\mu}^{IF} = \overline{\mathbf{Y}},$$

where $\hat{\mu}^{BF}$ and $\hat{\mu}^{IF}$ are Bayesian estimators under Model 3 and Model 4 respectively. $\overline{\mathbf{Y}}$ is the BLUE as shown before.

Proof. See Appendix D. □

Because $\hat{\mu}^{BF} = \hat{\mu}^{IF} = \overline{\mathbf{Y}}$, we have the same estimators for μ under Model 3 and Model 4 (i.e. the Bayesian estimators are same under the Dirichlet process oneway model with flat prior and the classical oneway model with flat prior). They are unbiased.

Since $1 > l_l \geqslant l$, the Bayesian estimator under Model 1 shrinks more. In addition, the bias of the Bayesian estimator under Model 1 is bigger than that of Bayesian estimator under Model 2.

Thus, if we want the estimators of Model 1 and Model 2 have the same bias, we should choose a bigger parameter ν in Model 1 than that in Model 2. Particularly, if we choose ν_1 in Model 1, then $l = l_l$ by choosing $\nu_2 = \frac{l}{1^T \Sigma_l^{-1}(1-l)}$ in Model 2. In other words, with these ν_1 and ν_2 the corresponding estimators have the same bias.

4.2.2 Comparison and Choice of the Parameter ν Based on MSE

The Model 1 and Model 2 in previous section depend on the parameter ν. Model 3 and Model 4 correspond to $\nu = \infty$. In addition, under these four models the MSE of the estimators can be written as

$$\mathrm{MSE}(\hat{\mu}) = g^2 \mathrm{Var}(\overline{\mathbf{Y}}) + (1-g)^2 \mu^2, \tag{4–19}$$

where $g \in (0,1)$ for Models 1and 2 and $g = 1$ for Models 3 and 4. Thus, we focus on this equation for the four models. We want to find a good choice of ν in order to make the MSE as small as possible.

In fact,

$$\frac{\partial \mathrm{MSE}(\hat{\mu})}{\partial g} = 2(\mu^2 + \mathrm{Var}(\overline{\mathbf{Y}}))(g - \frac{\mu^2}{\mathrm{Var}(\overline{\mathbf{Y}}) + \mu^2}) \begin{cases} \leq 0, & g \leq \frac{\mu^2}{\mathrm{Var}(\overline{\mathbf{Y}})+\mu^2}; \\ > 0, & g > \frac{\mu^2}{\mathrm{Var}(\overline{\mathbf{Y}})+\mu^2}. \end{cases}$$

Thus, $\hat{g} = \frac{\mu^2}{\mathrm{Var}(\overline{\mathbf{Y}})+\mu^2} (< 1)$ makes the MSE smallest.

For Model 1, $g(\nu) = \sum P(\mathbf{A}) \frac{1^\top \Sigma_\mathrm{A}^{-1} 1}{1^\top \Sigma_\mathrm{A}^{-1} 1 + \frac{1}{\nu}}$, which is a monotone and continuous function for ν and the range of this function is $(0,1)$. Thus, there is a unique $\hat{\nu}$, such that

$$\sum P(\mathbf{A}) \frac{\mathbf{1}^\top \mathbf{\Sigma}_\mathrm{A}^{-1} \mathbf{1}}{\mathbf{1}^\top \mathbf{\Sigma}_\mathrm{A}^{-1} \mathbf{1} + \frac{1}{\nu}} = \frac{\mu^2}{\mathrm{Var}(\overline{\mathbf{Y}} | \mathrm{Dirichlet\ model}) + \mu^2}. \tag{4–20}$$

If we know μ, then we can use the bi-section method to find this $\hat{\nu}$. With this choice of the parameter, we can make the MSE smallest for Model 1. If we use estimators for μ and $\mathrm{Var}(\overline{\mathbf{Y}}|\mathrm{Dirichlet\ model})$ in Eq.(4–20), we can get an approximate of $\hat{\nu}$.

For Model 2, $g(\nu) = \frac{1^\top \Sigma_{\mathrm{I}}^{-1} 1}{1^\top \Sigma_{\mathrm{I}}^{-1} 1 + \frac{1}{\nu}}$. We can obtain $\hat{\nu}$ from

$$\frac{\mathbf{1}^\top \mathbf{\Sigma}_{\mathrm{I}}^{-1} \mathbf{1}}{\mathbf{1}^\top \mathbf{\Sigma}_{\mathrm{I}}^{-1} \mathbf{1} + \frac{1}{\nu}} = \frac{\mu^2}{\mathrm{Var}(\overline{\mathbf{Y}}|\mathrm{normal\ model}) + \mu^2}. \tag{4–21}$$

Similarly, if we know μ, then we can use the bi-section method to find this $\hat{\nu}$. With this choice of the parameter, we can make the MSE smallest for Model 2. If we use estimators for μ and $\mathrm{Var}(\overline{\mathbf{Y}}|\mathrm{Dirichlet\ model})$ in Eq.(4–21), we can get an approximate of $\hat{\nu}$.

Models 3 and 4 are the limiting case of Models 1 and 2 respectively. Since the extreme values can be obtained by the interior points, the MSE with the extreme point ν under Models 1 and 2 is smaller than the MSE under Models 3 and 4 respectively.

The above discussion is based on μ, which is unknown in reality. However, we can use $\overline{\mathbf{Y}}$ in Eq. (4–20)-(4–21) to approximate μ to find the good choice of ν.

In addition, with this approximated ν the estimators for μ under Model 1-2 are

$$\hat{\mu}^B = \frac{\overline{\mathbf{Y}}^2}{\mathrm{Var}(\overline{\mathbf{Y}}|\mathrm{Dirichlet\ model}) + \overline{\mathbf{Y}}^2}\overline{\mathbf{Y}},\ \hat{\mu}^I = \frac{\overline{\mathbf{Y}}^2}{\mathrm{Var}(\overline{\mathbf{Y}}|\mathrm{normal\ model}) + \overline{\mathbf{Y}}^2}\overline{\mathbf{Y}}.$$

4.3 The MSE and Bayes Risks

We have obtained the Bayesian estimators under the four models in the previous sections. People might be interested in the question: which model/prior is better? The comparison of the models might need to compare the corresponding Bayesian estimators. We usually calculate the corresponding MSEs and Bayes risks to compare the Bayesian estimators. In this section, we will calculate the MSEs and Bayes risks of the Bayesian estimators to compare the four models. First, we will consider a special case—the oneway model. Then we move to a general case. We always use the sum of the squared error loss in this part.

4.3.1 Oneway Model

In the previous part,we get the estimators of μ under Models 1-4. In this part, we want to investigate more details about the MSEs and Baye risks of the estimators under the oneway model.

Since $\mathrm{MSE}(\hat{\mu}^B) = \mathrm{Var}(l\overline{\mathbf{Y}}) + \mathrm{Bias}(l\overline{\mathbf{Y}}) = l^2\mathrm{Var}(\overline{\mathbf{Y}}) + (1-l)^2\mu^2$, and $l^2\mathrm{Var}(\overline{\mathbf{Y}}) < \mathrm{Var}(\overline{\mathbf{Y}})$, due to the existence of $(1-l)^2\mu^2$, the mean squared errors of Bayesian estimators under Model 1 might cross with the mean squared error of the BLUE.

In addition, the relationship between $\mathrm{Var}(l\overline{\mathbf{Y}}|\mathrm{Dirichlet\ model})$ and $\mathrm{Var}(l_l\overline{\mathbf{Y}}|\mathrm{normal\ model})$ is dependent on the parameter ν, which is proven in the following theorem.

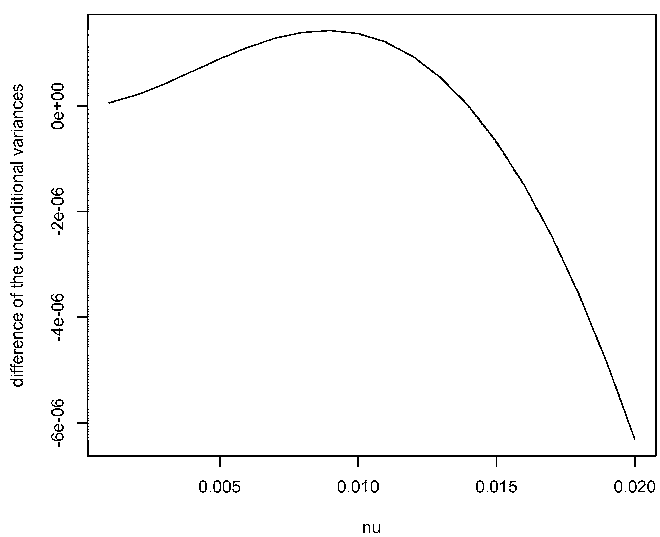

Figure 4-1. $\mathrm{Var}(l_l\overline{\mathbf{Y}}|\mathrm{normal\ model}) - \mathrm{Var}(l\overline{\mathbf{Y}}|\mathrm{Dirichlet\ model})$ for small ν, $\sigma = 1, \tau = .5$.

Theorem 15. *Although* $1 > l_l \geqslant l$, *when* ν *is big, e.g. when* $\nu \geq \frac{ctr+1}{tr}$, $\mathrm{Var}(l\overline{\mathbf{Y}}|\mathrm{Dirichlet\ model}) \geq \mathrm{Var}(l_l\overline{\mathbf{Y}}|\mathrm{normal\ model})$ *and* $MSE(\hat{\mu}^B|\mathrm{Dirichlet\ model}) \geq MSE(\hat{\mu}^l|\mathrm{normal\ model})$.

Proof. See Appendix E. □

The theorem tells us that although $1 > l_l \geqslant l$, $\mathrm{Var}(l\overline{\mathbf{Y}}|\mathrm{Dirichlet\ model}) - \mathrm{Var}(l_l\overline{\mathbf{Y}}|\mathrm{normal\ model})$ depends on ν. Figure 4-1 show the difference of $\mathrm{Var}(l\overline{\mathbf{Y}}|\mathrm{Dirichlet\ model}) -$

$\mathrm{Var}(l_l\overline{\mathbf{Y}}|\text{normal model})$ with different ν. When ν is small, $\mathrm{Var}(l_l\overline{\mathbf{Y}}|\text{normal model})$ is big. When ν is big, $\mathrm{Var}(l\overline{\mathbf{Y}}|\text{Dirichlet model})$ is big.

On the other hand, based on Eq.(4–19), the corresponding Bayes risks with squared error loss under Model 1 and 2 are

$$\begin{aligned}\text{Bayes Risk} &= \int (g^2\mathrm{Var}(\overline{\mathbf{Y}}) + (1-g)^2\mu^2)d\Lambda(\mu)\\ &= g^2\mathrm{Var}(\overline{\mathbf{Y}}) + (1-g)^2\nu\sigma^2\\ &= (\mathrm{Var}(\overline{\mathbf{Y}}) + \nu\sigma^2)g^2 - 2\nu\sigma^2 g + \nu\sigma^2,\end{aligned}$$

where $g = l = \sum P(\mathbf{A})\frac{1^\top\Sigma_A^{-1}1}{1^\top\Sigma_A^{-1}1+\frac{1}{\nu}}$ in Model 1 and $g = l_l = \frac{1^\top\Sigma_l^{-1}1}{1^\top\Sigma_l^{-1}1+\frac{1}{\nu}}$ in Model 2.

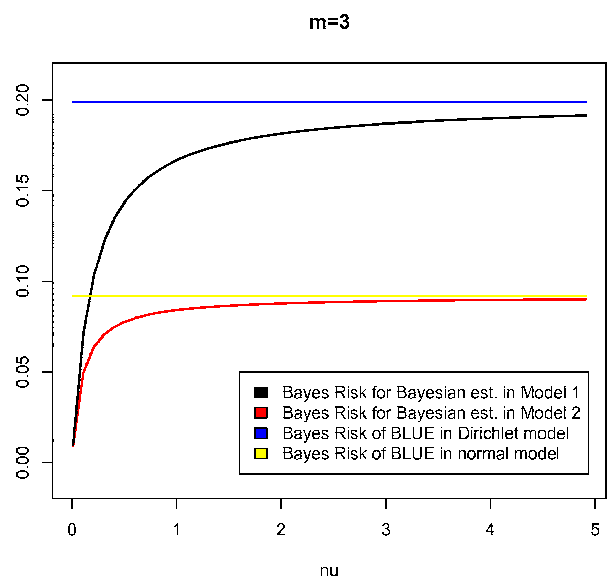

Figure 4-2. The Bayes risks of Bayesian estimators in Models 1-4 and the Bayes risk of BLUE. m=3.

Figure 4-2 shows the Bayes risks of $\hat{\mu}^B$, $\hat{\mu}^I$, $\hat{\mu}^{BF}$ and $\hat{\mu}^{IF}$. The Bayes risks in Figure 4-2 are calculated based on the assumption that the corresponding models used are the true models, i.e., the Bayes risk of $\hat{\mu}^B$ is the Bayes risk$(\hat{\mu}^B|\text{Dirichlet model})$ and the Bayes risk of $\hat{\mu}^I$ is the Bayes risk$(\hat{\mu}^I|\text{normal model})$. In other words, the Bayes risks in Figure 4-2 are calculated under different true models.

However, the true model should be one model in reality, which may be either the Dirichlet process oneway model or the normal oneway model. Thus, it is more meaningful to compare the corresponding MSEs and Bayes risks of the Bayesian estimators under the same true model.

Now, assume that the true model is the Dirichlet model (or the normal model). Then, we can compare the MSEs and Bayes risk of μ^B and μ^I. We have the following theorem.

Theorem 16. *Let $\hat{\mu}^B$ and $\hat{\mu}^I$ be same as before: $\hat{\mu}^B = I\overline{\mathbf{Y}}, \hat{\mu}^I = I_I\overline{\mathbf{Y}}$. We have the following results:*

(1) If the true model is the Dirichlet model, then $\mathrm{Var}(\hat{\mu}^B) \leq \mathrm{Var}(\hat{\mu}^I)$ *and* $\text{Bayes Risk}(\hat{\mu}^B) \leq \text{Bayes Risk}(\hat{\mu}^I)$.

In addition, define $\delta_D = \text{Bayes Risk}(\hat{\mu}^I) - \text{Bayes Risk}(\hat{\mu}^B)$, *then*

$$\delta_D = (\mathrm{Var}(\overline{\mathbf{Y}}|\text{Dirichlet model}) + \nu\sigma^2)(I_I - I)(I_I + I - 2\frac{\nu\sigma^2}{\mathrm{Var}(\overline{\mathbf{Y}}|\text{Dirichlet model}) + \nu\sigma^2});$$

(2) If the true model is the normal model, then $\mathrm{Var}(\hat{\mu}^B) \leq \mathrm{Var}(\hat{\mu}^I)$ *and* $\text{Bayes Risk}(\hat{\mu}^B) \geq \text{Bayes Risk}(\hat{\mu}^I)$.

Let $\delta_N = \text{Bayes Risk}(\hat{\mu}^B) - \text{Bayes Risk}(\hat{\mu}^I)$, *then*

$$\delta_N = (\mathrm{Var}(\overline{\mathbf{Y}}|\text{normal model}) + \nu\sigma^2)(I_I - I)(2\frac{\nu\sigma^2}{\mathrm{Var}(\overline{\mathbf{Y}}|\text{normal model}) + \nu\sigma^2} - I_I - I).$$

(3) $\delta_D \geq \delta_N$.

Proof. See Appendix F. □

This theorem measures and compares the increases in the Bayes risks when the model used is not the true model. If the increase is big, we know that the estimator used is not so good. If the increase is not big, we know that the estimator used is not so bad.

The fact that $\delta_D \geq \delta_N$ tells us that if the true model is the Dirichlet model, and we use the normal model as the true model, the increase in the Bayes risk is bigger. On the other hand, if the true model is the normal model and we use the Dirichlet model as the

true model, the increase in the Bayes risk δ_N is not big. The Dirichlet model has a kind of robust property. Thus, under this case $\hat{\mu}^B$ is better that $\hat{\mu}^I$, i.e., we should use Model 1.

Since $\hat{\mu}^{BF} = \hat{\mu}^{IF} = \overline{\mathbf{Y}}$, they have the same variance, same MSE and same Bayes Risks, i.e. there no difference in using them. Under this case, there is no difference in using Model 3 and Model 4.

As what we will show in the proof of Theorem 18, the Bayes risk of $\hat{\mu}^B$ is always smaller than the Bayes risk of $\hat{\mu}^{BF}$, which means that $\hat{\mu}^B$ is better than $\hat{\mu}^{BF}$. Thus, we should use a normal prior instead of the flat prior.

In summary, we should use the Dirichlet model with a normal prior on μ under the oneway model.

4.3.2 General Linear Mixed Model

In the previous section, we have considered the special case—the oneway model. In this part, we move to a general case. We still consider the four Bayesian setups: Models 1-4. We will compare these four model in parts A-C.

Part A: Compare Model 1 and Model 2

Consider the following example.

Example 17. Let $\boldsymbol{\beta} = [0, 1, 1, 1, 1]^T$,

and $\mathbf{X} = \begin{bmatrix} \mathbf{F} \\ \mathbf{I}_5 \\ \vdots \\ \mathbf{I}_5 \end{bmatrix}$, where $\mathbf{F} = \begin{bmatrix} 1 & & & \\ & 2 & & \\ & & \ddots & \\ & & & 5 \end{bmatrix}$.

Calculate the $\text{Bayes Risk}(\hat{\boldsymbol{\beta}}^B|\text{Dirichlet Model})$,

the $\text{Bayes Risk}(\hat{\boldsymbol{\beta}}^I|\text{normal Model})$,

$\text{MSE}(\hat{\boldsymbol{\beta}}^B|\text{Dirichlet Model})$,

and $\text{MSE}(\hat{\boldsymbol{\beta}}^I|\text{normal Model})$.

The results are shown in Figures 4-3-4-4. As discussed before, the Bayes risks and MSEs in the Figures 4-3-4-4 are calculated based on the assumption that the

corresponding models used are the true models, i.e., the Bayes risks and MSEs are calculated under different true models. We might be more interested in comparing the Bayesian estimators performance under the same true model.

Thus, we move to consider the four differences:

$\text{Bayes Risk}(\hat{\boldsymbol{\beta}}^B|\text{normal Model}) - \text{Bayes Risk}(\hat{\boldsymbol{\beta}}^I|\text{normal Model})$;

$\text{Bayes Risk}(\hat{\boldsymbol{\beta}}^I|\text{Dirichlet Model}) - \text{Bayes Risk}(\hat{\boldsymbol{\beta}}^B|\text{Dirichlet Model})$;

Compare $\text{MSE}(\hat{\boldsymbol{\beta}}^B|\text{normal Model}) - \text{MSE}(\hat{\boldsymbol{\beta}}^I|\text{normal Model})$;

$\text{MSE}(\hat{\boldsymbol{\beta}}^I|\text{Dirichlet Model}) - \text{MSE}(\hat{\boldsymbol{\beta}}^B|\text{Dirichlet Model})$.

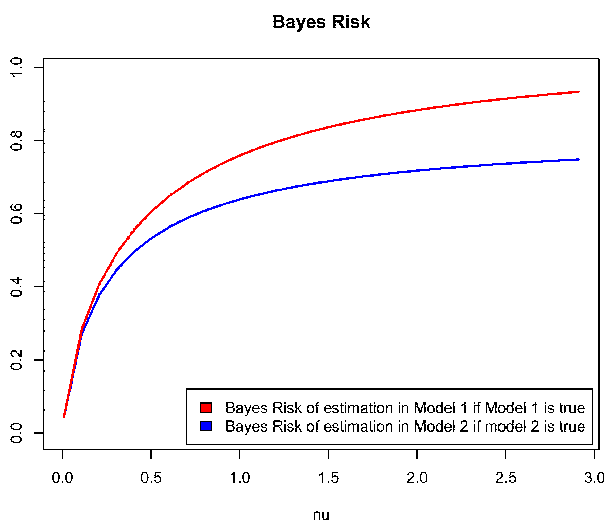

Figure 4-3. The Bayes Risks.

As in the Theorem 16, these four differences measure the increase in the Bayes risks/MSEs when we use a wrong model as the true model. If the difference is big, we know that the estimator used is not so good; if the difference is not big, we know that the estimator used is not bad.

The four differences are shown in the Figures 4-5-4-6. The Figures tells use that we should choose small ν, since the differences of the Bayes risks/MSEs are small when ν is small.

In addition, when ν is small, the increase in the Bayes risk/MSE by using the Dirichlet process mixed model is not big if the true model is the normal model. Similarly,

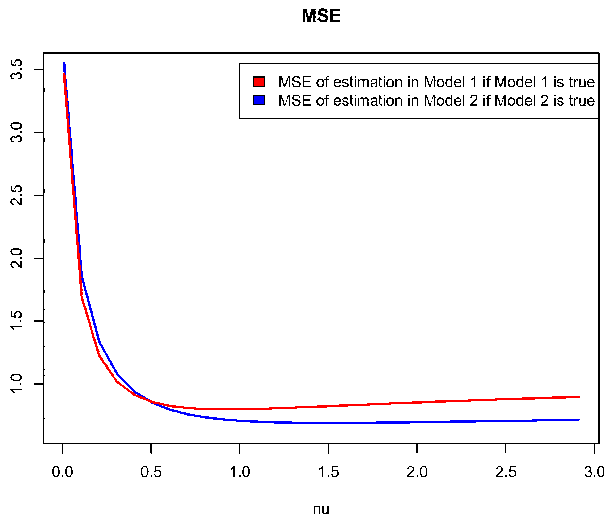

Figure 4-4. The MSEs under Different Models.

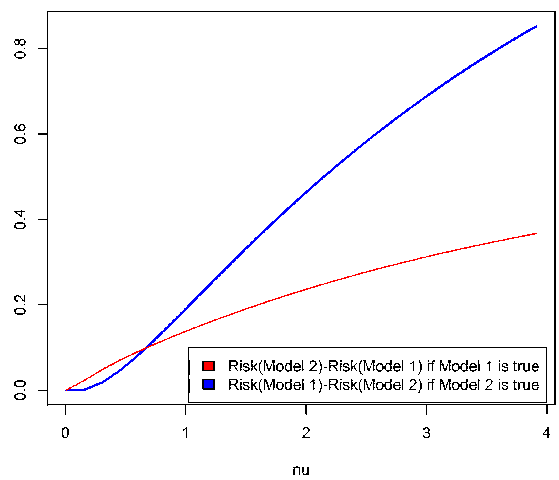

Figure 4-5. The Bayes Risks. $\sigma^2 = 5$.

when ν is big, the increase in the Bayes risk/MSE by using the normal model is not big if the true model is the Dirichlet process mixed model. Thus, the estimator by Model 1 is better if ν is small. If we choose ν to be big, then the estimator by the normal model (Model 2) is better.

Part B: Compare Model 3 and Model 4.

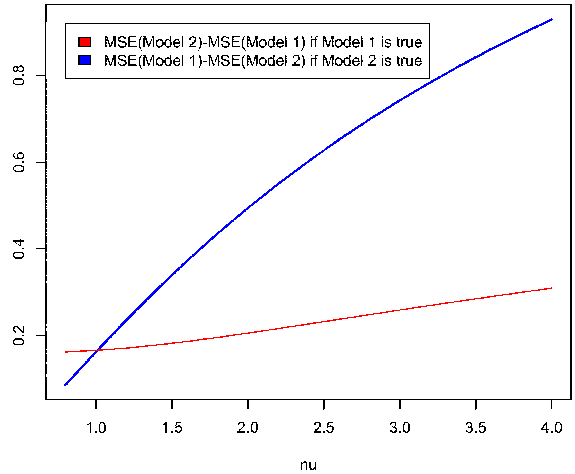

Figure 4-6. The MSEs.$\sigma^2 = 5$.

First compare the estimators under Model 3 and Model 4, i.e. compare the estimators $\boldsymbol{\beta}^{BF}$ and $\boldsymbol{\beta}^{IF}$. Note that both $\boldsymbol{\beta}^{BF}$ and $\boldsymbol{\beta}^{IF}$ are unbiased. $\boldsymbol{\beta}^{IF}$ is the BLUE under the normal model.

If OLS is the BLUE ($\mathbf{ZZ'H}$ and $\mathbf{ZJZ'H}$ are symmetric, where $\mathbf{H}$ is same as before), then $\boldsymbol{\beta}^{IF}$ is the OLS, and the BLUE is equal to the OLS under the Dirichlet model. Thus, $\boldsymbol{\beta}^{IF}$ is the BLUE under the Dirichlet model.

Then we have that $\mathrm{Var}_{\boldsymbol{\beta}}(\boldsymbol{\beta}^{BF}) - \mathrm{Var}_{\boldsymbol{\beta}}(\boldsymbol{\beta}^{IF})$ is always a nonnegative definite matrix, no matter what the true distribution of $\mathbf{Y}$ is. $\boldsymbol{\beta}^{IF}$ is better than $\boldsymbol{\beta}^{BF}$. That is, we should use the normal model when we use flat prior on $\boldsymbol{\beta}$ and the OLS is equal to the BLUE.

If OLS is not the BLUE, I only have numerical results currently.

Table 4-1. The MSEs with different σ^2. $\sigma^2 = \tau^2 = 1$.

True Model	MSE($\hat{\boldsymbol{\beta}}^{BF}|\sigma^2 = 1$)	MSE($\hat{\boldsymbol{\beta}}^{IF}|\sigma^2 = 1$)
Normal Model	1.257	0.828
Dirichlet Model	1.071	1.2325

Table 4-2. The MSEs with different σ^2. $\sigma^2 = \tau^2 = 5$.

True Model	MSE($\hat{\boldsymbol{\beta}}^{BF}|\sigma^2 = 5$)	MSE($\hat{\boldsymbol{\beta}}^{IF}|\sigma^2 = 5$)
Normal Model	6.287	4.142
Dirichlet Model	5.369	6.163

The Table 4-1-4-2 shows the MSEs under different models. From the results we find that the MSEs are close. If we consider the differences of MSEs as what we have done in part A, we will find that the normal model (Model 4) is better now.

In short, Model 4(the normal model) is better now.

Part C: Compare Model 1 and Model 3. Let us move to compare Model 1and Model 3. In the oneway model section, we have known that with "good" choice of the parameter ν, Model 1 is better that Model 3. Do we have similar results in the general cases?

Again consider the same β and $\mathbf{X}$ in Example 17. The corresponding MSEs of $\hat{\beta}^B$ and $\hat{\beta}^{BF}$ are shown in Figure 4-7. It is obvious that $\hat{\beta}^B$ is better under this case since the MSEs of $\hat{\beta}^B$ is always smaller. Thus, Model 1 is better.

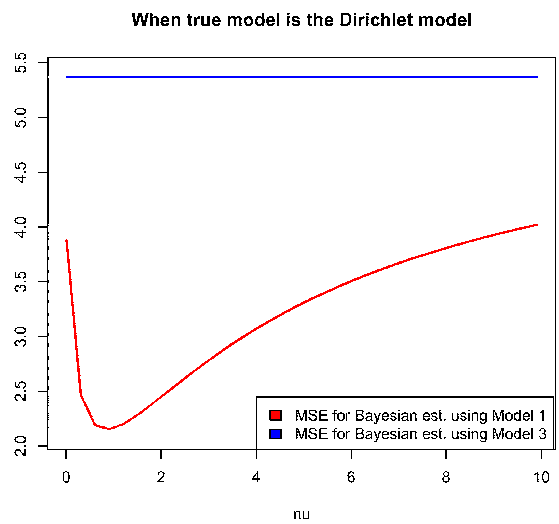

Figure 4-7. The MSEs. $\sigma^2 = 5$.

In short, when comparing Model 1 VS Model 2(Dirichlet+normal VS normal+normal), we should choose small ν and the Dirichlet model (Dirichlet+normal). When comparing Model 3 VS Model 4(Dirichlet+flat prior VS normal+flat prior), we should choose Model 4(the normal +flat prior). When comparing Model 1 VS Model 3(Dirichlet+normal VS Dirichlet+flat prior), we should choose Model 1 (Dirichlet+normal).

CHAPTER 5
MINIMAXITY AND ADMISSIBILITY

Under the classical normal mixed model, we know the minimax estimators of the fixed effects in some special cases. We want to investigate if there are still some minimax estimators in the Dirichlet process mixed models. In this section, we will discuss the minimaxity and admissibility of the estimators, and to show the admissibility of the confidence intervals. We only consider the cases under the squared error loss in this part.

5.1 Minimaxity and Admissibility of Estimators

First we consider a special case—the Dirichlet process oneway model: $\mathbf{Y} = \mathbf{1}\mu + \mathbf{A}\boldsymbol{\eta} + \boldsymbol{\varepsilon}$. The following theorem shows that $\overline{\mathbf{Y}}$ is minimax and admissible under the Dirichlet process oneway model.

Theorem 18. $\overline{\mathbf{Y}}$ *is minimax and admissible with the Dirichlet process random effect.*

Proof. Consider a sequence of prior distributions $\mu \sim \Lambda_n(\mu) = N(0, n\sigma^2)$. Let $l_n = \sum P(\mathbf{A})\frac{1^\top\Sigma_A^{-1}1}{1^\top\Sigma_A^{-1}1+\frac{1}{n}}$ and $r_0 = \mathrm{Var}(\overline{\mathbf{Y}})$. By the previous discussion, we know that $r_0 = c_0\sigma^2$, where c_0 is a constant free of σ^2.

Then the corresponding Bayes estimator is $\mu_n^B = l_n\overline{\mathbf{Y}}$.

By the similar calculation in previous section, the corresponding Bayes risk is

$$r_n = l_n^2\mathrm{Var}(\overline{\mathbf{Y}}) + n(1-l_n)^2\sigma^2 = (\mathrm{Var}(\overline{\mathbf{Y}}) + n\sigma^2)l_n^2 - 2n\sigma^2 l_n + n\sigma^2.$$

Since $n(1-l_n)^2 \to 0$ and $l_n \to 1$ as $n \to \infty$, $r_n \to r_0 = \mathrm{Var}(\overline{\mathbf{Y}})$.

By the inequality between the harmonic mean and arithmetic mean we have $\frac{1^\top\Sigma_A 1}{(rt)^2} \geq \frac{1}{1^\top\Sigma_A^{-1}1}$, where $\Sigma_A = \mathbf{I} + c\mathbf{B}\mathbf{A}\mathbf{A}'\mathbf{B}'$, since $\mathbf{1}^\top\boldsymbol{\Sigma}_A^{-1}\mathbf{1} = \sum_{i=1}^{rt}\frac{1}{ctr_{index(i)}+1}$.

Then,

$$\frac{\mathrm{Var}(\overline{\mathbf{Y}})}{\sigma^2} = \frac{1}{(rt)^2}\sum P(A)\mathbf{1}^\top\boldsymbol{\Sigma}_A\mathbf{1} \geq \sum P(\mathbf{A})\frac{1}{\mathbf{1}^\top\boldsymbol{\Sigma}_A^{-1}\mathbf{1}}.$$

Let $C_A = \mathbf{1}^\top \mathbf{\Sigma}_A^{-1}\mathbf{1}$. Then

$$\begin{aligned} r_n &= \mathrm{Var}(\overline{\mathbf{Y}})\{(\sum_A P(\mathbf{A})\frac{C_A}{C_A+\frac{1}{n}})^2 + \frac{\sigma^2}{n\mathrm{Var}(\overline{\mathbf{Y}})}\sum_A P(\mathbf{A})(\frac{1}{C_A+\frac{1}{n}})^2\} \\ &\le \mathrm{Var}(\overline{\mathbf{Y}})\{(\sum_A P(\mathbf{A})\frac{C_A}{C_A+\frac{1}{n}})^2 + \frac{1}{n}\sum_A P(\mathbf{A})\frac{1}{C_A+\frac{1}{n}}\} \\ &= \mathrm{Var}(\overline{\mathbf{Y}})\{(\sum_A P(\mathbf{A})\frac{C_A}{C_A+\frac{1}{n}})^2 + (1-\sum_A P(\mathbf{A})\frac{C_A}{C_A+\frac{1}{n}})\} \le \mathrm{Var}(\overline{\mathbf{Y}}). \end{aligned}$$

Thus, $r_n \to r_0 = \mathrm{Var}(\overline{\mathbf{Y}})$ and $r_n \le r_0$.

Since $\sup_\mu \mathbf{R}(\mu, \overline{\mathbf{Y}}) = \mathrm{Var}(\overline{\mathbf{Y}}) = r_0$, by the Theorem 5.1.12 in Lehmann and Casella (1998), $\overline{\mathbf{Y}}$ is minimax.

On the other hand, the above discussion show that conditions (a)-(c) of the Theorem 5.7.13 in Lehmann and Casella (1998)) are satisfied. thus, $\overline{\mathbf{Y}}$ is admissible. □

Now move to the model: $\mathbf{Y} = \mathbf{B}\boldsymbol{\beta} + \mathbf{BA}\boldsymbol{\eta} + \boldsymbol{\varepsilon}$, where $\mathbf{B}$ is the same as before. Then the BLUE (OLS) is $\underline{\overline{\mathbf{Y}}} = \{\overline{\mathbf{Y}}_1, \overline{\mathbf{Y}}_2, ..., \overline{\mathbf{Y}}_r\}$, where $\overline{\mathbf{Y}}_i = \frac{1}{t}\sum_s Y_{is}$, $i = 1, ..., r$.

Similarly with the squared error loss $\frac{1}{r}\sum_i(\delta_i - \beta_i)^2$, $\underline{\overline{\mathbf{Y}}}$ is minimax.

5.2 Admissibility of Confidence Intervals

In this section, we show the admissibility of the usual frequentist confidence interval.

Theorem 19. *The confidence interval for μ with the form $(\overline{\mathbf{Y}} - c_0, \overline{\mathbf{Y}} + c_0)$ is admissible in the Dirichlet oneway model.*

Proof. As discussed before, $\overline{\mathbf{Y}}|\mathbf{A} \sim N(\mu, \sigma_A^2)$, where $\sigma_A^2 = \frac{1}{n^2}(n\sigma^2 + \tau^2 t^2 \sum_l r_l^2)$. Let $f(y|\mathbf{A})$ be the normal density with mean μ and variance σ_A^2, and the marginal density for $\overline{\mathbf{Y}}$ be $f_{\overline{Y}}(y)$. Then

$$f_{\overline{Y}}(y) = \sum_A f(y|\mathbf{A})P(\mathbf{A}), \tag{5–1}$$

i.e. the density of $\overline{\mathbf{Y}}$ is the mixture of normal densities with the same mean μ.

Assume there is another interval $(g(\mathbf{Y}), h(\mathbf{Y}))$ satisfying:

$$(1*)\ h(\mathbf{Y}) - g(\mathbf{Y}) \leqslant 2c_0;$$

$$(2*)\ P_\mu(\mu \in (\overline{\mathbf{Y}} - c_0, \overline{\mathbf{Y}} + c_0)) = \sum P(\mathbf{A}) P_\mu(\mu \in (\overline{\mathbf{Y}} - c_0, \overline{\mathbf{Y}} + c_0)|A)$$
$$\leqslant P_\mu(\mu \in (g(\mathbf{Y}), h(\mathbf{Y}))) = \sum P(\mathbf{A}) P_\mu(\mu \in (g(\mathbf{Y}), h(\mathbf{Y}))|A),$$

where (2*) holds for every μ and the strict inequality in (2*) holding for at least one μ.

By this assumption, we know there is at least one matrix $\mathbf{A}_0$ such that

$$P_\mu(\mu \in (\overline{\mathbf{Y}} - c_0, \overline{\mathbf{Y}} + c_0)|\mathbf{A}_0) \leqslant P_\mu(\mu \in (g(\mathbf{Y}), h(\mathbf{Y}))|\mathbf{A}_0), \tag{5–2}$$

where Eq.(5–2) holds for every μ and the strict inequality holding for at least one μ. This is contradictory to the fact that for every $\mathbf{A}$, $\overline{\mathbf{Y}} \pm c_0$ is admissible for the normal density $f(y|\mathbf{A})$.

Thus, the assumption is false. Thus, there is no such interval $(g(\mathbf{Y}), h(\mathbf{Y}))$. That is, the confidence interval for μ with the form $(\overline{\mathbf{Y}} - c_0, \overline{\mathbf{Y}} + c_0)$ is admissible. □

For the model: $\mathbf{Y} = \mathbf{B}\boldsymbol{\beta} + \mathbf{A}\boldsymbol{\eta} + \boldsymbol{\varepsilon}$, $\overline{\mathbf{Y}}_i|\mathbf{A} \sim N(\mu, \widetilde{\sigma}^2_{A,i})$. By similar proof, we can show that the confidence intervals with the form $(\overline{\mathbf{Y}}_i - c_0, \overline{\mathbf{Y}}_i + c_0)$ for β_i is also admissible.

CHAPTER 6
CONCLUSIONS AND FUTURE WORK

The previous chapters discuss the linear mixed model with Dirichlet Process random effects from both the frequentist and Bayesian viewpoint. We first consider the Dirichlet process as a model for classical random effects, and investigate their effect on frequentist estimation in the linear mixed model. In my dissertation I discuss the relationship between the BLUE (Best Linear Unbiased Estimator) and OLS (Ordinary Least Squares) in Dirichlet process mixed models, and also give conditions under which the BLUE coincides with the OLS estimator in the Dirichlet process mixed model. In addition, I investigate the model from the Bayesian view, and discuss the properties of estimators under different model assumptions, compare the estimators under the frequentist model and different Bayesian models, and investigate minimaxity. Furthermore, we apply the linear mixed model with Dirichlet Process random effects to a real data set and get satisfactory results.

Literature and present work offer the potential to develop some further results. My future research plan focus on the Dirichlet mixed model from both the theoretical way and application.

(1) In the real example of Kyung et al. (2010), the length of the credible interval obtained by the Dirichlet mixed model is smaller than that by the classical mixed model. Is this always true in general cases? I would like to investigate more details to this question. The credible intervals by different Bayesian models might have different coverage probabilities. The Dirichlet mixed model gave shorter credible intervals in Kyung et al. (2010). Do these shorter intervals have higher or lower coverage probabilities? Are the coverage probabilities of the credible intervals by the Dirichlet mixed model always larger? I would like to compare the coverage probabilities of the credible intervals under the Dirichlet process mixed model and the classical normal mixed model in general.

(2) We have done some simulation studies and an application to a real data set in the previous chapters. The results are satisfactory. Thus, I would like to apply this model to more real problems, which might want a more flexible and possibly nonparametric structure, such as some genetic problems and problems in social science.

APPENDIX A
PROOF OF THEOREM 2

Proof. Let $h(r,m) = \sum_{i=1}^{r-1} im\frac{\Gamma(m+r-1-i)\Gamma(i)}{\Gamma(m+r)}$. We only need to prove that $h(r,m)$ is a decreasing in m on $m \geq \sqrt{(r-2)(r-1)}$ or $0 \leq m \leq 2$.

Let $h_i(r,m) = im\frac{\Gamma(m+r-1-i)\Gamma(i)}{\Gamma(m+r)}$. Then $h(r,m) = \sum_{i=1}^{r-1} h_i(r,m)$.

There are four cases for r.

Case 1: When $r = 2$, $h = h_1 = \frac{m}{m(m+1)} = \frac{1}{m+1}$, which is decreasing obviously.

Case 2: When $r = 3$, $h = h_1 + h_2 = \frac{m}{(m+1)(m+2)} + \frac{2m}{m(m+1)(m+2)} = \frac{1}{m+1}$, which is decreasing obviously.

Case 3: When $r = 4$, $h = h_1 + h_2 + h_3 = \frac{1}{m+3} + \frac{4}{(m+1)(m+2)(m+3)}$ which is decreasing obviously.

Case 4: $r \geq 5$. We will use mathematical induction method to prove that $h(r,m)$ is a decreasing on $m \geq \sqrt{(r-2)(r-1)}$ or $0 \leq m \leq 2$.

When $r = 2$, $h = h_1 = \frac{m}{m(m+1)} = \frac{1}{m+1}$, which is decreasing obviously.

Assume that for $r = s$, $h(s,m)$ is a decreasing on $m \geq \sqrt{(s-2)(s-1)}$ or $0 \leq m \leq 2$. i.e., $\frac{\partial h(s,m)}{\partial m} \leq 0$ on $m \geq \sqrt{(s-2)(s-1)}$ or $0 \leq m \leq 2$.

Now consider $r = s+1$.

We need to show that $h(s+1,m) = \sum_{i=1}^{s} h_i(s+1,m) = \sum_{i=1}^{s} im\frac{\Gamma(m+s-i)\Gamma(i)}{\Gamma(m+s+1)}$ is decreasing on $m \geq \sqrt{s(s-1)}$ or $0 \leq m \leq 2$.

By regular calculation, we have that $h_i(s+1,m) = (1 - \frac{i+1}{m+r})h_i(s+1,m)$, for $i = 1, ..., s-1$.

Then,

$$
\begin{aligned}
h(s+1,m) &= \sum_{i=1}^{s-1} h_i(s+1,m) + h_s(s+1,m) \\
&= h_s(s+1,m) + \sum_{i=1}^{s-1}(1 - \frac{i+1}{m+r})h_i(s,m) \\
&= \frac{m\Gamma(s+1)\Gamma(m)}{\Gamma(m+s+1)} + h(s,m) - \sum_{i=1}^{s-1} m\frac{\Gamma(m+s-1-i)\Gamma(i+2)}{\Gamma(m+s+1)}
\end{aligned}
$$

$$= h(s,m) + \frac{m\Gamma(s+1)\Gamma(m)}{\Gamma(m+s+1)} - \sum_{j=2}^{s} m\frac{\Gamma(m+s-j)\Gamma(j+1)}{\Gamma(m+s+1)}$$
$$= h(s,m) + \frac{m\Gamma(s+1)\Gamma(m)}{\Gamma(m+s+1)} - h(s+1,m) + h_1(s+1,m)$$

In another word,

$$h(s+1,m) = \frac{1}{2}[h(s,m) + \frac{m\Gamma(s+1)\Gamma(m)}{\Gamma(m+s+1)} + \frac{m\Gamma(m+s-1)}{\Gamma(m+s+1)}] \tag{A–1}$$

Thus, we only need to prove that $\frac{\partial\{\frac{\Gamma(s+1)\Gamma(m+1)}{\Gamma(m+s+1)} + \frac{m\Gamma(m+s-1)}{\Gamma(m+s+1)}\}}{\partial m} < 0$ when $m^2 \geq s(s-1)$ or $m \leq 2$..

In fact,

$$\frac{\partial\{\frac{\Gamma(s+1)\Gamma(m+1)}{\Gamma(m+s+1)} + \frac{m\Gamma(m+s-1)}{\Gamma(m+s+1)}\}}{\partial m}$$
$$= \frac{(s-1)s - m^2}{(s+m)^2(m+s-1)^2} - \Gamma(s+1)\frac{\sum_{i=1}^{s}\prod_{j=1,j\neq i}^{s}(m+j)}{(m+s)^2\cdots(m+1)^2}$$
$$= \frac{(s-1)s - m^2}{(s+m)^2(m+s-1)^2} - \frac{\Gamma(s+1)}{\prod_{j=1}^{s}(m+j)}\sum_{j=1}^{s}\frac{1}{m+j}.$$

it is obvious that when $m^2 \geq s(s-1)$, $\frac{\partial\{\frac{\Gamma(s+1)\Gamma(m+1)}{\Gamma(m+s+1)} + \frac{m\Gamma(m+s-1)}{\Gamma(m+s+1)}\}}{\partial m} < 0$.

Now consider the interval $0 \leq m \leq 2$. Since $\frac{(s-1)s-m^2}{(s+m)^2(m+s-1)^2}$ is decreasing, we have

$$\frac{\partial\{\frac{\Gamma(s+1)\Gamma(m+1)}{\Gamma(m+s+1)} + \frac{m\Gamma(m+s-1)}{\Gamma(m+s+1)}\}}{\partial m}$$
$$= \frac{(s-1)s - m^2}{(s+m)^2(m+s-1)^2} - \frac{\Gamma(s+1)}{\prod_{j=1}^{s}(m+j)}\sum_{j=1}^{s}\frac{1}{m+j}$$
$$\leq \frac{s-1}{s(s+1)^2} - \frac{\Gamma(s+1)}{\prod_{j=1}^{s}(2+j)}\sum_{j=1}^{s}\frac{1}{2+j},$$
$$= \frac{s-1}{s(s+1)^2} - \frac{2}{(s+2)(s+1)}\sum_{j=1}^{s}\frac{1}{2+j}$$
$$\leq \frac{(1-\ln 4)s^2 - (\ln 4 + 3)s + 2}{s(s+1)(s+2)} < 0.$$

Thus, if $r \geq 5$, $h(s,m)$ is a decreasing on $m^2 \geq (r-2)(r-1)$ or $m \leq 2$.

In another word, if $r \geq 5$, d is a decreasing on $m^2 \geq (r-2)(r-1)$ or $m \leq 2$. When $r \leq 5$, d is a decreasing function of m on the whole real line. □

APPENDIX B
PROOF OF THEOREM 3

Proof. We only need to show that $\mathbf{g}$ is the eigenvector of matrix $\mathbf{V}$ if and only if $\mathbf{g} \in \bigcup_{j=2}^{4} \mathbf{E}_j$.

1. Sufficient part. Assume $\mathbf{g} \in \bigcup_{j=2}^{4} \mathbf{E}_j$.

By regular calculation we know that $\mathbf{ZWZ'g} = \text{a constant } \times \mathbf{g}$, which means $\mathbf{g}$ is an eigenvector. The proof of this part is complete.

2. Necessary part. Assume $\mathbf{g}$ is an eigenvector of $\mathbf{ZWZ'}$. If we can show that $\bigcup_{j=2}^{4} \mathbf{E}_j$ contains all the eigenvectors, then $\mathbf{g} \in \bigcup_{j=2}^{4} \mathbf{E}_j$.

Now move to show that $\bigcup_{j=2}^{4} \mathbf{E}_j$ contains all the eigenvectors of $\mathbf{ZWZ'}$.

For every $\mathbf{h} \in \mathbf{E}_1$, by the assumption and algebra calculation, we know that $\mathbf{h}$ is also an eigenvector corresponding to a certain nonzero eigenvalue of $\mathbf{WZ'Z}$. The sum of the geometric multiplicities of all these distinct nonzero eigenvalues is r.

By the algebra theory, we know that if the $\mathbf{h}$ is an eigenvector of $(\mathbf{WZ'})\mathbf{Z} = \mathbf{WZ'Z}$ corresponding to a nonzero eigenvalue, then $\mathbf{Zh}$ is an eigenvector of $\mathbf{Z}(\mathbf{WZ'}) = \mathbf{ZWZ'}$ corresponding to the same eigenvalue with the same algebra multiplicity and geometric multiplicity.

So we know that every element in $\mathbf{E}_2 \bigcup \mathbf{E}_3$ is an eigenvector corresponding to a certain nonzero eigenvalue of $\mathbf{ZWZ'}$ and the sum of the geometric multiplicities of all these distinct nonzero eigenvalues is also r. $\mathbf{E}_2 \bigcup \mathbf{E}_3 \bigcup \{\mathbf{0}\}$ is the union of the corresponding eigenspaces.

For every $\mathbf{h} \in \mathbf{E}_4$, we have $\mathbf{ZWZ'h} = \mathbf{0}$, which is the eigenvector corresponding to eigenvalue 0. The geometric multiplicity for this eigenvalue is $n - r$. $\mathbf{E}_4 \bigcup \{\mathbf{0}\}$ is the corresponding eigenspace.

The total sum of the geometric multiplicities of all these distinct (zero and nonzero) eigenvalues is $n - r + r = n$. In another word, $\bigcup_{j=2}^{4} \mathbf{E}_j$ contains all the eigenvectors of $\mathbf{ZWZ'}$. Thus, $\mathbf{g} \in \bigcup_{j=2}^{4} \mathbf{E}_j$. The proof of necessary part is complete. $\square$

APPENDIX C
EVALUATION OF EQUATION (2-4)

Let $\mathbf{A} = [\mathbf{a}_1, ..., \mathbf{a}_r]'$.

It is straightforward to check that $\sum_A P(\mathbf{A})\mathbf{a}_i'\mathbf{a}_i = 1$, Moreover, for $i \neq j$,

$$d = E(\mathbf{a}_i'\mathbf{a}_j) = P(\text{i,j in the same cluster}) = P(\text{1,2 in the same cluster}).$$

We can consider that $(1, 2)$ forms a "new" unit, and we partition the $r-1$ subjects to get

$$\begin{aligned}
P(\text{1,2 in the same cluster}) &= \frac{\Gamma(m)}{\Gamma(m+r)} \sum_{k=1}^{r-1} \sum_{C:|C|=k, \sum r_i = r-1} \Gamma(r_1+1) m^k \prod_{j=2}^{k} \Gamma(r_j) \\
&= \frac{\Gamma(m+r-1)}{\Gamma(m+r)} \\
&\times \{ \frac{\Gamma(m)}{\Gamma(m+r-1)} \sum_{k=1}^{r-1} \sum_{C:|C|=k, \sum r_i = r-1} m^k r_1 \prod_{j=1}^{k} \Gamma(r_j) \} \\
&= \frac{\Gamma(m+r-1)}{\Gamma(m+r)} E_{m,r-1}[r_1],
\end{aligned}$$

where $E_{m,r-1}[r_1]$ means this expectation is corresponding to the partitions of $r-1$ subjects and C means a partition. $|C| = k$ means a partition divides the sample to k groups.

Using an idea similar to above, we calculate,

$$\begin{aligned}
E_{m,r-1}[r_1] \quad &= 1 \times P(r_1 = 1) + ... + (r-1) \times P(r_1 = r-1) \\
&= 1 \times P(r_2, ..., r_k \text{ form partitions of } r-1-1 \text{ subjects}) + ... \\
&+(r-2) \\
&\times P(r_2, ..., r_k \text{ form partitions of r-1-(r-2) subjects}) \\
&+(r-1)P(r_1 = r-1) \\
&= \frac{m\Gamma(m+r-2)}{\Gamma(m+r-1)} \times \\
&\left\{ \frac{\Gamma(m)}{\Gamma(m+r-2)} \sum_{k=2}^{r-1} \sum_{C:|C|=k-1, \sum_{j=2}^{k} r_j = r-2} m^{k-1} \prod_{j=2}^{k} \Gamma(r_j) \right\} \times \Gamma(1)
\end{aligned}$$

$$
\begin{aligned}
&+\cdots+(r-1)\times\frac{m\Gamma(m+r-1-r+1)\Gamma(r-1)}{\Gamma(m+r-1)}\\
=&\frac{m\Gamma(m+r-2)\Gamma(1)}{\Gamma(m+r-1)}+2\times\frac{m\Gamma(m+r-1-2)\Gamma(2)}{\Gamma(m+r-1)}\\
&+\cdots+(r-1)\times\frac{m\Gamma(m+r-1-r+1)\Gamma(r-1)}{\Gamma(m+r-1)}\\
=&\sum_{i=1}^{r-1} im\frac{\Gamma(m+r-1-i)\Gamma(i)}{\Gamma(m+r-1)}
\end{aligned}
$$

When $m \to 0$,

$$
\begin{aligned}
d &= \sum_{i=1}^{r-1} im\frac{\Gamma(m+r-1-i)\Gamma(i)}{\Gamma(m+r)}\\
&= [\sum_{i=1}^{r-2} im\frac{\Gamma(m+r-1-i)\Gamma(i)}{\Gamma(m+r)}+(r-1)\frac{m\Gamma(m)\Gamma(r-1)}{\Gamma(r)}]\\
&\to 1.
\end{aligned}
$$

When $m \to \infty$,

$$
\begin{aligned}
d &= \sum_{i=1}^{r-1} im\frac{\Gamma(m+r-1-i)\Gamma(i)}{\Gamma(m+r)}\\
&< \sum_{i=1}^{r-1} i(m+r-1-i)\frac{\Gamma(m+r-1-i)\Gamma(i)}{\Gamma(m+r)}\\
&< \sum_{i=1}^{r-1} i\frac{\Gamma(m+r-i)\Gamma(i)}{\Gamma(m+r)}\\
&< \sum_{i=1}^{r-1} i\frac{\Gamma(i)}{m+r-1}\\
&\to 0.
\end{aligned}
$$

APPENDIX D
PROOF OF THEOREM 14

Proof. Since $\Sigma_A = \mathbf{I} + c\mathbf{BAA'B'}$,

$$\Sigma_A^{-1} = \mathbf{I} - \mathbf{BA}(\mathbf{A'B'BA} + \frac{\mathbf{I}}{c})^{-1}\mathbf{A'B'},$$

where $(\mathbf{A'B'BA} + \frac{\mathrm{I}}{c})^{-1} = \mathrm{diag}\{\frac{c}{ctr_j+1}\}$, r_j is the sum of j-th column of matrxi $\mathbf{A}$.

Denote $index(i)$ as the function to identify the cluster that the i-th observation of response belongs to. For example, $index(3) = 2$ means that the 3rd observation of response corresponding to the 2nd cluster, i.e, $y_3 = \mu + \psi_2 + \epsilon_{23}$. Then,

$$\begin{aligned}
\mathbf{1}^\top \mathbf{\Sigma}_\mathrm{A}^{-1} &= \mathbf{1}^\top[\mathbf{I} - \mathbf{BA}\mathrm{diag}\{\frac{c}{ctr_j+1}\}\mathbf{A'B'}] \\
&= \underline{1}^T - (tr_1, ..., tr_k)\mathrm{diag}\{\frac{c}{ctr_j+1}\}\mathbf{A'B'} \\
&= \underline{1}^T - (\frac{ctr_1}{ctr_1+1}, \frac{ctr_2}{ctr_2+1}, ..., \frac{ctr_k}{ctr_k+1})\mathbf{A'B'} \\
&= \underline{1}^T - (\frac{ctr_{index(1)}}{ctr_{index(1)}+1}, \frac{ctr_{index(2)}}{ctr_{index(2)}+1}, ..., \frac{ctr_{index(n)}}{ctr_{index(n)}+1}) \\
&= (\frac{1}{ctr_{index(1)}+1}, \frac{1}{ctr_{index(2)}+1}, ..., \frac{1}{ctr_{index(n)}+1})
\end{aligned}$$

$$\frac{\mathbf{1}^\top \mathbf{\Sigma}_\mathrm{A}^{-1}}{\mathbf{1}^\top \mathbf{\Sigma}_\mathrm{A}^{-1}\mathbf{1} + \frac{1}{\nu}} = \frac{(\frac{1}{\mathrm{ctr_{index(1)}}+1}, \frac{1}{\mathrm{ctr_{index(2)}}+1}, ..., \frac{1}{\mathrm{ctr_{index(n)}}+1})}{\sum_{\mathrm{i=1}}^{\mathrm{n}} \frac{1}{\mathrm{ctr_{index(i)}}+1} + \frac{1}{\nu}}. \tag{D–1}$$

For a certain $(index(1), ..., index(n))$, let $(index(1)', ..., index(n)')$ be a possible permutation. If $\mathbf{A}_1$ is corresponding to $(index(1), ..., index(n))$ and $\mathbf{A}_2$ is corresponding to $(index(1)', ..., index(n)')$, then $P(\mathbf{A}_1) = P(\mathbf{A}_2)$.

Set $f = \sum_{i=1}^{n} \frac{1}{ctr_{index(i)}+1}$, we have:

$$\begin{aligned}
&\sum_{permutations\ of\ (index(1),...,index(n))} \frac{\mathbf{1}^\top \mathbf{\Sigma}_\mathrm{A}^{-1}}{\mathbf{1}^\top \mathbf{\Sigma}_\mathrm{A}^{-1}\mathbf{1} + \frac{1}{\nu}} \\
&= \sum_{permutations\ of\ (index(1),...,index(n))} \frac{(\frac{1}{ctr_{index(1)}+1}, \frac{1}{ctr_{index(2)}+1}, ..., \frac{1}{ctr_{index(n)}+1})}{f + \frac{1}{\nu}} \\
&= \#\ of\ permutations \times \frac{1}{f + \frac{1}{\nu}}(f_1, ..., f_1)
\end{aligned}$$

where f_1 is a constant. Since $f = \frac{n}{\sigma^2} - \frac{1}{\sigma^2}\sum_{i=1}^{n}\frac{ctr_{index(i)}}{ctr_{index(i)}+1}$, we have $\sum f_1 = f$, i.e, $f_1 = \frac{f}{n}$

Thus,

$$\sum_{permutations\ of\ (index(1),...,index(n))} \frac{\mathbf{1}^\top \mathbf{\Sigma}_\mathsf{A}^{-1}}{\mathbf{1}^\top \mathbf{\Sigma}_\mathsf{A}^{-1}\mathbf{1} + \frac{1}{\nu}}\mathbf{Y} = \frac{\#\ of\ permutations}{n}\frac{f}{f+\nu}(1,...,1)\mathbf{Y}$$
$$= \#\ of\ permutations \times \frac{f}{f+\frac{1}{\nu}}\overline{Y}.$$

Let $\mathbf{S} = \{\text{all possible } (index(1), ..., index(n))\}$.

By previous definition we have

$l = \sum h_i = \{\sum_{(index(1),...,index(n))\in \mathsf{S}} P(\mathbf{A})\frac{\sum_{i=1}^{n}\frac{1}{ctr_{index(i)}+1}}{\sum_{i=1}^{n}\frac{1}{ctr_{index(i)}+1}+\frac{1}{\nu}}\}$.

Let set $\mathbf{S}_1$ be the largest subset of $\mathbf{S}$, which satisfies the condition: for every $(index(1), ..., index(n)) \in \mathbf{S}_1$, its permutations do not belong to $\mathbf{S}_1$. Thus, we have:

$$\begin{aligned}
\hat{\mu}^B &= \{\sum_A P(\mathbf{A})\frac{\mathbf{1}^\top \mathbf{\Sigma}_\mathsf{A}^{-1}}{\mathbf{1}^\top \mathbf{\Sigma}_\mathsf{A}^{-1}\mathbf{1} + \frac{1}{\nu}}\}\mathbf{Y} \\
&= \sum P(\mathbf{A})\{\frac{(\frac{1}{ctr_{index(1)}+1}, \frac{1}{ctr_{index(2)}+1}, ..., \frac{1}{ctr_{index(n)}+1})}{\sum_{i=1}^{n}\frac{1}{ctr_{index(i)}+1}+\frac{1}{\nu}}\}\mathbf{Y} \\
&= \{\sum_{(index(1),...,index(n))\in \mathsf{S}_1} P(\mathbf{A})(\#\ of\ permutations) \times \frac{f}{f+\frac{1}{\nu}}\}\overline{\mathbf{Y}} \\
&= \{\sum_{(index(1),...,index(n))\in \mathsf{S}} P(\mathbf{A})\frac{\sum_{i=1}^{n}\frac{1}{ctr_{index(i)}+1}}{\sum_{i=1}^{n}\frac{1}{ctr_{index(i)}+1}+\frac{1}{\nu}}\}\overline{\mathbf{Y}} \\
&= l\overline{\mathbf{Y}}.
\end{aligned}$$

Thus, $\hat{\mu}^B = l\overline{\mathbf{Y}}$. In addition, $l < 1$.

Similarly,

$$\hat{\mu}^l = \frac{\mathbf{1}^\top \mathbf{\Sigma}_\mathsf{I}^{-1}}{\mathbf{1}^\top \mathbf{\Sigma}_\mathsf{I}^{-1}\mathbf{1} + \frac{1}{\nu}}\mathbf{Y} = \frac{\underline{1}^\top - (\frac{\mathsf{ct}}{\mathsf{ct}+1}, \frac{\mathsf{ct}}{\mathsf{ct}1+1}, ..., \frac{\mathsf{ct}}{\mathsf{ct}1+1})}{\mathbf{n} - \sum_{\mathsf{i}=1}^{\mathsf{n}}\frac{\mathsf{ct}}{\mathsf{ct}+1} + \frac{1}{\nu}}\mathbf{Y} = l_l\overline{\mathbf{Y}}, \tag{D–2}$$

where $l_l = \frac{\sum_{i=1}^{n}\frac{1}{ct+1}}{\sum_{i=1}^{n}\frac{1}{ct+1}+\frac{1}{\nu}}$.

Since $f = n - \sum_{i=1}^{n}\frac{ctr_{index(i)}}{ctr_{index(i)}+1} \leq \sum_{i=1}^{n}\frac{1}{ct+1}$, $l \leq l_l$.

Since $\hat{\mu}^{BF} = \sum P(\mathbf{A})(\mathbf{X}^\top \mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{\Sigma}^{-1}\mathbf{Y}$ and $\hat{\mu}^{IF} = (\mathbf{X}^\top \mathbf{\Sigma}_{\mathrm{I}}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{\Sigma}_{\mathrm{I}}^{-1}\mathbf{Y}$, by the similar calculation for $\hat{\mu}^{B}$ and $\hat{\mu}^{I}$ we can also get

$$\hat{\mu}^{BF} = \hat{\mu}^{IF} = \overline{\mathbf{Y}},$$

where they are corresponding to $\nu = \infty$ in Eq. (D–1) and (D–2). □

APPENDIX E
PROOF OF THEOREM 15

Proof. Since

$$\mathrm{Var}(l\overline{\mathbf{Y}}|\text{Dirichlet model}) = l^2 \sum_{\mathsf{A}} P(\mathbf{A}) \frac{n\sigma^2 + \sigma^2 \sum_{i=1}^{n} ctr_{index(i)}}{n^2},$$

by the expression of l, we have:

$$\mathrm{Var}(l\overline{\mathbf{Y}}|\text{Dirichlet model}) \geq \sum_{\mathsf{A}} P(\mathbf{A}) \frac{\sigma^2}{n} \sum_{i=1}^{n} (ctr_{index(i)} + 1) [\frac{\sum_{i=1}^{n} \frac{1}{ctr_{index(i)}+1}}{\frac{1}{\nu} + \sum_{i=1}^{n} \frac{1}{ctr_{index(i)}+1}}]^2.$$

Consider a function $g_A(x_1, ..., x_n) = \sum_{i=1}^{n} x_i [\frac{\sum_{i=1}^{n} \frac{1}{x_i}}{\frac{1}{\nu} + \sum_{i=1}^{n} \frac{1}{x_i}}]^2$, where $ct + 1 \leq x_i \leq ctr + 1$. Then when $\nu \geq \frac{ctr+1}{tr}$, for every j,

$$\begin{aligned}
\frac{\partial g_A}{\partial x_j} &= \frac{\sum_{i=1}^{n} \frac{1}{x_i}}{(\frac{1}{\nu} + \sum_{i=1}^{n} \frac{1}{x_i})^2} [\sum_{i=1}^{n} \frac{1}{x_i} - 2(\sum x_i) \frac{\frac{1}{\nu} \frac{1}{x_j^2}}{\frac{1}{\nu} + \sum_{i=1}^{n} \frac{1}{x_i}}] \\
&\geq \frac{\sum_{i=1}^{n} \frac{1}{x_i}}{(\frac{1}{\nu} + \sum_{i=1}^{n} \frac{1}{x_i})^2} [\sum_{i=1}^{n} \frac{1}{x_i} - 2(\sum x_i) \frac{\frac{tr}{ctr+1} \frac{1}{(ctr+1)^2}}{\frac{2tr}{ctr+1}}] \\
&\geq 0,
\end{aligned}$$

which means that g is a increasing function in x_j.

Thus,

$$\begin{aligned}
\mathrm{Var}(l\overline{\mathbf{Y}}|\text{Dirichlet model}) &\geq \sum_{\mathsf{A}} P(\mathbf{A}) \frac{\sigma^2}{n} g_A(r_{index(1)}, .., r_{index(n)}) \\
&\geq \sum_{\mathsf{A}} P(\mathbf{A}) \frac{\sigma^2}{n} g_A(ct + 1, ..., ct + 1) \\
&= \mathrm{Var}(l\overline{\mathbf{Y}}|\text{normal model}).
\end{aligned}$$

Thus, when $\nu \geq \frac{2(ctr+1)}{tr}$, we have $\mathrm{Var}(l\overline{\mathbf{Y}}|\text{Dirichlet model}) \geq \mathrm{Var}(l_l\overline{\mathbf{Y}}|\text{normal model})$.

Since $l \leq l_l$, we have $MSE(\hat{\mu}^B) \geq MSE(\hat{\mu}^l)$. □

APPENDIX F
PROOF OF THEOREM 16

Proof. Since $l \leq l_l$, then we always have $\mathrm{Var}(\hat{\mu}^B|\text{true model}) \leq \mathrm{Var}(\hat{\mu}^l|\text{true model})$, no matter what the true model is.

Part 1: Assume the true model is the Dirichlet model.

For $\hat{\mu}^B$ and $\hat{\mu}^l$, the Bayes risks have the form

$$\text{Bayes Risk} = \int (g^2\mathrm{Var}(\overline{\mathbf{Y}}) + (1-g)^2\mu^2)d\Lambda(\mu) = (\mathrm{Var}(\overline{\mathbf{Y}}) + \nu\sigma^2)g^2 - 2\nu\sigma^2 g + \nu\sigma^2,$$

which can be considered as a quadratic function of g with the axis of symmetry $g_0 = \frac{\nu\sigma^2}{\nu\sigma^2+\mathrm{Var}(\overline{\mathrm{Y}})}$. For $\hat{\mu}^B$, $g = l$ and for $\hat{\mu}^l$, $g = l_l$.

By the proof of the Theorem 18, we know that $\frac{\mathrm{Var}(\overline{\mathrm{Y}})}{\sigma^2} \geq \sum P(\mathbf{A})\frac{1}{1^\top\Sigma_\mathrm{A}^{-1}1}$ under the Dirichlet model.

Then by Jensen's inequality, we have

$$\begin{aligned} g_0 = \frac{\nu\sigma^2}{\nu\sigma^2 + \mathrm{Var}(\overline{\mathbf{Y}})} &\leq \frac{1}{1 + \frac{1}{\nu}\sum P(\mathbf{A})\frac{1}{1^\top\Sigma_\mathrm{A}^{-1}1}} \\ &\leq \sum P(\mathbf{A})\frac{1}{1 + \frac{1}{\nu}\frac{1}{1^\top\Sigma_\mathrm{A}^{-1}1}} \\ &= l \leq l_l. \end{aligned}$$

Thus when the true model is the Dirichlet model, $\text{Bayes Risk}(\hat{\mu}^B) \leq \text{Bayes Risk}(\hat{\mu}^l)$.

In addition, $\delta_D = \text{Bayes Risk}(\hat{\mu}^l) - \text{Bayes Risk}(\hat{\mu}^B) = (\mathrm{Var}(\overline{\mathbf{Y}}) + \nu\sigma^2)(l_l - l)(l_l + l - 2\frac{\nu\sigma^2}{\mathrm{Var}(\overline{\mathrm{Y}})+\nu\sigma^2})$.

Part 2: Assume the true model is the normal model. By the similar argument,

$$g_0 = \frac{\nu\sigma^2}{\nu\sigma^2 + \mathrm{Var}(\overline{\mathbf{Y}})} = \frac{1}{1 + \frac{1}{\nu}\frac{1+tc}{n}} = l_l \geq l.$$

Thus when the true model is the normal model, $\text{Bayes Risk}(\hat{\mu}^B) \geq \text{Bayes Risk}(\hat{\mu}^l)$.

In addition, $\delta_N = \text{Bayes Risk}(\hat{\mu}^B) - \text{Bayes Risk}(\hat{\mu}^l) = (\mathrm{Var}(\overline{\mathbf{Y}}) + \nu\sigma^2)(l_l - l)(2\frac{\nu\sigma^2}{\mathrm{Var}(\overline{\mathrm{Y}})+\nu\sigma^2} - l_l - l)$.

Part 3.

By the above calculation, we have:

$$
\begin{aligned}
\delta_D - \delta_N &= 4\nu\sigma^2(l_l - l) \\
&\times [(l + l_l)(\frac{2\nu\sigma^2 + \mathsf{Var}(\overline{\mathbf{Y}}|\text{Dirichlet model}) + \mathsf{Var}(\overline{\mathbf{Y}}|\text{normal model}))}{4\nu\sigma^2} - 1] \\
&= 4\nu\sigma^2(l_l - l) \\
&\times [(l + l_l)\frac{\nu\sigma^2 + \mathsf{Var}(\overline{\mathbf{Y}}|\text{Dirichlet model}) + \nu\sigma^2 + \mathsf{Var}(\overline{\mathbf{Y}}|\text{normal model})}{4\nu\sigma^2} - 1] \\
&\geq 4\nu\sigma^2(l_l - l)[(l + l_l)(\frac{1}{4}\frac{1}{l} + \frac{1}{4}\frac{1}{l_l}) - 1] \\
&= 4\nu\sigma^2(l_l - l)[\frac{1}{4}\frac{l_l}{l} + \frac{1}{4}\frac{l}{l_l} - \frac{1}{2}] \\
&\geq 0.
\end{aligned}
$$

Thus $\delta_D \geq \delta_N$. □

REFERENCES

AFSHARTOUS, D. and WOLF, M. (2007). Avoiding 'data snooping' in multilevel and mixed effects models. *J. R. Statist. Soc. A*, **170** 1035–1059.

ANTONIAK, C. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *Annals of Statistics*, **2** 1157–1174.

BLACKWELL, D. and MACQUEEN, J. (1973). Ferguson distributions via polya urn schemes. *Annals of Statistics*, **1** 353–355.

BROWN, K. G. (1976). Asymptotic behavior of minque-type estimations of variance components. *The Annals of Statistics*, **4** 746–754.

BURR, D. and DOSS, H. (2005). A bayesian semi-parametric model for random effects metaanalysis. *J. Amer. Statist. Assoc.*, **100** 242–251.

CHAUBEY, Y. P. (1980). Application of the method of minque for estimation in regression with intraclass covariance matrix. *Sankhya: The Indian Journal of Statistics*, **42** 28–32.

DAS, K., JIANG, J. and RAO, J. (2004). Mean squared error of empirical predictor. *The Annals of Statistics*, **32** 818–840.

DAY, N. E. (1969). Estimating the components of a mixtures of normal distributions. *Biometrika*, **56** 463.

DEMIDENKO, E. (2004). *Mixed Models: Theory and Applications*. Wiley.

ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90** 577–588.

FERGUSON, T. (1973). A bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1** 209–230.

GILL, J. and CASELLA, G. (2009). Nonparametric priors for ordinal bayesian social science models: Specification and estimation. *J. Amer. Statist. Assoc.*, **104** 453–464.

HARVILLE, D. (1976). Extension of the gauss-markov theorem to include the estimation of random effects. *The Annals of Statistics*, **4** 384–395.

HARVILLE, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *ournal of the American Statistical Association*, **72** 320–338.

HUANG, S. and LU, H. H. (2001). Extended gauss-markov theorem for nonparametric mixed-effects models. *Journal of Multivariate Analysis*, **76** 249–266.

JIANG (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. Springer.

KACKAR, R. and HARVILLE, D. (1981). Unbiasedness of two-stage estimation and prediction procedures for mixed linear models. *Comm. Statist. A-Theory Methods*, **10** 1249–1261.

KACKAR, R. and HARVILLE, D. (1984). Approximations for standard errors of estimators of fixed and random effect in mixed. *Journal of the American Statistical Association*, **79** 209–230.

KORWAR, R. M. and HOLLANDER, M. (1973). Contributions to the theory of dirichlet processes. *Ann. Probab.*, **1** 705–711.

KYUNG, M., GILL, J. and CASELLA, G. (2009). Characterizing the variance improvement in linear dirichlet random effects models. *Statistics and Probability Letters*, **79** 2343–2350.

KYUNG, M., GILL, J. and CASELLA, G. (2010). Estimation in dirichlet random effects models. *The Annals of Statistics*, **38** 979–1009.

LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*. Springer.

LIU, J. S. (1996). Nonparametric hierarchical bayes via sequential imputations. *The Annals of Statistics*, **24** 911–930.

LO, A. Y. (1984). On a class of bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics*, **12** 351–357.

MACEACHERN, S. N. and MULLER, R. (1998). Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, **7** 223–238.

MATHEW, T. (1984). On nonnegative quadratic unbiased estimability of variance componenes. *The Annals of Statistics*, **12** 1566–1569.

MCCULLOCH, C. E. and SEARLE, S. R. (2001). *Generalized, Linear and Mixed Models*. Wiley-Interscience.

NEAL, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **9** 249–265.

PETERS, B. and WALKER, H. F. (1978). An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions. *SIAM J. of Appl. Math.*, **35** 362–378.

PUKELSHEIM, F. (1981). On the existence of unbiased nonnegative estimates of variance convariance components. *The Annals of Statistics*, **9** 293–299.

PUNTANEN, S. and STYAN, G. (1989). The equality of the ordinary least squares estimator and the best linear unbiased estimator. *The American Statistician*, **43** 153–161.

RAO, C. R. (1979). Minqe theory and its relation to ml and mml estimation of variance components. *Sankhya: The Indian Journal of Statistics*, **41** 138–153.

RAO, P. S. R. S. (1977). Theory of the minque: A review. *Sankhya: The Indian Journal of Statistics*, **39** 201–210.

RAO, R. S. R. S. and CHAUBEY, V. P. (1978). Three modifications of the principle of the minque. *Comm. Statist. Meth.*, **A7**.

ROBINSON, G. (1991). That blup is a good thing: the estimation of random effects. *Statistical Science*, **6** 15–51.

SEARLE, S., CASELLA, G. and MCCULLOCH, C. (2006). *Variance Components*. John Wiley and Sons, New Jersey.

SETHURAMAN, J. (1994). A constructive definition of dirichlet priors. *Statistical Sinica*, **4** 638–650.

XU, L. and JORDAN, M. I. (1996). On convergence properties of the em algorithm for gaussian mixtures. *Neural Computation*, **8** 129–151.

YOUNG, T. Y. and CORALUPPI, G. (1969). On estimation of a mixture of normal density functions. *Adaptive Processes (8th) Decision and Control, 1969 IEEE Symposium on* 63–63.

ZYSKIND, G. (1967). On canonical forms non-negative covariance matrices and best and simple least squares linear estimators in linear models. *Ann. Math. Statist*, **38** 1092–1109.

ZYSKIND, G. and MARTIN, F. B. (1969). On best linear estimation and a general gauss-markov theorem in linear models with arbitrary nonnegative covariance structure. *SIAM Journal on Applied Mathematics*, **17** 1190–1202.

BIOGRAPHICAL SKETCH

Chen Li was born in P. R. China. She received her bachelor's degree in applied mathematics from Tongji University in 2002 at Shanghai, P. R. China. She then completed her PhD in Applied Mathematics at Tongji University in 2007. The research on applied mathematics focused on the numerical methods for partial differential equations. After that, she joined the Department of Statistics at the University of Florida. She received the PhD in Statistics in 2012. The research on statistics was to discuss the linear mixed model with the Dirichlet Process random effects from both the frequentist way and the Bayesian way.