Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Tielin Qin Date

Bayesian Analysis for Repeated Compositional Data and Approaches for Correcting Measurement Errors in General Multivariate Linear Model

By

Tielin Qin
Doctor of Philosophy
Biostatistics

Vicki S. Hertzberg, Ph.D.
Advisor

Qi Long, Ph.D.
Committee Member

Robert H. Lyles, Ph.D.
Committee Member

Aneesh K. Mehta, M.D.
Committee Member

Tianwei Yu, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Bayesian Analysis for Repeated Compositional Data and Approaches for Correcting Measurement Errors in General Multivariate Linear Model

By

Tielin Qin
B.S., Beijing Medical University, 1996
MSPH, Emory University, 2005

Advisor: Vicki S. Hertzberg, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2011

Abstract

Bayesian Analysis for Repeated Compositional Data and Approaches for Correcting Measurement Errors in General Multivariate Linear Model

By

Tielin Qin

Compositional data can be viewed as the positive vectors whose components are the proportion or percentage of whole. In this dissertation, we are motivated by the need to examine the different subpopulation of white blood cells in the Protective Immunity Project (PIP) study conducted at the Emory Transplant Center. The first research question is how to modeling the white cell compositions over time. The data obtained from this study is the compositional data with repeated measurements. We develop a Bayesian approach for the analysis of the repeat-measured compositional data. Our results have been demonstrated that the Bayesian methodology can be used to analyzed repeat-measured compositional data. We use MCMC for model inference and show that the method is practical in high dimensional problems.

Another research question motivated from the PIP study is how to get the correct estimates when the measurement errors exist on the total cell count data. In the medical studies, some variables of interest are difficult to obtain, and surrogate variables are used instead. However, these surrogate variables may contain measurement errors. We propose the likelihood-based estimators for general multivariate linear model when the non-linear measurement errors exist in the response variables. The observed response variables are related to the true values through a non-linear regression model, and the parameters in the measurement error model are estimated by using independent, external calibration data. The pseudo-MLE is used for model inference to avoid computational problems. Our proposed models provide a tool to correct for measurement errors in response variables in longitudinal data.

Finally, we propose a Bayesian approach for correcting the measurement error in the general multivariate linear model when the non-linear measurement errors exist in the response variables. We outline how the estimations of the parameters of interest can be carried out in a Bayesian framework using Gibbs sampling and Metropolis Algorithm. In the Bayesian approach, we impute the values of the unobservable variable Y by sampling from their conditional distribution given all the observed data and other parameters. Therefore, using Bayesian approach can avoid numerical integrations which may be tedious and extensive.

Bayesian Analysis for Repeated Compositional Data and Approaches for Correcting Measurement Errors in General Multivariate Linear Model

By

Tielin Qin
B.S., Beijing Medical University, 1996
MSPH, Emory University, 2005

Advisor: Vicki S. Hertzberg, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2011

Acknowledgements

I first would like to thank my advisor Dr. Vicki S. Hertzberg for her patient and valuable guidance and support through my research journey. Without her constant help, I would not to reach where I am now. I would like also thank my committee members Drs. Aneesh Mehta, Qi Long, Robert Lyles and Tianwei Yu for their constructive comments and suggestions. I would also like to thank all the faculty, staff, and students in the Department of Biostatistics and Bioinformatics for their help through those years.

I want to express my special thanks to my wife Lihui Liang for her unselfish support during my study at Emory University.

Finally , I would like to deeply thank my parents for their unconditional love, support and encouragement.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Overview

Flow cytometry is a powerful technique for analyzing multiple parameters of cells within heterogeneous populations, and can conduct multi-parametric analysis for thousands of cells per second simultaneously. Flow cytometry can be used in the diagnosis of diseases, especially blood cancers, and many other fields in both research and clinical practice. This dissertation is motivated by a real study in which data were obtained from flow cytometry.

Our motivation is the Protective Immunity Project (PIP) conducted at the Emory Transplant Center(Larsen and Ahmed [2005]). In the PIP study, the investigators enrolled 60 patients aged 18-59 years old who were scheduled to undergo renal transplantation at Emory University transplant center. They also enrolled 20 age-, sex- and race-matched healthy volunteers into the control group. All subjects enrolled in this study were followed for two years, and multiple blood samples were collected at baseline, 3 months, 6 months, 9 months, 12 months, 18 months and 24 months. The blood samples were analyzed with flow cytometry. There are two research questions motivated by the PIP study. The first research question motivated by the PIP study is how to model the white cell compositions over time. Since the PIP study is a longitudinal study in which every enrolled subject was followed for two years, the data obtained from this study is compositional in nature with repeated measurements. In the PIP study, the data sometimes only come as a set of percentages that sum to 100%. No total cell counts were obtained in this situation. Therefore, the standard analysis, such as multinomial model, is not appropriate if the total count is not available. For the percentage data without total counts available, we define as the compositional data. We are motivated by this type of data in the PIP study, and develop a Bayesian approach for repeated compositional data in chapter 3. For the

compositional data, the summation of all components equals to 1. Because of this constriction, standard statistical methods are not appropriate for the compositional data analysis. Another research question motivated in part from the PIP study is how to get the correct estimates when measurement errors exist on the white blood cell count data. In the PIP study, the variables of interest are the absolute counts of sub-categories of lymphocytes. However, the recorded data obtained from flow cytometer may contain measurement errors.

1.2 An Introduction of Flow Cytometry

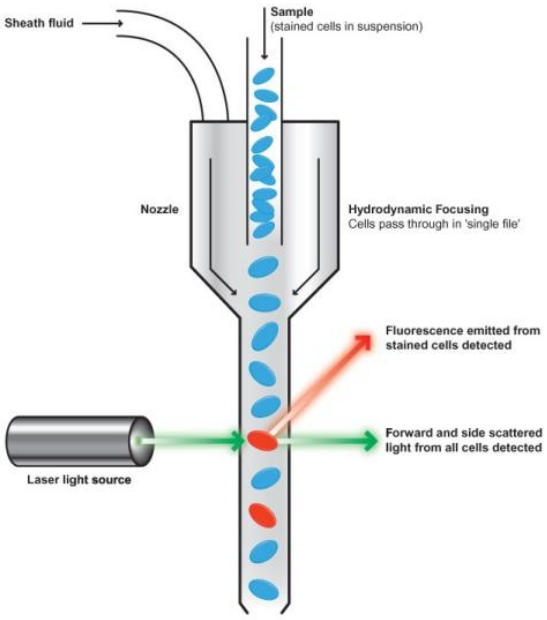


Figure 1.1: Flow Cytometry System

Flow cytometry is a modern technique which can be used for the analysis of multiple parameters of microscopic particles, for example, cells and chromosomes. The applications of flow cytometry range from immunophenotyping, to ploidy analysis,

to cell counting. Figure 1.1 shows how the flow cytometry works. First, the sample cells are suspended in the fluid and then form a stream of fluid. The stream passes through an electronic detection apparatus. A laser source emits laser beams directed onto the stream of fluid. As a cell passes through the laser light, the laser will be scattered. The scattered leaser lights are recorded by a number of detectors. One detector records the scattered light which has the same direction with the light beam (Forward Scatter or FSC) and several detectors record the scattered light perpendicular to it (Side Scatter or SSC) and one or more detectors record fluorescent light. The light signals recorded by the detectors are processed by a computer connected with the flow cytometer. The data generated by flow-cytometer can be plotted. Based on the light intensity, some distinct populations can be separated in these plots by creating a series of subset extractions. The process to identify the distinct subcategories is called "Gating". Specific gating protocols had been developed for diagnostic and clinical purposes especially in relation to hematology. Figure 1.2 shows two dot plots from some four parameter data derived from human peripheral blood leucocytes. From the left plot, we can see that the side scattered light vs. forward scattered light can define three distinct populations; these are the granulocytes, monocytes and lymphocytes, labeled G, M and L respectively. In the right plot, the cells can be defined based on the anti-CD4-PE and anti-CD8-FITC expression(Ormerod [2000, 2008]).

1.3 The Motivation Example

Compositional data can be viewed as the positive vectors whose components are the proportion or percentage of whole. The constraint is that their sum should be some constant, for example, equal to 1 for proportion, 100 for percentage. A typical example of compositional data is geochemical compositions of rocks. Commonly the

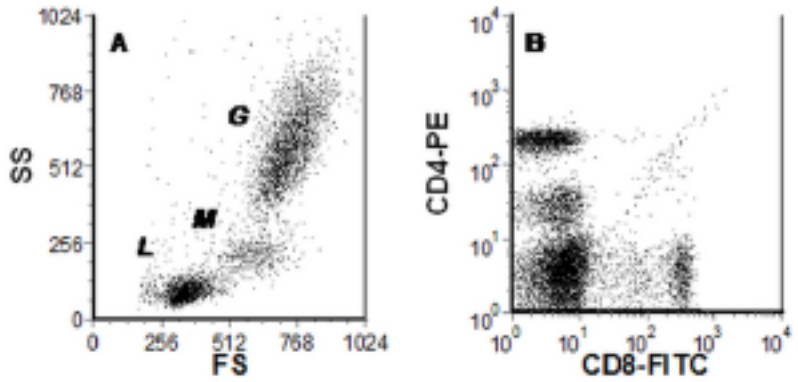


Figure 1.2: Figure shows two dot plots from data derived from human blood leucocytes. A: The light scatter (SS versus FS) defines three distinct populations; these are the granulocytes, monocytes and lymphocytes, labelled G, M and L. B: The cells were labelled with anti-CD4-PE and anti-CD8-FITC, both proteins are expressed on T lymphocytes.

mineral compositions of rocks are expressed as percentage by weight. One may want to describe the variation of compositions from specimen to specimen, and determine if the compositions of one kind of rock differ from another kind of rock. In economics study, an important aspect of the study of consumer demand is the analysis of household budget. In this case, the compositions of household expenditures, that is, the proportion of total expenditures to different commodity groups are critical to household budget analysis. One may want to examine if the compositions of expenditures depend on the total amount spent, or if there are differences between high-income and low-income household in their expenditure compositions.

In this paper, we are motivated by the need to examine the different subpopulation of white blood cells in the Protective Immunity Project (PIP) conducted at the Emory Transplant Center (Larsen and Ahmed [2005]). The PIP study contains three complementary studies. The goal of the first study is to characterize the impact of immunosuppression regimens on protective immunity over time in renal

transplant patients. This is an ongoing study began in 2005, and ended in 2011. The investigators enrolled 60 patients aged 18-59 years old who had renal transplantation at Emory University transplant center. They also enrolled 20 age-, sex- and race-matched healthy volunteers into the control group. All subjects enrolled in this study were followed for two years, and multiple blood samples were collected at baseline, 3 months, 6 months, 9 months, 12 months, 18 months and 24 months. The blood samples were analyzed with flow cytometry. Whole blood was passed through the flow cytometer to determine blood composition. Lymphocytes were isolated and the percentage of CD3+ lymphocytes, also called T lymphocytes, was recorded. The CD3+ cells can also be broken into four subcategories based on different cell surface markers: CD4+CD8- (T Helper cells), CD4-CD8+ (cytotoxic T cells), CD4-CD8-, and CD4+CD8+. The T helper cells and the cytotoxic T cells can also be further broken down based on the chosen surface markers. The T helper cells can be further broken into four categories: CCR7+CD45RA- (central memory), CCR7+CD45RA+ (naive), CCR7-CD45RA- (effector memory), and CCR7-CD45RA+. The percentages were recorded by flow cytometer. The cytotoxic T cells can also be further broken into 4 categories: CCR7+CD45RA- (central memory), CCR7+CD56RA+ (naive) CCR7-CD45RA- (effector memory), and CCR7-CD45RA+ (effector memory RA). Figure 1.3 shows the hierarchy of blood cell types in PIP study. In this study, we want to know the impact of immunosuppression regimens on lymphocyte compositions in renal transplant patients with respect to how cell compositions change over time.

There are two research questions motivated by the PIP study. The first research question motivated by the PIP study is how to model the white blood cell compositions over time. Since the PIP study is a follow-up study in which every enrolled subject was followed for two years, the data obtained from this study is the composi-

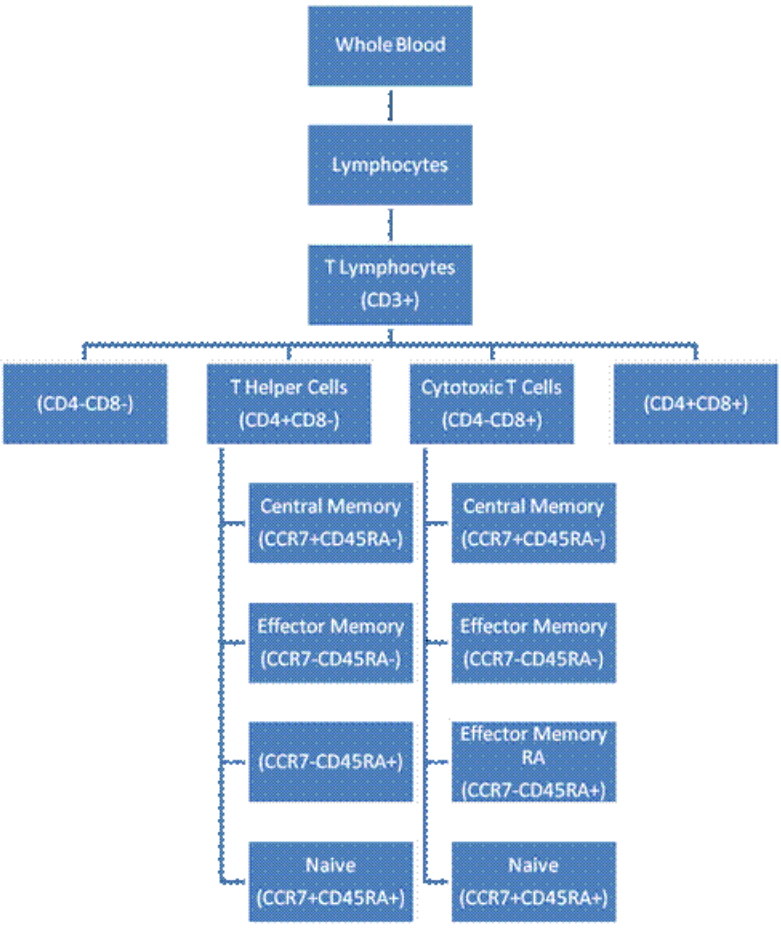


Figure 1.3: Hierarchy of blood cell types in PIP study

tional data with repeated measurements. The data only contains the percentages of cell subcategories. No total cell counts is obtained in this situation. Since the repeat measurements made on the same individual are typically correlated, special attention is needed when we analyze the repeated compositional data. The statistical model for this kind of data will be discussed in chapter 3.

Another research question motivated in part from the PIP study is how to get the correct estimates when measurement error exists with respect to the white blood cell count data. In medical studies, some variables of interest may be difficult to obtain, and surrogate variables are recorded and used instead. However, these surrogate

variables may contain measurement errors. In the PIP study, the whole blood samples were passed through the flow cytometer to determine blood composition. The counts and percentage of subcategories of lymphocytes were recorded based on the cell surface markers. The variables of interest are the true counts of subcategories of lymphocytes. However, the recorded data obtained from flow cytometer may contain measurement errors. We propose a likelihood based method for correcting measurement errors for a general multivariate linear model in chapter 4. We also propose a Bayesian approach for correcting measurement errors for a general multivariate linear model in chapter 5.

1.4 Proposed Research

In this dissertation, the goal is to solve two research questions raised from the PIP study. The first is how to analyze the compositional data with repeated measurements. The second research question is how to correct the measurement errors in the dependent variables in the general multivariate linear model.

For the first research question, we propose a Bayesian approach for repeated compositional data. We use a multivariate logistic normal model, and use MCMC approach for model inference. Under the logistic normal distribution, and by relying on the additive logratio transformation, we transform the compositional data into multivariate normal distributed data. Because of the non-linear additive logratio transformation, we use a Bayesian approach for model inference.

For the measurement errors in the dependent variables, we propose two approaches for correcting measurement errors in the dependent variables in the general multivariate linear model. The parameters in the measurement error model are estimated by using independent, external calibration data. We use the likelihood-based method to

correct measurement errors. The pseudo-maximum likelihood estimators and their asymptotic properties are developed. A simulation study is conducted to compare the performance of the pseudo MLE approach and the simply adjusted/imputed data approach.

In this dissertation, we also propose a Bayesian approach to adjust the measurement errors in the response variables in the general multivariate linear model. In the Bayesian approach, the unobserved true variable Y is treated as missing data, and can be imputed by drawing from the full conditional distribution of Y given all other parameters and data. Markov Chain Monte Carlo (MCMC) methods are used to compute Bayesian quantities.

Chapter 2

BACKGROUND

2.1 An Introduction of Composition Data Analysis

To begin with compositional data analysis, we need to give a formal definition of compositional data.

Definition 1 *Composition (Aitchison [1986])*
A composition x of K-parts is a $k \times 1$ *vector with positive components* $x_1, \ldots, x_k$ *whose sum is 1 .*

2.1.1 Ternary diagrams to display compositional data

A ternary diagram is a convenient way to display 3-part compositional data.

In Figure 2.1, the equilateral triangle with vertices 1, 2, 3 has unit height. For any point in this triangle, the perpendiculars x_1, x_2, x_3 from the point to the sides

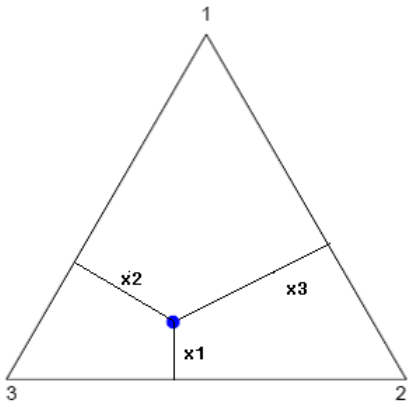


Figure 2.1: Graphical display of a three-part compositions in a Ternary diagram

opposite 1, 2, 3 satisfy

$$x_i \geq 0 (i = 1, 2, 3), x_1 + x_2 + x_3 = 1$$

For any vector (x_1, x_2, x_3) satisfy above equation, there is a unique point in triangle 123. The relationship between 3-part compositions and points in the triangle is one-to-one. The larger a component x_i is, the further the point is away from the opposite side, and then the closer the point is to the vertex i.

2.1.2 Algebra for Compositions

The first task for modeling compositional data is to define a suitable sample space. If our concern is specifying a density function on the sample space, then we need to emphasize the dimensionality of the composition. This concern leads to the definition of the simplex sample space.

Definition 2 *Simplex sample space (Aitchison [1986])*
The (k-1)-dimensional simplex is the set defined by

$$S^{k-1} = \{(x_1, \ldots, x_k) : x_1 > 0, \ldots, x_k > 0; x_1 + \ldots + x_k = 1\}$$

Compositional data are positive vectors with constant sum, and describe the relative proportion of each component in k categories. Suppose we have a K-part composition, $\underline{z} = (z_1, z_2, \ldots, z_k)$, where the z_i $(i = 1, \ldots, K)$ are the components, proportions of the available unit, and $z_i > 0$ for all $i = 1, \ldots, K$, and $\sum_{i=1}^{k} z_i = 1$. The sample space associated with K-part compositions is the (k-1) dimensional unit simplex (∇^{k-1}). Aitchison [1982, 1986] introduced the Logistic Normal distribution into compositional data analysis. He used the additive logratio transformation to take the compositional data from the (k-1)-dimensional simplex (∇^{k-1}) to (k-1)-dimensional Euclidean space (R^{k-1}).

Perturbation operator for compositional data

Aitchison [1986] defined the perturbation operator in simplex sample space. He described a simple motivating example to introduce this operator. Suppose we have two compositions $\underline{x}$ and $\underline{X}$ on a similar, but differential scaling relationship $X_1 = p_1 x_1, \ldots, X_k = p_k x_k$ reflects the composition change from $\underline{x}$ to $\underline{X}$. Such a unique differential scaling can always be found by taking $p_i = X_i / x_i$ $(i = 1, \ldots, K)$. The relative amounts of the components of $\underline{X}$ can be expressed as $x_1 \frac{p_1}{c}, \ldots, x_k \frac{p_k}{c}$, where $c = \sum_{i=1}^{k} x_i p_i$. Each p_i must be between zero and one. One property of this operator is that only the relative size of the components of $\underline{p}$ affects the resulting composition $\underline{X}$. Therefore, we can consider $\underline{x}$ and $\underline{p}$ are both proportion vectors, and the "addition" of $\underline{x}$ and $\underline{p}$ is also a composition.

Before we define perturbation operator, we need to define the closure operation first.

Definition 3 *Closure operation (C) (Aitchison [1986])*
Suppose $\underline{p}$ *is a k-dimensional vector in positive Euclidean space* (R_+^k). *The closure operation* $C(\underline{p})$ *is defined as*

$$[C(p)]_i = \frac{p_i}{\sum_{j=1}^{k} p_i}$$

where $[C(p)]_i$ *denotes the* i^{th} *element of the k-vector.*

Definition 4 *Perturbation Operation (Aitchison [1986])*
Let $\underline{x}$ *be a k-part composition and* $\underline{u}$ *be a k-part with positive elements. Then the operation*

$$\underline{X} = \underline{u} \oplus \underline{x} = C(x_1 u_1, \ldots, x_k u_k)$$

is defined as a perturbation with the original composition $\underline{x}$ *being operated on the perturbing* $\underline{u}$ *to form a perturbed composition* $\underline{X}$. *The perturbation operator can be considered as an "addition" operator for compositional data.*

Aitchison [1986] showed several simple properties of perturbation operation.

Property 1 *The operation* $\underline{u}\oplus$ *is a a one-to-one transformation between* (∇^{k-1}) *and* (∇^{k-1}). *The inverse transformation is the perturbation* $\underline{u}^{-1}\oplus$, *where*

$$\underline{u}^{-1} = \left(\frac{1}{u_1}, \frac{1}{u_2}, \ldots, \frac{1}{u_k}\right)$$

Property 2 *Since* $\underline{u} \oplus \underline{x} = C(\underline{u}) \oplus \underline{x}$*, the effect of any perturbing vector* $\underline{u} \in R_+^{k-1}$ *is the same as that of the perturbing vector* $C(\underline{u}) \in \nabla^{k-1}$*. Therefore, we can restrict the perturbing vectors to the simplex* ∇^{k-1} *without loss of any generality.*

Property 3 *The operation* $\oplus \underline{I}_{k-1}$ *is the identity operator for any* $\underline{u} \in \nabla^{k-1}$*, where* $\underline{I}_{k-1} = \left(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}\right)$*. We have* $\underline{u} \oplus \underline{I}_{k-1} = \underline{u}$ *for any* $\underline{u}$*.*

Property 4 *The operation* $\oplus$ *is commutative. For* $\underline{u}$ *and* $\underline{a}$ *in* ∇^{k-1}*,*

$$\underline{u} \oplus \underline{a} = \underline{a} \oplus \underline{u}$$

Property 5 *The operation* $\oplus$ *is associative.*

$$\underline{u}_2 \oplus (\underline{u}_1 \oplus \underline{x}) = \underline{u}_1 \oplus (\underline{u}_2 \oplus \underline{x})$$

where $\underline{u}_1, \underline{u}_2$ *are any perturbing vectors.*

Scalar Multiplication for compositional data

Definition 5 *Scalar multiplication (Aitchison [1986])*
Define scalar multiplication of a composition $\underline{u}$ *by a in the following way*

$$\underline{u}^a = C(u_1^a, u_2^a, \ldots, u_k^a)$$

where $a \in R$ *be any scalar, and* $\underline{u} \in \nabla^{k-1}$ *is a k-component composition.*
This defines a "multiplication" operator in simplex sample space.

Definition 6 *Inner product for compositions (Billheimer and Guttorp [1995])*
For $\underline{u}$, $\underline{z} \in \nabla^{k-1}$, *let* $\underline{\theta} = \phi(\underline{u})$, *and* $\underline{\eta} = \phi(\underline{z})$. *Define*

$$\langle \underline{u}, \underline{z} \rangle = \underline{\theta}' N^{-1} \underline{\eta}$$

as the inner product of $\underline{u}$ *and* $\underline{z}$.
We define $N = [I_{k-1} + j_{k-1} j'_{k-1}]$, *where* I_{k-1} *is a (k-1)-dimensional identity matrix, and* j_{k-1} *is a (k-1) column vector of 1's, and* $N^{-1} = [I_{k-1} - \frac{1}{k} j_{k-1} j'_{k-1}]$.

Definition 7 *Norm on* $(k-1)$ *dimensional Simplex sample space* ∇^{k-1}. *(Aitchison [1986])*
Define $||u||$, *the norm of composition* $\underline{u}$, *for* $\underline{u} \in \nabla^{k-1}$ *as* $\langle \underline{u}, \underline{u} \rangle^{1/2}$.

Aitchison [1986] showed that the inner product and norm are invariant to permutations of components of $\underline{u}$. The definition of the matrix N^{-1} ensures this invariance. Also note that the norm is a sum of squares of log-ratios. The norm defined above measures the distance of a composition from I_{k-1}, the "center" of ∇^{k-1}.

2.1.3 Logistic Normal Distribution

The first direction to analyze compositional data is to use the Dirichlet distribution, since it is the most familiar class of distributions on the simplex sample space. However, (Aitchison [1982]) has pointed out, it is inadequate for the description of variability in compositional data. One major disadvantage with the Dirichlet distribution is that the correlation structure of a Dirchlet composition is wholly negative,

but the correlation in compositional data may be positive. Another disadvantage of Dirichlet class is the strong independence structure because of the relationship between the Dirichlet and gamma classes. These disadvantages make the Dirichlet distribution unsuitable for compositional data analysis.

Additive logratio transformation

The idea of perturbation operation leads to the Logistic Normal distribution. Aitchison and Shen [1980] first introduced the Logistic Normal (LN) distribution for modeling compositional data. Aitchison [1982] established many mathematical and statistical properties of LN distribution. The methods he used for modeling rely on the the additive logratio transformation. He used the additive logratio transformation to take the compositional data from the (k-1)-dimensional simplex (∇^{k-1}) to (k-1)-dimensional Euclidean space (R^{k-1}). Aitchison models the compositional data through the $(k-1)$ multivariate normal distribution by using additive logratio transformation. The main benefit is the rich covariance structure from the multivariate normal assumption, and this allows modeling the dependence between pairs of the k elements. However, interpretation of parameter estimates on the multivariate log-odds scale is difficult.

Definition 8 *The additive logratio transformation of $\underline{z}$ from (k-1)-dimensional simplex to (k-1)-dimensional Euclidean space (R^{k-1})is defined as*

$$\phi(\underline{z}) = \left[log\left(\frac{z_1}{z_k}\right), log\left(\frac{z_2}{z_k}\right), \ldots, log\left(\frac{z_{k-1}}{z_k}\right)\right]$$

where $\phi(.)$ is denoted as additive logratio transformation.

The additive logratio transformation is one-to-one and the inverse of this transformation $\phi^{-1}(.)$ transforming $\underline{y}$ from (k-1)-dimensional Euclidean space (R^{k-1}) to (k-1)-dimensional simplex is defined by

$$z_i = \frac{exp(y_i)}{\sum_{j=1}^{k-1} exp(y_j) + 1}, (i = 1, 2, \ldots, k-1)$$

$$and\ z_k = \frac{1}{\sum_{j=1}^{k-1} exp(y_j) + 1}$$

where $y_i = log\left(\frac{z_i}{z_k}\right)$.

Density Function

A k-part composition, $\underline{z}$, is said to have a Logistic Normal distribution $L^{k-1}(\mu, \Sigma)$ when $\underline{y} = log\left(\frac{z_{-k}}{z_k}\right)$ has a k-1 dimensional Multivariate Normal distribution with mean $\underline{\mu}$ and covariance matrix Σ, where $z_{-k} = (z_1, z_2, \ldots, z_{k-1})$.

The density function for $L^{k-1}(\underline{\mu}, \Sigma)$ is

$$f(\underline{z}|\underline{\mu}, \Sigma) = \left(\frac{1}{2\pi}\right)^{\frac{k-1}{2}} |\Sigma|^{-\frac{1}{2}} \left(\frac{1}{\prod_{i=1}^{k} z_i}\right) exp\left[-\frac{1}{2}(\phi(\underline{z}) - \underline{\mu})'\Sigma^{-1}(\phi(\underline{z}) - \underline{\mu})\right] \qquad (2.1)$$

where $\underline{\mu}$ is the mean parameter vector in R^{k-1}, and Σ is a $(k-1) \times (k-1)$ variance-covariance matrix. The i^{th} element of $\underline{\mu}$, μ_i , can be interpreted as $E\left[log(\frac{z_i}{z_k})\right]$, and the ij^{th} element of Σ, σ_{ij}, can be interpreted as $cov\left[log(\frac{z_i}{z_k}), log(\frac{z_j}{z_k})\right]$. The $\left(\frac{1}{\prod_{i=1}^{k} z_i}\right)$ term is the Jacobian of the additive logratio transformation. Aitchison [1986] showed that the logistic normal density function is invariant with respect to the permutations of the components. Therefore, the density, and any inference based on the density, is not affected by the ordering of components in $\underline{z}$.

Interpretation of parameters

The mean parameter $\underline{\mu}$ in the LN distribution can be expressed as a composition through the inverse additive logratio transformation. We define

$$\phi^{-1}(\underline{\mu}) = \underline{\xi} \tag{2.2}$$

where $\underline{\xi} \in \nabla^{k-1}$.

$\underline{\xi}$ is a composition in simplex sample space. Therefore, interpretation of $\underline{\xi}$ in the simplex is much easier than interpreting $\underline{\mu}$ on the multivariate logratio scale in Euclidean space (R^{k-1}). However, Billheimer and Guttorp [1995] pointed that some statistical properties of $\underline{\mu}$ are lost with the transformation to the simplex. When $\underline{\mu}$ is the mean and mode of the multivariate normal logit, the ϕ^{-1} transformation does not preserve these properties. However, the ϕ^{-1} transformation is monotone in each of the (k-1) components of $\underline{\mu}$. As a consequence, ordering of the values is preserved under this transformation. Therefore, $\underline{\xi} = \phi^{-1}(\underline{\mu})$ can be interpreted as the component-wise multivariate median for the LN distribution in ∇^{k-1}. This interpretation is useful when we get point estimates of parameters, and can be treated as the "center" for the asymmetric LN distribution.

Aitchison [1986] introduced the definition of logratio linear model to incorporate the effect of explanatory variables.

Definition 9 *A set of N independent K-part compositions is said to follow a logratio linear model if the logratio data matrix $\underline{Y}$ can be expressed in the form*

$$\underline{Y} = \underline{A\Theta} + \underline{E},$$

where the covariate matrix $\underline{A}$, of order $N \times p_m$ and full rank p_m, is a matrix of known

constants, the parameter matrix $\underline{\Theta}$ *is of order* $p_m \times d$ *and the* $N \times d$ *error matrix* $\underline{E}$ *is assumed to consist of independent row vectors, each distributed as* $N^k(0, \Sigma)$ *(Aitchison [1986]; section 7.6).*

The matrix $\underline{\Theta}$ is the "regression" parameter matrix of the logratio linear model, and the "error" logratio covariance matrix is Σ. Estimation of $\underline{\Theta}$ and Σ under the model is standard, either by maximum likelihood under the normality assumption or by multivariate least squares.

For example, if we have a scalar covarate $x_j, j = 1, 2, \ldots, n$, then $\underline{\mu}_j$ in the density function can be replaced by $\underline{\beta}_0 + \underline{\beta}_1(x_j - \bar{x})$. $\underline{\beta}_0$ and $\underline{\beta}_1$ are parameter vectors in R^{k-1}, and $\bar{x}$ is the mean of the observed covariate values. We can interpret $\underline{\beta}_0$ as the location when $x_j = \bar{x}$, and $\underline{\beta}_1$ as the change of location for one unit change of x under this parameterization.

By using the inverse additive logratio transformation, the linear regression expression $\underline{\mu}_j = \underline{\beta}_0 + \underline{\beta}_1(x_j - \bar{x})$ can be expressed as the perturbation of compositions. Taking the inverse additive logratio transformation on both sides of the equation,

$$\phi^{-1}(\underline{\mu}_j) = \phi^{-1}(\underline{\beta}_0) \oplus \phi^{-1}(\underline{\beta}_1)^{(x_j - \bar{x})}.$$

we have the perturbation of compositions

$$\underline{\xi}_j = \underline{\xi} \oplus \gamma^{\underline{\mu}_j}.$$

where $\xi_j = \phi^{-1}(\underline{\mu}_j), \underline{\xi} = \phi^{-1}(\underline{\beta}_0)$, and $\underline{\gamma} = \phi^{-1}(\underline{\beta}_1)$. The $\underline{\xi}$ is the overall location on the simplex. The location $\underline{\xi}_j$ is the overall location $\underline{\xi}$ perturbed by $\underline{\gamma}$, the regression composition parameter. $\underline{\gamma}$ is the regression parameter on simplex, and it is the amount a location shifted, via a perturbation, when the covariate variables changed in one

unit. The deviations in $\underline{\gamma}$ from the identity composition, $\underline{I}_{k-1}$, indicate the direction and magnitude of the change. If $\underline{\gamma}$ equals $\underline{I}_{k-1}$, then it shows that the covariate variables have no effect on the compositions.

2.1.4 State-Space model for Discrete Compositions

Aitchison's Logistic Normal model using observed compositions as data treats compositions as continuous variables. However, when the observed data involve discrete variables, for example, the count data, Aitchison's model might underestimate the actual variability of the observed data. In this situation, including the discrete observations in the model might better reflect the actual variability of the observed compositions.

Billheimer, Guttorp, and Fagan [2001] proposed a State-Space Model which combined the logistic model for continuous compositions and the conditional multinomial observations distribution. In this model, he posits a latent composition vector, z, associated with the observed count data, y. For observation $\underline{y}$, a k-vector of counts, given $m = \sum_{i=1}^{k}[y]_i$, and $\underline{z}$ (the unobserved composition vector, latent variable), $\underline{y}$ follows the multinomial distribution with probability mass function

$$p(\underline{y}|\underline{z}, m) = \frac{m!}{\prod_{i=1}^{k}[y]_i!} \prod_{i=1}^{k}[z]_i^{[y]_i} \tag{2.3}$$

where $[.]_i$ denotes the i^{th} element of the vector. $\underline{z}$ follows the logistic normal distribution, $L_{k-1}(\underline{\mu}, \Sigma)$.

By combining the multinomial distribution and logistic normal distribution, the joint

density function of $\underline{y}$, $\underline{z}$ is:

$$p(\underline{y},\underline{z}|\sum_{i=1}^{k} y_i = m, \underline{\mu}, \Sigma) = \frac{m!}{\prod_{i=1}^{k}[y]_i!}\prod_{i=1}^{k}[z]_i^{[y]_i-1}\left(\frac{1}{2\pi}\right)^{\frac{k-1}{2}}|\Sigma|^{-\frac{1}{2}} exp\left[-\frac{1}{2}(\phi(\underline{z})-\underline{\mu})'\Sigma^{-1}(\phi(\underline{z})-\underline{\mu})\right] \quad (2.4)$$

Billheimer, Guttorp, and Fagan [2001] uses Markov Chain Monte Carlo (MCMC) for model inference about the unknown logistic normal distribution parameters and the latent composition vectors. The conditional distribution of $\underline{z}$ (given current values of $\underline{\mu}$ and Σ) is sampled by Metropolis-Hastings algorithm, and the LN distribution parameters $\underline{\mu}$ and Σ can be updated by Gibbs sampling step.

2.2 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) method is a class of algorithms for sampling from approximate distributions and then correcting those samples to better approximate the target posterior distribution. The samples are dawn based on constructing a Markov chain. Markoc chain is a sequence of random variables $\theta^1, \theta^2, \ldots$, for which, at any time t, the distribution of θ^t given all previous θ's depends only on the most recent value, θ^{t-1} (Gelman, Carlin, Stern, and Rubin [2004]). Grenander [1983], and Geman and Geman [1984] first used MCMC method for Bayesian inference. After that, many researchers expanded the use of MCMC in Bayesian inference (Gelfand and Smith [1990], Gelfand, Hills, Racine-Poon, and Smith [1990]; Besag, Green, Higdon, and Mengersen [1995] for a review of the methodology; and Tierney [1994]). Good introductions to MCMC are given by Gelman, Carlin, Stern, and Rubin [2004], Carlin and Louis [2000], and Gilks, Richardson, and Spiegelhalter [1996].

2.2.1 Monte Carlo Integration

The original Monte Carlo method was developed by physicists to use random number generation to compute integrals. Suppose we have a complex integral for which the exact integration does not exist,

$$\int_a^b h(x)dx$$

If we can decompose h(x) into the product of a function f(x) and a probability density function p(x) defined over the interval (a, b), then the integral becomes

$$\int_a^b h(x)dx = \int_a^b f(x)p(x)dx = E_{p(x)}[f(x)].$$

Thus, the integral can be expressed as an expectation of f(x) over the density function p(x). If we draw a large number of random variables $(x_1, \ldots, x_n)$ from the density p(x), the we have

$$\int_a^b h(x)dx = E_{p(x)}[f(x)] = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

This method is called Monte Carlo integration.

2.2.2 The Gibbs Sampler

The Gibbs sampler (Geman and Geman [1984]) can be viewed as a special case of the Metropolis-Hastings algorithm where the proposal θ_i^* is from the full conditional distribution of θ_i. In this algorithm, the proposal θ_i^* is always accepted. Suppose we have the parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$. At each iteration t, an ordering of the d subvectors of θ is chosen, and each θ_j^t is sampled from the conditional distribution given all the other components of θ:

$$p(\theta_j|\theta_{-j}^{t-1}, y),$$

where θ_{-j}^{t-1} is all the components of θ, except for θ_j, at their current values:

$$\theta_{-j}^{t-1} = (\theta_1^t, \ldots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \ldots, \theta_d^{t-1}).$$

Therefore, each subvector θ_j is updated conditional on the latest values of the other components of θ.

We can transition from θ^t to θ^{t+1} by using the following steps.

Given θ^t, generate

1. $\theta_1^{t+1} \sim f_1(\theta_1|\theta_2^t, \ldots, \theta_d^t)$
2. $\theta_2^{t+1} \sim f_2(\theta_2|\theta_1^{t+1}, \theta_3^t, \ldots, \theta_d^t)$
3. $\vdots$
4. $\theta_d^{t+1} \sim f_d(\theta_d|\theta_1^{t+1}, \theta_2^{t+1}, \ldots, \theta_{d-1}^t)$

The densities $f_1, \ldots, f_d$ are full conditional densities. One of the nice features of the Gibbs sampler is that they are the only densities used for sampling, so all updates can be univariate if desired. One does not need to propose an entire vector of parameters, as with the MH algorithm. Another feature is that there is no accept or reject step in the algorithm.

2.2.3 The Metropolis-Hastings algorithm

Many methods have been developed for constructing and sampling from transition distributions for posterior distributions. The Metropolis-Hastings algorithm is a general term for drawing samples from Bayesian posterior distributions(Metropolis and Ulam [1949], Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller [1953], Hastings [1970]). The Metropolis-Hastings algorithm uses a random walk, and then applies an

acceptance/rejection rule to converge to the specified target posterior distribution. The algorithm is as follows [Gelman, Carlin, Stern, and Rubin, 2004].

1. Choose a starting point θ^0 satisying $p(\theta^0|y) > 0$, from a starting distribution $p_0(\theta)$.

2. For $t = 1, 2, \ldots$

   (a) Sample a proposal θ^* from a jumping distribution (or proposal distribution) at time t, $J_t(\theta^*|\theta^{t-1})$.

   (b) Calculate the ratio of the densities,

$$r = \frac{p(\theta^*|y)/J_t(\theta^*|\theta^{t-1})}{p(\theta^{t-1}|y)/J_t(\theta^{t-1}|\theta^*)}$$

   (c) Define

$$\theta^t = \begin{cases} \theta^* & \text{with probability min(r,1)} \\ \theta^{t-1} & \text{otherswise.} \end{cases}$$

A further simplification occurs when the proposal distribution is symmetric, so that $J_t(\theta^*|\theta^{t-1}) = J_t(\theta^{t-1}|\theta^*)$. As a result,

$$r = \frac{p(\theta^*|y)}{p(\theta^{t-1}|y)}$$

and the resulting algorithm is simply called the Metropolis algorithm. An example is a random walk MH sampler, in which $J(\theta^*|\theta_{t-1}, y)$ is a normal distribution with mean θ_{t-1} and variance σ_y^2.

The algorithm accepts the proposal θ^* with the probability $min(r, 1)$. If $\theta_t = \theta_{t-1}$, the jump is not accepted.

2.3 The Limitations of Existing Methods

2.3.1 Compositional Data Analysis

The natural first thought for analyzing compositional data is to use the Dirichlet distribution. However, since every Dirichlet composition can be viewed as a composition formed from a basis of independent gamma distributed components, this class has a very strong independent structure which makes it unappropriate for data with dependence. Aitchison [1982, 1986] did the fundamental work for compositional data analysis. He introduced the Logistic Normal distribution into compositional data and established many mathematical and statistical properties of LN distribution. The strategy he used is the additive logratio transformation which can transform the compositional data from the (k-1)-dimensional simplex (∇^{k-1}) to (k-1)-dimensional Euclidean space (R^{k-1}). Aitchison models the compositional data through the $(k-1)$ multivariate normal distribution by using additive logratio transformation. However, as Aitchison [1986] and others (e.g. Pawlowski and Burger [1992]) describe, the interpretation of parameter estimates on the multivariate log-odds scale is difficult.

Billheimer, Guttorp, and Fagan [2001] introduced an algebra for compositions that includes addition, scalar multiplication, and a metric for differences in compositions. The algebra aids interpretation of treatment effects, treatment interactions, and covariates. Also Billheimer et al. [2001] presented a hierarchical statistical model which combines the logistic normal for continuous compositions with a conditional multinomial distribution. This method provides a tool for analyzing compositional count data at a single time point. However, in our motivation example, PIP study, the measurements are compositional data observed over a period of time. Because of the special properties of compositional data and associated simplex sample space, the

traditional statistical methods for multivariate repeated measurements data are not appropriate. There is no other author discussing the statistical methods for repeated compositional data. The statistical model for repeated compositional data will be presented in chapter 3.

2.3.2 Measurement Error in Longitudinal Data

Measurement error problems in predictor variables have recently received extensive attention by researchers. There are many papers about how to correct measurement errors in regressors in different applications (Carroll and Stefanski [1990]; Brown and Fuller [1990]; Byar and Gail [1989]; Fuller [1987]). However, there is little literature about the problem with measurement errors in the response variable. The main reason is that the problem can be handled with standard methodology if the measurement errors in response variable are additive errors in which the observed values = true value + error, where the error has mean 0. When the measurement errors are additive with constant variance, the errors can be ignored in regression analysis because they can be thought as an extra variance component.

However, it has been recognized that for many situations, the additive error model is often not appropriate(Buonaccorsi [1989, 1990a,b, 1991], Buonaccorsi and Tosteson [1993], Buonaccorsi [1996], Carroll, Gail, and Lubin [1993], Pepe [1992], Rosner, Spiegelman, and Willett [1990], Tosteson, Stefanski, and Schafer [1989]). A specific example was described by Buonaccorsi and Tosteson [1993]. In this example, the serum neopterin level is measured by a radioimmunoassay, and the observed value is a standardized radioactive count. The measurement error model in this example is a four-parameter logistic model. In a series of papers, Buonaccorsi [1991, 1996] and Buonaccorsi and Tosteson [1993] discussed how to correct bias when the mea-

surement errors are not additive. Buonaccorsi [1996] considered the measurement error in the response variable in the general linear model. He described the full and pseudo-maximum likelihood estimators under distributional assumptions.

Longitudinal data are obtained when subjects are followed over a period of time, and for each subject, some variables are measured at multiple time points. The primary goal of a longitudinal study is to characterize the change in response over time and the factors that affect the change. Because of the repeated measures on individuals, the researchers can capture within-individual change. In our motivation example, the PIP study, the white blood cell counts from the patients' blood samples were measured over a period of time. We assume that the true cell count values should follow a normal distribution since the cell counts range from several hundreds to several thousands. Therefore, longitudinal analysis should be used because of the correlation of repeated measurements in one individual. One of the approach for analyzing longitudinal data is the general multivariate model.

However, Buonaccorsi [1996] only discussed the measurement error for a general linear model. His methods didn't cover the measurement error in the response in longitudinal/repeated-measures problems. There is no other author discussing the measurement error in response only in longitudinal data although the longitudinal studies are popular in medical studies and clinical trials, etc. We propose a likelihood based method to correct for measurement errors in general multivariate linear model that will be presented in chapter 4. We also propose a Bayesian approach to correct for measurement errors in general multivariate linear model that will be presented in chapter 5.

Chapter 3

BAYESIAN ANALYSIS OF REPEATED COMPOSITIONAL DATA

3.1 Introduction

Multiple measurements over time on a response variable on the same experimental unit or the same subject lead to repeated measurements (or longitudinal) data. Such data are very common in many field such as biomedical, pharmaceutical, industrial engineering, business etc. If multiple measurements are made over time on more than one response variables on the same subject or experimental unit, we call this type of data "multivariate repeated measures data". We need to pay special attention for this type of data since the measurements made over time on the same individual are typically correlated.

In our motivating example, the Protective Immunity Project (PIP study), the patients' blood samples were collected at different time points. Each blood sample was analyzed by Flow cytometer, and the subpopulation of blood cells were recorded. Therefore, the data in this study is also the multivariate repeated measures data. However, the measurements on multiple response variables at each time point are compositional data. The sample space associated with the K-part compositions is the (k-1) dimensional unit simplex (∇^{k-1}). Because of the special properties of compositional data and associated simplex sample space, the traditional statistical methods for multivariate repeated measurements data are not appropriate. The goal of PIP study is to evaluate the change of white blood cell (Lymphocytes) compositions over time. Billheimer, Guttorp, and Fagan [2001] proposed a hierarchical statistical model combining Aitchison's (1982, 1986) logistic normal (LN) distribution with a conditional multinomial model. They used Markov Chain Monte Carlo (MCMC) approach for model inference. His method provided a tool for analyzing compositional count data at a single time point. His method did not cover the longitudinal/repeated measures compositional data. There is no other author discussing the compositional

data analysis in longitudinal/repeated measures situations although the longitudinal studies represent one of the principle research strategies used in medical and social science research (Goldstein [1979], Nesselroade and Baltes [1979]). In this chapter, we propose a multivariate logistic normal moodel, and use the MCMC approach for model inference. We use the algebra introduced by Billheimer et al. [2001] for parameter interpretation. We also use Aitchison's Logistic normal distribution to model the compositional data. Relying on the additive logratio transformation, this approach transforms the compositional data into multivariate normal distribution. Because of the non-linear ALR transformation, the likelihood methods become impractical. Bayesian methods become more attractive under this situation. Hence we also propose Bayesian approach for our model inference, and demonstrate its characteristics.

3.2 Model Structure

Suppose we have n subjects, and they were followed for t time points, and for each subject at time t, the composition has k components. We assume the composition vector for subject j, at time point t, $\underline{z}_{jt}$ is a k-part composition. Since the measurements made at different time points on the same individual may be correlated, we assume the composition vector for subject j, $\underline{z}_j$ follows the logistic normal distribution (LN) with t(k-1) dimensions. So the joint density function of $\underline{z}_{j1}, \underline{z}_{j2}, \ldots, \underline{z}_{jt}$ is

$$p(\underline{z}_{j1}, \underline{z}_{j2}, \ldots, \underline{z}_{jt} | \underline{\mu}_1, \underline{\mu}_2, \ldots, \underline{\mu}_t, \Omega) =$$
$$\left(\frac{1}{2\pi}\right)^{\frac{t(k-1)}{2}} |\Omega|^{-\frac{1}{2}} \frac{1}{\prod_{t=1}^{T} \prod_{i=1}^{k} z_{jti}} exp\left[-\frac{1}{2}\left(\phi(\underline{z}_j) - \underline{\mu}\right)^T \Omega^{-1} \left(\phi(\underline{z}_j) - \underline{\mu}\right)\right] \tag{3.1}$$

where $\underline{\mu} = \begin{pmatrix} \underline{\mu}_1 \\ \vdots \\ \underline{\mu}_T \end{pmatrix}$, is a t(k-1)-dimension mean vector, and Ω is the $t(k-1) \times t(k-1)$ dimension variance-covariance matrix. Also please note that the $\frac{1}{\prod_{t=1}^{T}\prod_{i=1}^{k} z_{jti}}$ term in the density function is the Jacobian of the additive logratio transformation.

3.3 Model Inference

We can use Bayesian approach for making inference about model parameters.

To implement Bayesian analysis, first we need to write down the likelihood function of the model. The likelihood function is

$$\begin{aligned} L(\underline{z}|\underline{\mu}, \Omega) &= \prod_{j=1}^{n} p(\underline{z}_j|\underline{\mu}) \\ &= \prod_{j=1}^{n} \frac{1}{\prod_{t=1}^{T}\prod_{i=1}^{k} z_{jti}} (\frac{1}{2\pi})^{\frac{(k-1)t}{2}} |\Omega|^{-\frac{1}{2}} \\ &\quad \times exp\left[-\frac{1}{2}\left(\phi(\underline{z}_j) - \underline{\mu}\right)^T \Omega^{-1} \left(\phi(\underline{z}_j) - \underline{\mu}\right)\right] \end{aligned} \tag{3.2}$$

where $\phi(\underline{z}_j) = \begin{pmatrix} \phi(\underline{z}_{j1}) \\ \vdots \\ \phi(\underline{z}_{jt}) \end{pmatrix}$, $\underline{z}_{jt} \in \nabla^{k-1}$, and Ω is the $t(k-1) \times t(k-1)$ dimension variance-covariance matrix.

Next, we need to specify the prior distributions for $\underline{\mu}$ and Ω. let $\underline{\mu}$ have the t(k-1) dimensional multivariate Normal distribution with mean vector $\underline{\eta}$, and variance-covariance matrix Q. Also we assume that $\Omega^{-1} \sim$ Wishart (Ψ^{-1}, ρ), where Ψ is a $t(k-1) \times t(k-1)$ positive definite matrix, and ρ denotes the degrees of freedom. The

value of ρ is set to $t(k-1)$.

We can choose the hyperparameters as

$$\begin{aligned} \underline{\eta} &= 0_{t(k-1)} \\ Q &= aN \\ \Psi &= cN \end{aligned} \tag{3.3}$$

where N is a $t(k-1) \times t(k-1)$ positive definite variance-covariance matrix, and we typically choose $N = I_{t(k-1)} + j_{t(k-1)} j'_{t(k-1)}$. $0_{t(k-1)}$ is the t(k-1)-vector of 0s, $I_{t(k-1)}$ is the t(k-1) identity matrix, and $j_{t(k-1)}$ is the t(k-1)-vector of 1s. a and c are scalars. The dispersion matrix, N, specifies a "null" variance-covariance matrix between log-ratio transformed compositions. That is, a priori one may consider the compositional elements "independent except for the summation constriant" (Aitchison [1986], Billheimer and Guttorp [1997]). The value of ρ is the smallest (least informative) that still can maintain a proper Wishart distribution. Thus, the prior distribution for the mean vector $\underline{\eta}$ is centered at $I_{t(k-1)}$ and disperse over the simplex space. The prior distribution for Ω is centered at the "null" precision matrix (i.e., compositions formed from independent bases; see Billheimer and Guttorp [1995]). a and c are scalars. The value of a is selected to allow the 95% prior probability contour for $\underline{\xi}$=inverse alr($\underline{\mu}$) to reach at least 0.05 for each component. The value of c is chosen so that the observed variance of simulated compositions approximates that observed in the data. These values specify proper, but diffuse prior distributions(Billheimer and Guttorp [1995]). During the analysis, several widely varying hyperparameters will be tried to ensure that inferences are data dependent.

Combining the likelihood and the prior distribution, we can get the posterior

distribution as

$$\begin{aligned}\pi(\underline{\mu},\Omega|\underline{z}) \quad \propto \quad & \\ & \prod_{j=1}^{n}\left[(\frac{1}{2\pi})^{\frac{(k-1)t}{2}}|\Omega|^{-\frac{1}{2}}exp\left[-\frac{1}{2}\left(\phi(\underline{z}_j)-\underline{\mu}\right)^T\Omega^{-1}\left(\phi(\underline{z}_j)-\underline{\mu}\right)\right]\right] \\ & \times|Q|^{-\frac{1}{2}}exp\left[-\frac{1}{2}(\underline{\mu}-\underline{\eta})^TQ^{-1}(\underline{\mu}-\underline{\eta})\right] \\ & \times|\Psi|^{\frac{\rho}{2}}|\Omega|^{-\frac{\rho-t(k-1)-1}{2}}exp\left[-\frac{1}{2}tr(\Psi\Omega^{-1})\right] \qquad (3.4)\end{aligned}$$

We can get the full conditionals for $\underline{\mu}$ and Ω^{-1} from the posterior density function.

$$\begin{aligned}\pi(\underline{\mu}|\ldots) \quad \propto \quad & exp\left[-\frac{1}{2}\left(\underline{\mu}-\underline{\eta}\right)^TQ^{-1}\left(\underline{\mu}-\underline{\eta}\right)\right] \\ & \times exp\left[-\frac{1}{2}\sum_{j=1}^{n}\left(\phi(\underline{z}_j)-\underline{\mu}\right)^T\Omega^{-1}\left(\phi(\underline{z}_j)-\underline{\mu}\right)\right] \qquad (3.5)\end{aligned}$$

$$\begin{aligned}\pi(\Omega^{-1}|\ldots) \quad \propto \quad & |\Omega|^{-\frac{(\rho-t(k-1)+n-1)}{2}} \\ & \times exp\left\{-\frac{1}{2}\left[\sum_{j=1}^{n}(\phi(\underline{z}_j)-\underline{\mu})^T\Omega^{-1}(\phi(\underline{z}_j)-\underline{\mu})+tr(\Psi\Omega^{-1})\right]\right\} \qquad (3.6)\end{aligned}$$

With the observed compositional data $\underline{z}_j$, $\underline{\mu}$ and Ω^{-1} specified, we can implement MCMC method by using Gibbs sampling methods. We can update the values of $\underline{\mu}$ and Ω accordingly based on $\underline{z}_j$. Because the conditional distributions of $\underline{\mu}$ and Ω are available in closed form (multivariate normal distribution and inverse Wishart distribution, respectively), we can use Gibbs sampling method to update $\underline{\mu}$ and Ω. By using some algebra, we find that the conditional distribution of $\underline{\mu}$ is multivariate

normal with mean

$$(Q^{-1} + n\Omega^{-1})^{-1}(n\Omega^{-1}\frac{1}{n}\sum_{j=1}^{n}\phi(\underline{z}_j) + Q^{-1}\underline{\eta})$$

and variance-covariance matrix

$$(Q^{-1} + n\Omega^{-1})^{-1}.$$

The conditional distribution of the variance-covariance matrix Ω is an inverse Wishart distribution with parameter matrix $(V + \Psi)$, where

$$V = \left[\sum_{j=1}^{n}\phi(\underline{z}_j) - \underline{\mu}\right]\left[\sum_{j=1}^{n}\phi(\underline{z}_j) - \underline{\mu}\right]^T$$

and $n + \rho - t * (k-1)$ degrees of freedom.

3.3.1 Incorporate The Effect of Covariates

The mean parameter $\underline{\mu}$ can depend on the explanatory variables. Therefore the effect of covariates can be incorporated into our model. Suppose we have a scalar covariate X, where X is a time independent covariate, and does not change over time. $\underline{\mu}$ can be replaced in the model by $\underline{\alpha} + \underline{\beta}X$. In this expression, $\underline{\alpha}$ and $\underline{\beta}$ are the vectors in $R^{t*(k-1)}$.

The likelihood function becomes

$$\begin{aligned} L(z|\underline{\alpha}, \underline{\beta}, \Omega) &= \prod_{j=1}^{n} p(\underline{z}_j|\underline{\alpha}, \underline{\beta}, \Omega) \\ &= \prod_{j=1}^{n} \frac{1}{\prod_{t=1}^{T}\prod_{i=1}^{k} z_{jti}}(\frac{1}{2\pi})^{\frac{(k-1)t}{2}}|\Omega|^{-\frac{1}{2}} \end{aligned}$$

$$exp\left[-\frac{1}{2}\left(\phi(\underline{z}_j)-(\underline{\alpha}+\underline{\beta}X)\right)^T\Omega^{-1}\left(\phi(\underline{z}_j)-(\underline{\alpha}+\underline{\beta}X)\right)\right] \tag{3.7}$$

where $\phi(\underline{z}_j)=\begin{pmatrix}\phi(\underline{z}_{j1})\\ \vdots \\ \phi(\underline{z}_{jt})\end{pmatrix}$, $\underline{z}_{jt}\in\nabla^{k-1}$, and Ω is the $t(k-1)\times t(k-1)$ dimension variance-covariance matrix.

We can define the prior distributions as

$$\Omega^{-1}\sim Wishart(\Psi^{-1},\rho) \tag{3.8}$$

where Ψ is a $t(k-1)\times t(k-1)$ positive definite matrix, and ρ denotes the degrees of freedom. The value of ρ is set to $t(k-1)$.

$$\begin{aligned}\underline{\alpha} &\sim N(\underline{\alpha}_0,\sigma_\alpha I)\\ \underline{\beta} &\sim N(\underline{\beta}_0,\sigma_\beta I)\end{aligned} \tag{3.9}$$

Then we can define the hyperparameters as

$$\begin{aligned}\underline{\beta}_0 &= 0_{t(k-1)}\\ \underline{\alpha}_0 &= 0_{t(k-1)}\\ \sigma_\alpha &= \sigma_\beta = 10\\ \Psi &= cN, where \quad c=0.1\end{aligned} \tag{3.10}$$

During the analysis, we can try several widely varying hyperparameters to check out

the sensitivity of the model inferences to ensure that the inference is data dependent.

Next, by combining the likelihood function and the prior distributions, we can get the posterior distribution

$$\begin{aligned}\pi(\underline{\alpha},\underline{\beta},\Omega|\underline{z}) \quad \propto \\ & \prod_{j=1}^{n}\left[(\frac{1}{2\pi})^{\frac{(k-1)t}{2}}|\Omega|^{-\frac{1}{2}}exp\left[-\frac{1}{2}\left(\phi(\underline{z}_j)-(\underline{\alpha}+\underline{\beta}X)\right)^T\Omega^{-1}\left(\phi(\underline{z}_j)-(\underline{\alpha}+\underline{\beta}X)\right)\right]\right] \\ & \times|\sigma_{\underline{\beta}}|^{-\frac{1}{2}}exp\left[-\frac{1}{2}(\underline{\beta}-\underline{\beta}_0)^T(\sigma_\beta I)^{-1}(\underline{\beta}-\underline{\beta}_0)\right] \\ & \times|\sigma_{\underline{\alpha}}|^{-\frac{1}{2}}exp\left[-\frac{1}{2}(\underline{\alpha}-\underline{\alpha}_0)^T(\sigma_\alpha I)^{-1}(\underline{\alpha}-\underline{\alpha}_0)\right] \\ & \times|\Psi|^{\frac{\rho}{2}}|\Omega|^{-\frac{\rho-t(k-1)-1}{2}}exp\left[-\frac{1}{2}tr(\Psi\Omega^{-1})\right] \end{aligned} \tag{3.11}$$

The full conditional distributions for $\underline{\alpha}$, $\underline{\beta}$ and Ω^{-1} can be deduced from the posterior density function.

$$\begin{aligned}&(\underline{\beta}|\ldots)\sim \\ &N\left\{[\sum_{j=1}^{n}X_j^2\Omega^{-1}+\frac{I}{\sigma_\beta}]^{-1}(\Omega^{-1}\sum_{j=1}^{n}X_j(\phi(\underline{z}_j)-\underline{\alpha})+\frac{\underline{\beta}_0 1}{\sigma_\beta}),\quad [\sum_{j=1}^{n}X_j^2\Omega^{-1}+\frac{I}{\sigma_\beta}]^{-1}\right\}\end{aligned} \tag{3.12}$$

$$\begin{aligned}&(\underline{\alpha}|\ldots)\sim \\ &N\left\{[n\Omega^{-1}+\frac{I}{\sigma_\alpha}]^{-1}(\Omega^{-1}[\sum_{j=1}^{n}\phi(\underline{z}_j)-\underline{\beta}\sum_{j=1}^{n}X_j]+\frac{\underline{\alpha}_0 1}{\sigma_\alpha}),\quad [n\Omega^{-1}+\frac{I}{\sigma_\alpha}]^{-1}\right\}\end{aligned} \tag{3.13}$$

$$(\Omega^{-1}|\ldots)\sim Wishart(V+\Psi,df) \tag{3.14}$$

where

$$V = \left[\sum_{j=1}^{n} \phi(\underline{z}_j) - (\underline{\alpha} + \underline{\beta} X)\right] \left[\sum_{j=1}^{n} \phi(\underline{z}_j) - (\underline{\alpha} + \underline{\beta} X)\right]^T$$

$$df = n + \rho - t * (k - 1).$$

3.3.2 Model Diagnostics

It is crucial to check the model in Bayesian data analysis, because the Bayes prior-to-posterior inferences condition on the whole probability model, and can be very misleading when the model is poor. Therefore, a Bayesian analysis should include at least some methods for model checking to find out the fit of the model to the data and the plausibility of the model for the purpose for which the model is used. We can conduct model checking in three ways.

1. Examining the sensitivity of inferences to reasonable changes in the prior distribution and the likelihood. We can try several widely varying hyperparameters to check out the sensitivity of the model inference to ensure that the inference is data dependent.

2. Checking that the posterior inference is reasonable, given the substantive context of the model.

3. Checking that whether the model fits the data. We will use Gelman, Meng, and Stern [1996]'s method to address this model checking method.

Gelman et al. [1996] use simulated values of a discrepancy measure $D(.,.)$ from the posterior predictive distribution, and compare to the same discrepancy measure $D(.,.)$ from the observed data. Let $\underline{\theta}$ denotes all the d-dimensional unknown parameters, $\underline{z}$ denote observed data, and H for our assumed model. $\underline{z}^{rep}$ denotes the *replicated*

data that could have been observed. $\underline{z}^{rep}$ is a replication like $\underline{z}$. For example, if we have a explanatory variable X in the model, $\underline{z}$ and $\underline{z}^{rep}$ should have identical values of X. The reference distribution of the future observation $\underline{z}^{rep}$ is its posterior predictive distribution,

$$P(\underline{z}^{rep}|\underline{z}) = \int P(\underline{z}^{rep}|\underline{\theta})P(\underline{\theta}|\underline{z})d\underline{\theta}$$

For a selected discrepancy, $D(\underline{z};\underline{\theta})$, its reference distribution is derived from the joint posterior distribution of $\underline{z}^{rep}$ and $\underline{\theta}$,

$$P(\underline{z}^{rep},\underline{\theta}|H,\underline{z}) = P(\underline{z}^{rep}|H,\underline{\theta})P(\underline{\theta}|H,\underline{z})$$

Then, the tail-area probability of D under its posterior reference distribution is defined as

$$P_B = P[D(\underline{z}^{rep};\underline{\theta}) \geq D(\underline{z};\underline{\theta})|H,\underline{z})]$$

The Bayesian tail-area probability is the probability that the replicated data could be more extreme than the observed data as measured by the discrepancy $D(.;.)$.
We can use simulation to compute the posterior predictive distribution. Suppose we already have J simulations from the posterior density of $\underline{\theta}$ during model inference, $J = 1, 2, \ldots, J$.

1. For each simulated $\underline{\theta}^j$, we draw one $\underline{z}^{rep,j}$ from the predictive distribution, $P(\underline{z}^{rep,j}|H,\underline{\theta}^j)$.

2. Calculate $D(\underline{z};\underline{\theta}^j)$ and $D(\underline{z}^{rep,j};\underline{\theta}^j)$.

The estimated p-value is just the proportion of these J simulations that the $D(\underline{z}^{rep,j};\underline{\theta}^j) \geq D(\underline{z};\underline{\theta}^j)$ $J = 1, 2, \ldots, J$. The idea of this test is that if the model is suitable, this probability would be about 0.5, and the model would be question-

able if this probability is either close to 1 or 0. The comparison of $D(\underline{z};\underline{\theta}^j)$ and $D(\underline{z}^{rep,j};\underline{\theta}^j)$ can also be displayed as a scatterplot or a histogram of the difference, $D(\underline{z};\underline{\theta}^j) - D(\underline{z}^{rep,j})$. If the model is appropriate, the scatterplot of the values $D(\underline{z};\underline{\theta}^j)$ vs. $D(\underline{z}^{rep,j};\underline{\theta}^j)$ should be symmetric about the 45^o line, and the histogram should include 0.

3.4 Data Analysis and Results

Recall that one of the goals of the PIP study is to characterize the impact of immunosuppression regimens on lymphocyte compositions in renal transplant patients, the change of white blood cell compositions over time. Therefore, the percentage of lymphocyte compositions were recorded over time for renal transplant patients and control subjects. The PIP study began in 2005 and ended in 2011. The investigators enrolled 60 patients aged 18-59 years old who had renal transplantation at Emory University transplant center. They also enrolled 20 age-, sex- and race-matched healthy volunteers into control groups. All subjects enrolled in this study are followed for two years, and multiple blood samples were collected at baseline, 3 months, 6 months, 9 months, 12 months, 18 months and 24 months. The blood samples were analyzed with Flow cytometry. Right now we have data from 36 subjects. Among them, 28 subjects are renal transplantation recipients, and 8 subjects are health controls. The blood samples of these subjects were collected at baseline, 3 months, 6 months, 9 months, 12 months, and 18 months respectively, and were analyzed by Flow cytometry. The T lymphocytes (CD3+ cells) were broken into four subcategories based on the cell surface markers: CD4+CD8- (T Helper cells), CD4-CD8+ (cytotoxic T cells), CD4-CD8-, and CD4+CD8+. The percentage of these subcategories were recorded in the data set. Actually, we have 3-part compositions (k=3) for T lymphocytes

(CD3+ cells). These three components are CD4+CD8- (T Helper cells), CD4-CD8+ (cytotoxic T cells), and CD4-CD8-. The time points are 0,3,6,9,12 and 18 months (t=6).

If the percentages of all components dont add up exactly to 100% in the real data, we normalize data to make sure the summation is 100%. For a k-part composition $(z_1, z_2, \ldots, z_k)$, the observed summation $N = z_1 + z_2 + \ldots z_k$, which is not 100%. We recalculate the new percentage of each component based on the following equation.

$$z_{1new} = z_1/N, z_{2new} = z_2/N, \ldots, z_{knew} = z_k/N.$$

After the recalculation, the summation of new percentages $(z_{1new}, z_{2new}, \ldots, z_{knew})$ is exactly 100%.

The MCMC method is employed for model inference. Estimations of the parameter values are made over 12,000 MCMC iterations, in which the first 2,000 iterations were discarded as the "burn-in" phase. The point estimates of the parameters and their 95% credible regions are presented in Table 3.1.

The meaning of the parameter μ_1 is $log(\frac{z_1}{z_3})$ as we defined before. In this data analysis, μ_1 represents the logarithm of the ratio of the percentage of the first component (CD4+CD8- cells) to the percentage of the third component (CD4-CD8- cells). Similarly, μ_2 represents the log ratio of the percentage of the second component (CD4-CD8+ cells) to the percentage of the third component (CD4-CD8- cells). Figure 3.1 and Figure 3.2 show the point estimates and their 95% credible regions of μ_1 and μ_2 over times. From these figures, we can see that the ratio of the percentage of CD4+CD8- cells to the percentage of CD4-CD8- cells is decreased after the transplantation in the treatment group, while this ratio does not change much in the control group. The ratio of the percentage of CD4-CD8+ cells to the percentage of

CD4-CD8- cells does not change too much in both treatment group and control group. These results indicate the changes of cell compositions after the transplantation.

Table 3.1: Bayesian Inference for Flow Cytometry data

	Bayesian Approach		MLE Approach	
Parameter	Point estimate	95% Credible Region	Point estimate	95% CI
μ_{c11}	2.99	(2.66, 3.30)	3.03	(2.79, 3.26)
μ_{c12}	1.94	(1.36, 2.50)	1.95	(1.42, 2.44)
μ_{c21}	3.00	(2.65, 3.34)	3.08	(2.80, 3.36)
μ_{c22}	1.95	(1.54, 2.35)	1.95	(1.56, 2.34)
μ_{c31}	2.79	(2.32, 3.21)	2.93	(2.66, 3.20)
μ_{c32}	1.89	(1.44, 2.31)	1.91	(1.54, 2.28)
μ_{c41}	3.04	(2.63, 3.40)	3.16	(2.90, 3.42)
μ_{c42}	1.99	(1.65, 2.35)	1.94	(1.71, 2.17)
μ_{c51}	2.78	(2.08, 3.44)	3.07	(2.75, 3.39)
μ_{c52}	1.89	(1.29, 2.50)	1.81	(1.31, 2.31)
μ_{c61}	2.89	(2.21, 3.52)	3.14	(2.84, 3.44)
μ_{c62}	1.99	(1.35, 2.64)	1.91	(1.40, 2.42)
μ_{t11}	2.84	(2.66, 3.01)	2.86	(2.63, 3.10)
μ_{t12}	1.65	(1.33, 1.96)	1.68	(1.49, 1.87)
μ_{t21}	1.70	(1.51, 1.89)	1.71	(1.43, 1.99)
μ_{t22}	1.54	(1.31, 1.76)	1.54	(1.35, 1.73)
μ_{t31}	1.59	(1.35, 1.83)	1.60	(1.33, 1.87)
μ_{t32}	1.44	(1.20, 1.67)	1.43	(1.27, 1.60)
μ_{t41}	1.66	(1.45, 1.87)	1.67	(1.41, 1.93)
μ_{t42}	1.66	(1.46, 1.84)	1.66	(1.52, 1.79)
μ_{t51}	1.77	(1.38, 2.16)	1.79	(1.47, 2.11)
μ_{t52}	1.62	(1.27, 1.95)	1.62	(1.42, 1.82)
μ_{t61}	1.65	(1.27, 2.01)	1.66	(1.37, 1.95)
μ_{t62}	1.56	(1.19, 1.93)	1.56	(1.26, 1.87)

The interpretation of the parameters $\underline{\mu}$ is not easy to understand since $\underline{\mu}$ is in the log-ratio scale. Referring the the earlier notation, the Monte Carlo estimates for the expected values of composition proportions ($\underline{\theta}$) can also be obtained based on the samples from the posterior distribution (Iyengar and Dey [1998]).

For $j = 1, \ldots, k-1$,

$$\hat{\theta}_j = \frac{1}{B}\sum_{s=1}^{B} \frac{exp(\hat{\mu}_{j,s})}{1 + exp(\hat{\mu}_{1,s}) + \ldots + exp(\hat{\mu}_{k-1,s})} \tag{3.15}$$

while the estimate of the k^{th} component is

$$\hat{\theta}_k = \frac{1}{B}\sum_{s=1}^{B} \frac{1}{1 + exp(\hat{\mu}_{1,s}) + \ldots + exp(\hat{\mu}_{k,s})} \tag{3.16}$$

In these equations, B represents the number of MCMC iterations which the estimates are based on, and $\hat{\mu}_{j,s}$ represents the value of the unknown parameter μ_j at the s^{th} MCMC iteration. The point estimates for the expected values of the composition proportions are showed in Table 3.2.

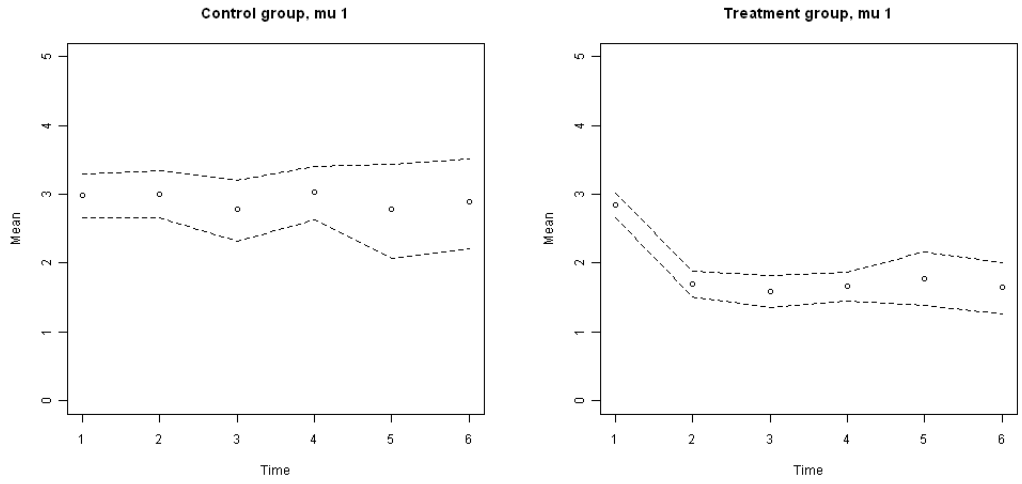
Table 3.2 shows that the percentage of CD4+CD8- cells is 73.28% at baseline in the treatment group, and is 71.17% at baseline in the control group. However, the percentage of CD4+CD8- cells drops to 49.21% after 3 months of transplantation in the treatment group, whereas the percentage is 71.31% at 3 months in the control group. From this table, we can see that the percentage of CD4+CD8- cells decreased after the transplantation in the treatment group, and the percentages of CD4-CD8+ cell and CD4-CD8- cell both increased in the treatment group. However, the ratio of CD4-CD8+ cells to CD4-CD8- cells is similar. This is reflected in the point estimates of μ_2 in table 2.1 which do not change much after transplantation in the treatment group. The cell compositions do not change much over time in the control group.

Figure 3.3 and 3.4 show the point estimates of cell compositions in the ternary diagrams for control group and treatment group respectively. A ternary diagram is a convenient way to display 3-part compositional data. From Figure 3.4, we can see that the cell compositions move from CD4+CD8- vertex toward the CD4-CD8+

vertex after the transplantation. According to the definition of the ternary diagram for compositional data, this indicates the increase of the percentage of CD4-CD8+ cells and the decrease of the percentage of CD4+CD8- cells. The figure also shows a little increase of the percentage of CD4-CD8- cells after transplantation. Figure 3.5 and 3.6 show the point estimate of cell composition in the ternary diagrams for the control group and the treatment group respectively. The point estimates of the cell compositions at each time are displayed separately. Figure 3.7 to Figure 3.12 show the point estimates and their 95% credible regions from time 1 to time 6. The red line represents the contour line of the highest 95% MCMC realizations for the control group, and the blue line represents the contour line of the highest 95% MCMC realizations for the treatment group. From these figures, we can see that the 95% credible regions of the control group and the treatment group overlap at baseline (Time=1), indicating no difference of cell compositions between the control group and the treatment group at the baseline. The 95% credible region of the treatment group separates from the 95% credible region of the control group after transplantation (at 3 months, 6 months and 9 months after baseline). These plots indicate the difference of cell compositions between the control group and the treatment group after transplantation.

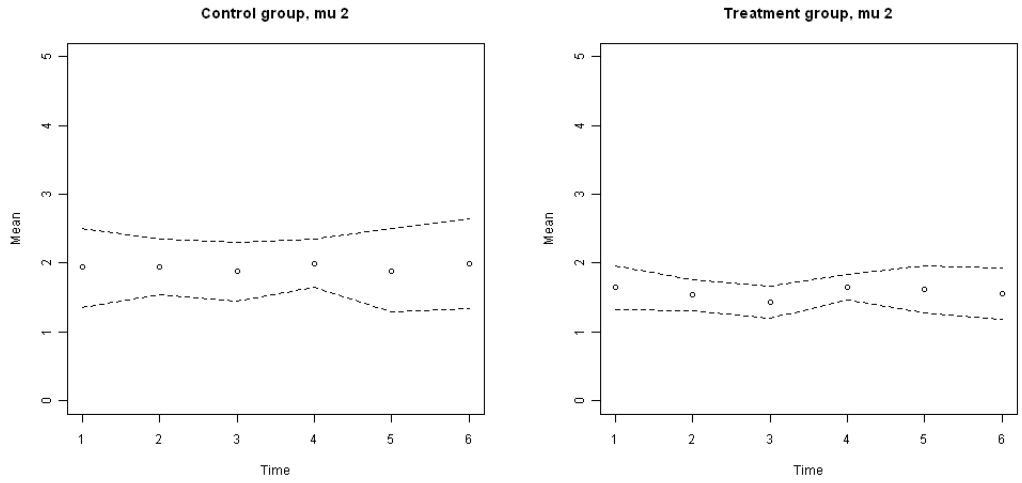
3.4.1 Model Diagnostics

It is crucial to check the model in Bayesian data analysis, because the Bayes prior-to- posterior inferences condition on the whole probability model, and can be very misleading when the model is poor. We used the methods we discussed in section 3.3.2 to evaluate the adequacy of the fit of our model to the data and the plausibility of the model.



(a) Control Group (b) Treatment Group

Figure 3.1: Point estimates and their 95% credible regions of μ_1



(a) Control Group (b) Treatment Group

Figure 3.2: Point estimates and their 95% credible regions of μ_2

First, we conduct a sensitivity analysis of our model. During the data analysis, we have tried several widely varying hyperparameters. For the hyperparameter a, c, the values of 0.1, 0.5 and 1.0 are tried in the MCMC inference. The resulting parameter

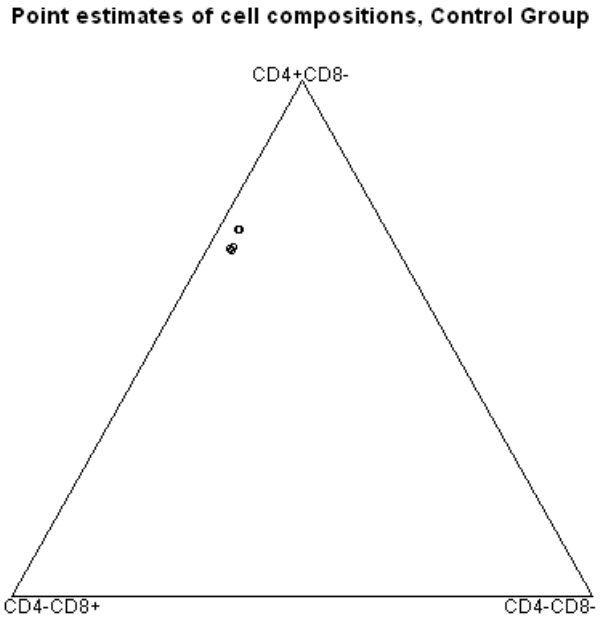


Figure 3.3: Graphical display of point estimates of cell compositions for control group at different time points in Ternary diagram

inferences are very similar even though the values of the hyperparameters changed widely. Table 3.3 shows the point estimates of the percentage of CD4+CD8- cells in the treatment group based on different values of the hyperparameters a and c. From this table, we can see that even the hyperparameters change from 0.1 to 1, the point estimates of the percentage of CD4+CD8- cells are very close. Therefore, the model inferences are not dependent on the values of the hyperparameters we chose. These results showed that the results of model inferences are data dependent.

Next, we use the posterior predictive p-value to evaluate the fit of the posterior distribution of our Bayesian model. We use the method of Gelman, Meng, and Stern [1996] to calculate the posterior predictive p-value. The discrepancy function we used

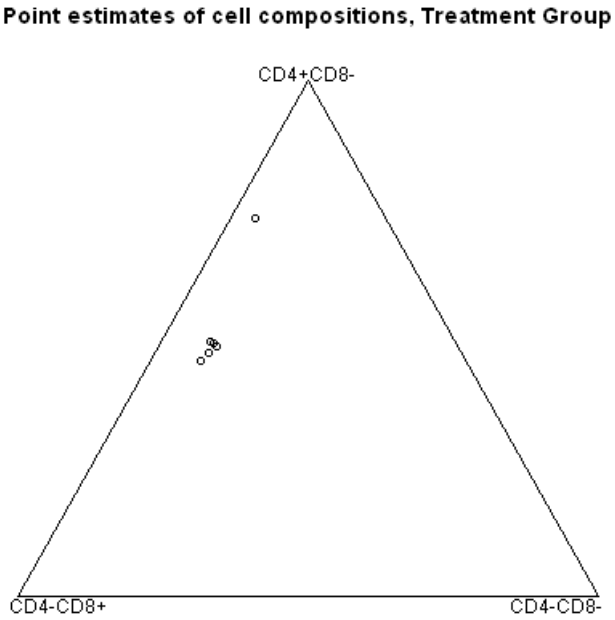


Figure 3.4: Graphical display of point estimates of cell compositions for treatment group at different time points in Ternary diagram

in this analysis is

$$D(z, \hat{\theta}) = \sum_{i=1}^{n} \frac{[z_i - E(z_i|\hat{\theta})]^2}{E(z_i|\hat{\theta})} \tag{3.17}$$

Based on 10,000 simulation iterations, the posterior predictive p-value, p_B is 0.188, which is the probability that the replicated data (z^{rep}) could be more extreme than the observed data (z^{obs}), as measured by the discrepancy function $D(z, \hat{\theta})$. Figure 3.13 shows the scatterplot of replicated vs. observed discrepancies (D(rep, θ) vs. D(obs, θ)) under the joint posterior distribution; the p-value is calculated as the proportion of points in the upper-left half of the plot. Figure 3.14 shows the histogram of 10,000 simulations from the difference of the replicated discrepancy (D(rep, θ)) and the

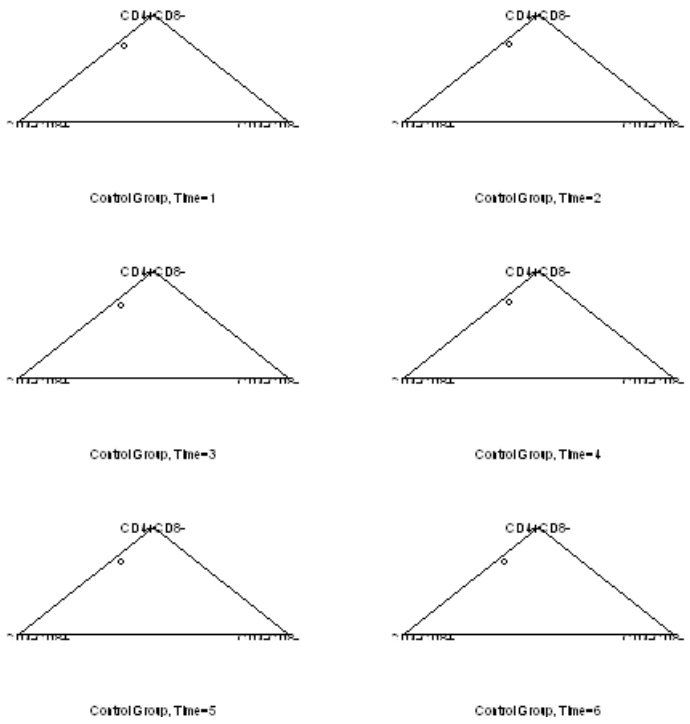


Figure 3.5: Graphical display of point estimate of cell compositions for control group in Ternary diagram

observed discrepancy (D(rep, θ) - D(obs, θ)). If the model is reasonable, the histogram should include 0. Figure 3.15 shows the scatterplot of 10,000 simulations from the difference of the replicated discrepancy (D(rep, θ)) and the observed discrepancy (D(rep, θ) - D(obs, θ)). Based on the result of posterior predictive p-value, p_B, and the histogram and scatterplots shown, we can conclude that there is no systematic differences between the replicated data generated under the model and the observed data. Therefore, our model fits the data well.

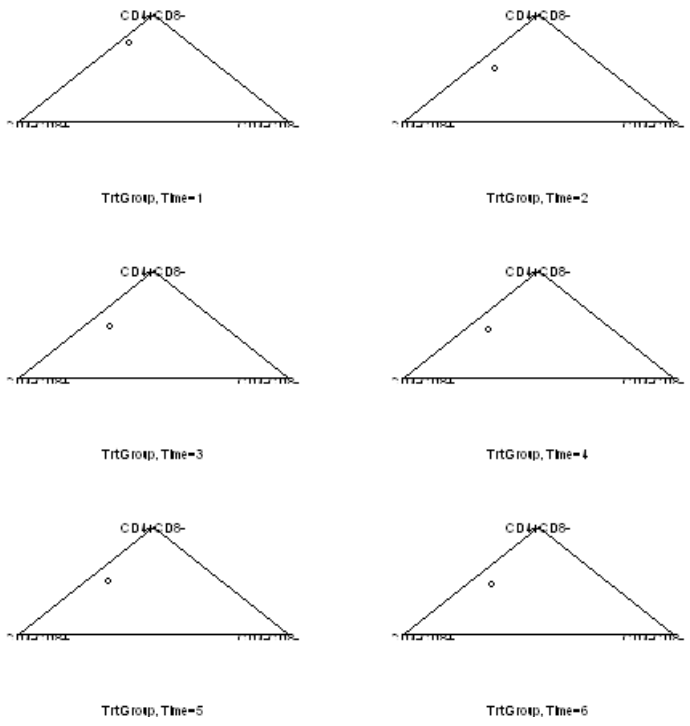


Figure 3.6: Graphical display of point estimate of cell compositions for treatment group in Ternary diagram

3.5 Simulation Study

Simulation studies were conducted to evaluate the performance of the proposed Bayesian method for repeated compositional data. Suppose we have n subjects for treatment group and control group respectively, and these subjects were followed for t time points. And for each sample, the composition has k components. We assume the composition vector for subject j, at time point t, $\underline{z}_{jt}$ is a k-part composition. Also, we generate a scalar covariate X, then the mean parameter $\underline{\mu}$ can depend on the explanatory variables, and can be replaced in the model by $\underline{\alpha}+\underline{\beta}X$. The covariate X takes the integer values between -10 and 10. In this simulation study, we set the

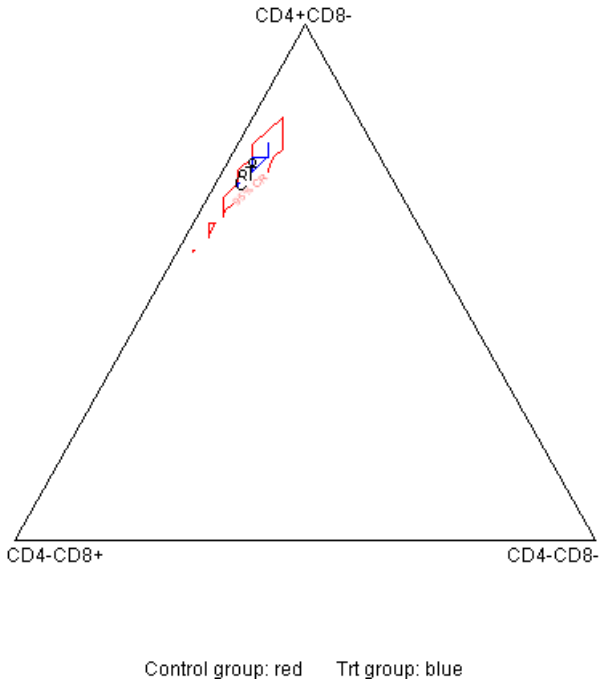


Figure 3.7: 95% credible regions for cell composition estimates, time=1

composition dimension, k=3, and the time points, t=3. We define the true values of $\underline{\alpha}_c = \underline{\alpha}_t = (2.00, 1.00)$, $\underline{\beta}_c = (0, 0)$ and $\underline{\beta}_t = (1.00, 0.50)$, respectively. We conduct 500 simulations on the sample size 500 in each group. For each simulated data set, 5000 Monte Carlo iterations were conducted, and the first 1000 iterations were used for "burn-in", and the subsequent 4,000 MonteCarlo realizations were collected for the posterior distribution inference. Simulation results of the sample size n=500 are presented in Tables 3.4-3.6. The true values of parameters, the point estimates, the standard deviations and the 95% coverage rate are displayed in the tables.

The simulation results show that the estimates based on the Bayesian approach model perform well with reasonably small bias and standard deviation.

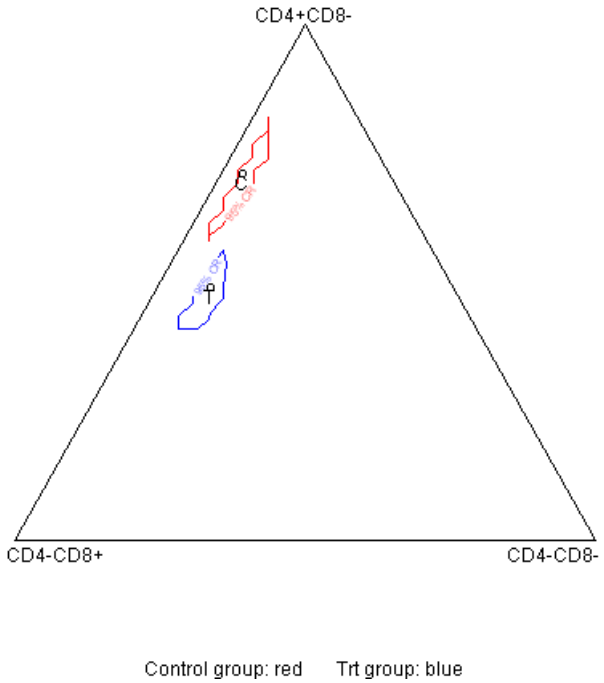


Figure 3.8: 95% credible regions for cell composition estimates, time=2

3.6 Discussion

Compositional data are expressed as a non-negative vector with unit-sum constraint. Because of the constraint, the sample space of compositional data is the Simplex space. Aitchison [1982, 1986] gave the fundamental works for compositional data analysis. He introduced the Logistic Normal distribution (LN) as an analysis tool for compositional data, and established its mathematical and statistical properties. Aitchison's methods rely on the additive logratio transformation to take the compositional data from the (k-1)-dimensional simplex space to the (k-1)-dimensional Euclidean space. Billheimer, Guttorp, and Fagan [2001] proposed a hierarchical model for compositional discrete data. His method combined Aithison's LN distribution and

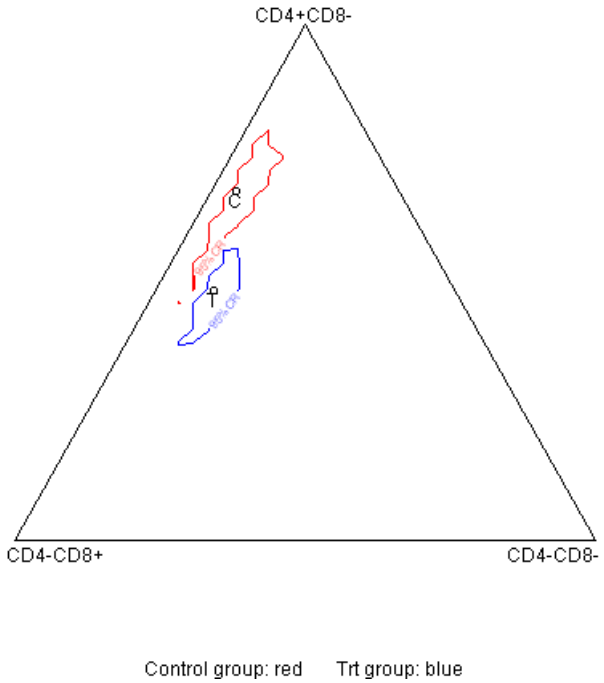


Figure 3.9: 95% credible regions for cell composition estimates, time=3

a conditional multinomial model. However, both Aitchison and Billheimer's methods deal with the compositional data at a single time point. In the medical studies or clinical trials, multiple measurements are obtained when the enrolled patients are followed by a period of time. Because the repeated measurements made over time on the same subject are typically correlated, either method can not be applied directly on the repeat-measured compositional data. Our proposed model extends Billheimer's hierarchical model to longitudinal/repeat-measures problems. The proposed model provides a tool to analyze the compositional data with repeated measurements. By using the algebra for compositions developed by Aichison and Billheimer, we can interpret the model parameter estimates and credible regions in terms of compositions.

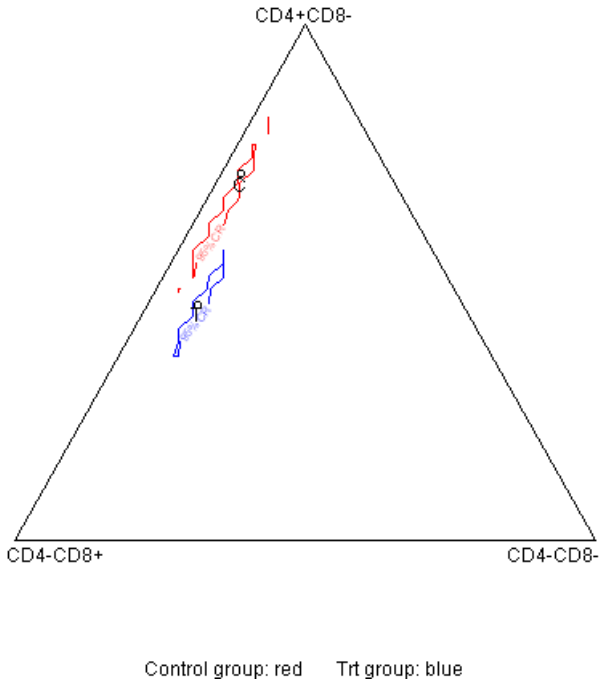


Figure 3.10: 95% credible regions for cell composition estimates, time=4

Since the proportions are the nature scale of measurement for composition data, interpretation in this way may help researchers have a better understanding from the statistical modeling results.

In this chapter, we developed a Bayesian approach for the analysis of the repeat-measured compositional data. Our results demonstrate that the Bayesian methodology can be used to analyzed repeat-measured compositional data. We use a Markov Chain Monte-Carlo method for model inference and show that the method is practical in high dimensional problems.

We use Aitchison's Logistic normal distribution to model the compositional data. This distribution provides a powerful approach for compositional data. Relying on

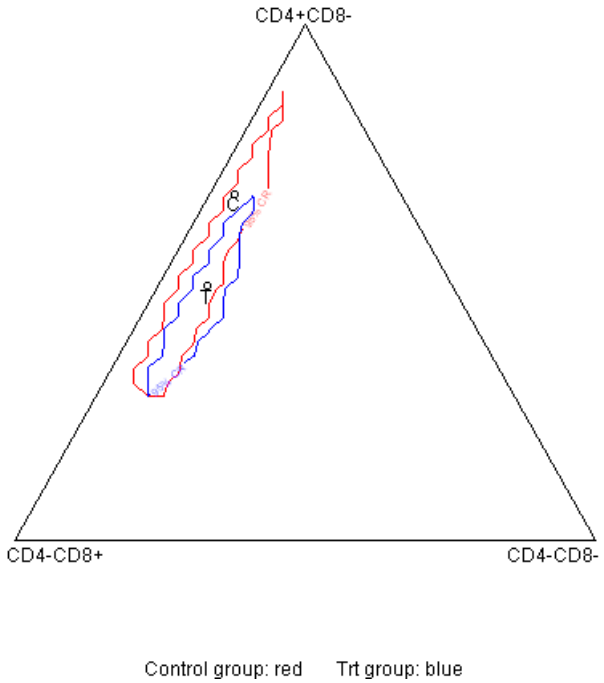


Figure 3.11: 95% credible regions for cell composition estimates, time=5

the additive logratio transformation, this approach transforms the compositional data into multivariate normal data for which powerful statistical methods have been developed. Also, this approach provides ability to describe the complicated variance-covariance structure between components of the compositional data. Although the logistic normal model has some strong properties for compositional data, it also has some weaknesses. A serious shortcoming is that the logistic normal distribution requires that all components must be positive. A zero component will change a k composition into a $(k-1)$ composition. Also, the additive logratio transformation is not defined if one or more components are zero either. This restriction of "no zero" may be a severe limitation in the application if one or more components are

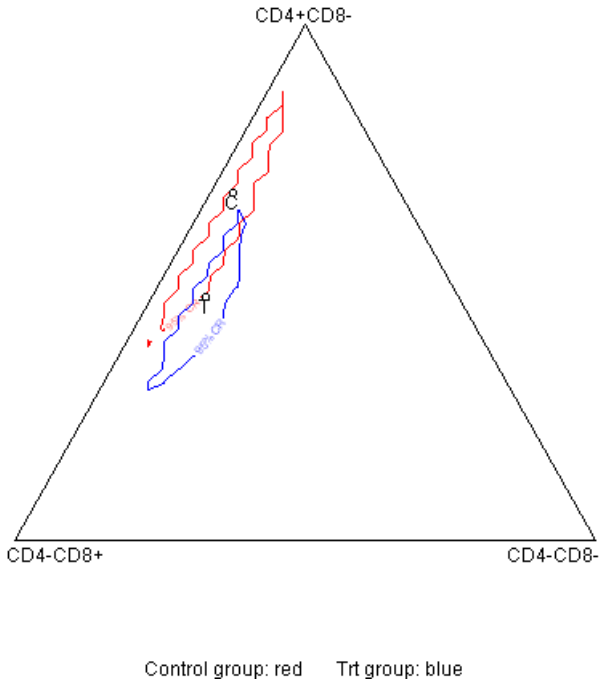


Figure 3.12: 95% credible regions for cell composition estimates, time=6

known to be absent, or inference of absence is important. Aitchison [1986] proposed a method that the zero value can be replaced by a small nonzero value to overcome the problem. However, his method is not subcompositionally coherent. Fry, Fry, and McLaren [2000], Martin-Fernandez, Barcelo-Vidal, and Pawlowsky-Glahn [2000] proposed independently a nonparametric replacement procedure for zero value problem.

If we have really small proportions in the data, we should pay attention to the small proportions, and should avoid picking the small proportion as the reference component (the last component, z_k) to assure the estimates are numerically stable. Aitchison's logistic normal distribution requires that all components are strictly positive because we can not take logarithms of zero. To deal with zero components, Aitchison proposed

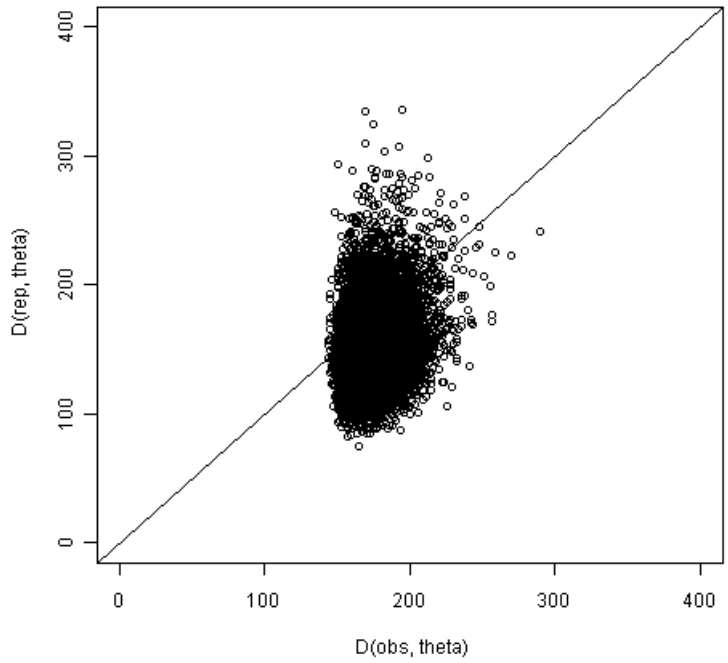


Figure 3.13: Scatterplot of replicated vs. observed discrepancies (D(rep, θ) vs. D(obs, θ)) under the joint posterior distribution; the p-value is estimated by the proportion of points above the $45^{\emptyset}$ line.

that the zero components can be replaced by some really small positive values. In this circumstance, he suggested that it will always be wise to perform a sensitivity analysis to determine the effects of the different small positive values on the conclusions of the analysis. For example, if we replaced zero components by 0.0005, we need also try other replacement values 0.001, 0.0025, 0.00001 and 0.000001 to see the change of parameters estimates, and to see if we could get the same conclusion (Aitchison [1986]). Thus, if we have really small values in the data, we need to pay attention to the numerical stability of our model.

There are alternative statistical models for compositional data. Some researchers proposed the use of the Liouville distribution (Smith and Rayens [2002], Iyenger and

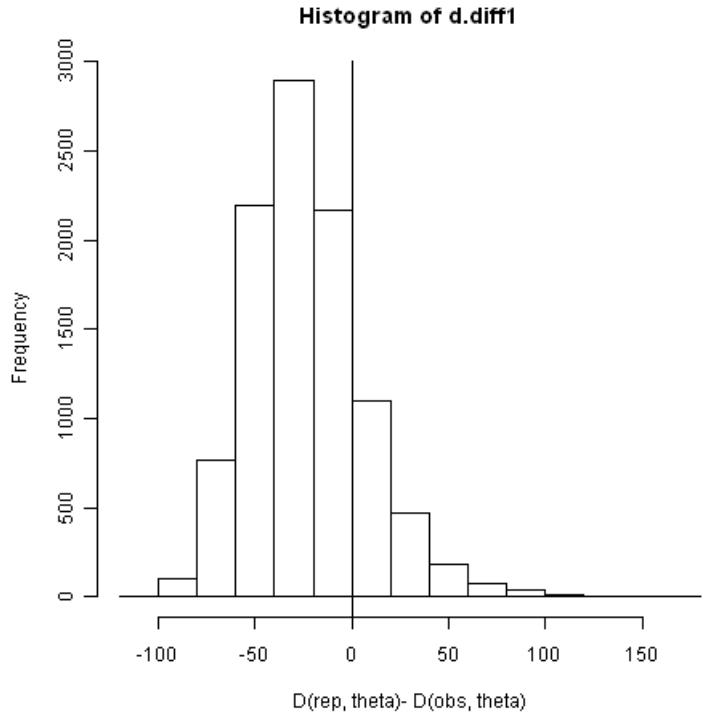


Figure 3.14: Histogram of 10,000 simulations from the difference of the replicated discrepancy (D(rep, θ)) and the observed discrepancy (D(rep, θ) - D(obs, θ)). Under the model, the histogram should include 0.

Dey [2002]. For detailed description of the Liouville and Dirichlet distributions, see Gupta and Richards [2001].) Barndorff-Nielsen and Jogensen [1991] proposed the S^- distribution. The S^- distribution is constructed from the conditional distribution of identical Gaussian random variables, given their sum. This distribution has attractive mathematical properties compared to the LN distribution. However, it suffers the same inflexibility as the Dirichlet distribution. The S- class of distributions is closed under marginalization, unlike the LN distribution. Stephens [1982] proposed the von Mises distribution for compositional data analysis. He used the square root to transform composition components. This model received little attention because of the complexity of the von Mises distribution, and also the the von Mises distribution

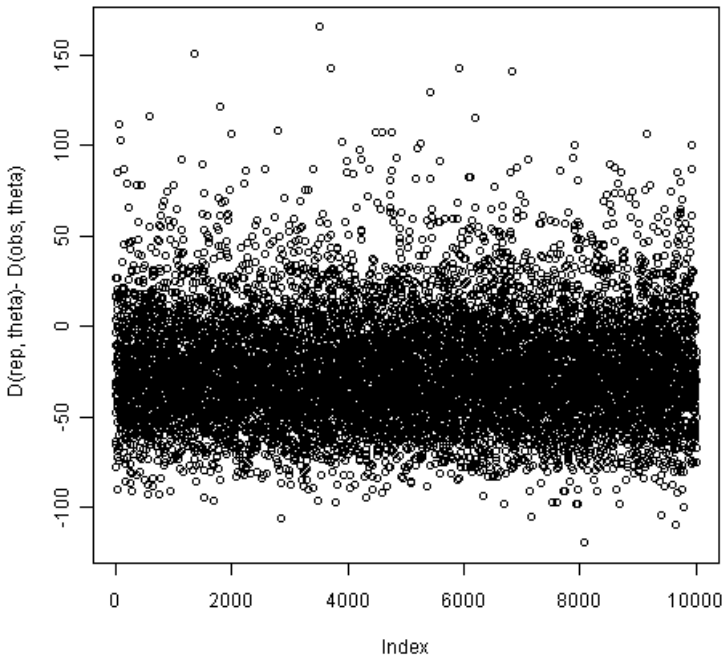


Figure 3.15: Scatterplot of 10,000 simulations from the difference of the replicated discrepancy (D(rep, θ)) and the observed discrepancy (D(rep, θ) - D(obs, θ)).

has difficulty to interpret the real world of the process.

Several researchers have extended Aitchison's standard Logistic normal model. Billheimer and Guttorp [1997] proposed a lattice spatial process for the error terms which can include spatial correlation for compositions with a spatial index. Tjelmeland and Lund [2003] proposed a multivariate geostatistical model in which they used a logistic Gaussian field for the spatial composition process. Brunsdon and Smith [1998] used a vector ARMA for modeling the error term in compositional time series analysis, and their method can induce serial correlation in the series of compositions from the repeated surveys. Abbitt and Breidt [2001] proposed a measurement error model, for soil composition.

Except for Brunsdon and Smith [1998], all the models presented were studied using a Bayesian inference approach. Because of the non-linear inverse ALR transformation, Bayesian method becomes more attractive comparing to the likelihood method. In this paper, The Bayesian approach and MLE method have the similar results. However, Bayesain is more attractive when we deal with more complicated settings. For example, Rayens and Srinivasan [1991a,b] extended Aitchison's method by incorporating the Box-Cox transformation as a generalization of the log-ratio function. In this situation, likelihood method becomes impractical. Also when we introduce covariates in the model, the complexity of the model increase rapidly. Therefore, Bayesian approach is more attractive when we deal with more complicated settings. We also use a Bayesian approach for our model inference.

Multivariate repeated measures data, or multivariate longitudinal data, are the data that multiple measurements are made over time on a collection of response variables on each subject or unit. In our motivation example, PIP study, the percentages of patients' white blood compositions were collected over time. Therefore, the data in this study are also multivariate repeated measures data, and the measurements on multiple response variables at each time point are compositional data. The sample space associated with the K-part compositions is the $(k-1)$ dimensional unit simplex (∇^{k-1}). By using the additive logratio transformation (ALR), we transformed the longitudinal compositional data into multivariate normal distribution data. We used the general multivariate linear model for the transformed data. For the composition with k components, and observed for t time points, the ALR transformed data follows a multivariate normal distribution with $t * (k-1)$ dimension. The general multivariate linear model accounts for all the potential sources of variability that have an impact on the covariance among repeated measurements on the same individual.

That is, the model does not distinguish between-subject and within-subject sources of variability. Therefore, we need assume only that the variance-covariance matrix Ω is an arbitrary positive definite matrix. We use a Bayesian approach for model inference, and it is straightforward to define the prior distribution of the unstructured variance-covariance marix Ω as an inverse Wishart distribution, the most common prior for a covariance matrix. However, this prior is restrictive and lacks flexibility. Some researchers have proposed some prior distributions for a covariance matrix with more flexibility (Leonard and Hsu [1992], Daniels and Kass [1999], Barnard, McCulloch, and Meng [2000]). However, since these prior distributions lack conjugacy, these priors may have computational difficulty during parameter estimations. Yang and Berger [1994] proposed a noninformative prior, and Everson and Morris [2000] proposed a constrined Wishart prior.

For a K-part composition with t time points, we have a $t*(k-1) \times t*(k-1)$ variance-covariance matrix Ω. For an unstructured Ω, the number of parameters of Ω is $(t*(k-1) \times (t*(k-1)+1))/2$. When the number of parameters becomes large, estimation of the parameters in the covariance matrix becomes computationally burdensome. To overcome this problem, some dimension-reduction technics can be used in future research. The covariance matrix Ω can be re-parameterized as

$$T\Omega T' = D$$

where T is a unique unit lower triangular matrix having 1's on its diagonal and D is diagonal with positive diagonal entries. The (T, D) re-parametrization covariance matrix provides flexibility in specifying prior beliefs and also offers conditional conjugacy for computations. Daniels and Pourahmadi [2002] proposed a conditionally conjugate prior distributions for covariance matrices by using the Cholesky decom-

position and the unconstrained re-parametrization technics. Their priors can shrink covariance matrix toward a particular structure with considerable flexibility.

The general multivariate linear model accounts for all the potential sources of variability that have an impact on the covariance among repeated measurements on the same individual. That is, the model does not distinguish between-subject and within-subject sources of variability. The general multivariate linear model is appropriate for balanced longitudinal designs. When each subject is observed at the same t time points and there is no theoretical or empirical assumption for a special covariance structure, we only need to assume the covariance matrix is a positive definite covariance matrix. This approach is attractive even when some observations are missing or the design is moderately unbalanced across subjects. However, when the data are highly unbalanced (e.g. subjects are observed at different sets of times), or incomplete (e.g. missing data), the general multivariate model with unrestricted covariance structure may not be appropriate for this kind of data set. A powerful approach to modeling unbalanced longitudinal data is the linear mixed model (Laird and Ware [1982]). There is no requirement for the balance in the data in the linear mixed model, and the model can analyze the between and within individual variation. Comparing to the general multivariate model, the major limitation of the linear mixed model is the requirement of the special form assumed for the covariance matrix.

Table 3.2: Point estimates for the composition proportions

		Bayesian Approach			MLE Approach		
Group	Time(Months)	CD4+CD8-	CD4-CD8+	CD4-CD8-	CD4+CD8-	CD4-CD8+	CD4-CD8-
Control	0	0.7117	0.2518	0.0365	0.7213	0.2439	0.0348
	3	0.7131	0.2511	0.0358	0.7305	0.2361	0.0334
	6	0.6780	0.2792	0.0428	0.7071	0.2552	0.0377
	9	0.7109	0.2546	0.0345	0.7472	0.2211	0.0317
	12	0.6712	0.2856	0.0432	0.7527	0.2124	0.0349
	18	0.6747	0.2865	0.0388	0.7498	0.2179	0.0323
Treatment	0	0.7328	0.2241	0.0431	0.7325	0.2256	0.0419
	3	0.4921	0.4179	0.0900	0.4946	0.4163	0.0891
	6	0.4858	0.4153	0.0989	0.4884	0.4132	0.0984
	9	0.4585	0.4547	0.0868	0.4588	0.4545	0.0867
	12	0.4936	0.4224	0.0840	0.4964	0.4206	0.0830
	18	0.4732	0.4352	0.0916	0.4770	0.4325	0.0905

Table 3.3: Sensitivity Analysis: Point estimates of the percentage of CD4+CD8- cell in the treatment group based on different hyperparameters a and c

	Months					
	0	3	6	9	12	18
a=0.1, c=0.1	0.7354	0.4858	0.4778	0.4569	0.4832	0.4589
a=0.5, c=0.1	0.7328	0.4921	0.4858	0.4585	0.4936	0.4732
a=1.0, c=0.1	0.7323	0.4933	0.4872	0.4586	0.4951	0.4750
a=0.1, c=0.5	0.7331	0.4838	0.4772	0.4565	0.4828	0.4581
a=0.5, c=0.5	0.7324	0.4920	0.4855	0.4587	0.4932	0.4733
a=1.0, c=0.5	0.7322	0.4934	0.4871	0.4592	0.4944	0.4742
a=0.1, c=1.0	0.7302	0.4827	0.4765	0.4561	0.4814	0.4580
a=0.5, c=1.0	0.7315	0.4917	0.4854	0.4582	0.4925	0.4726
a=1.0, c=1.0	0.7315	0.4928	0.4865	0.4584	0.4944	0.4743
a=10, c=10	0.7307	0.4934	0.4866	0.4582	0.4959	0.4747

Table 3.4: Simulation results based on 500 simulated datasets, sample size=500 in each group, Time point=1

Parameter	**Time Point=1**				
	Bayesian Approach			MLE Approach	
	True value	Estimates	95% Coverage	Estimates	95% Coverage
α_{c1}	2.00	2.00(0.050)	0.880	2.00(0.048)	0.893
α_{c2}	1.00	1.00(0.050)	0.942	1.00(0.051)	0.927
β_{c1}	0.00	0.0087(0.18)	0.884	0.0076(0.14)	0.902
β_{c2}	0.00	0.0023(0.17)	0.854	0.0045(0.21)	0.864
α_{t1}	2.00	2.00(0.028)	0.948	2.01(0.021)	0.939
α_{t2}	1.00	1.00(0.028)	0.926	1.00(0.019)	0.918
β_{t1}	1.00	1.00(0.0054)	0.934	1.02(0.0042)	0.924
β_{t2}	0.50	0.50(0.0053)	0.952	0.51(0.0039)	0.946

[a]Standard Deviations of the point estimates are shown in the bracket.

Table 3.5: Simulation results based on 500 simulated datasets, sample size=500 in each group, Time point=2

Parameter	**Time Point=2**				
	Bayesian Approach			MLE Approach	
	True value	Estimates	95% Coverage	Estimates	95% Coverage
α_{c1}	2.00	2.00(0.051)	0.944	2.00(0.049)	0.918
α_{c2}	1.00	1.00(0.050)	0.928	1.01(0.052)	0.931
β_{c1}	0.00	0.0099(0.19)	0.834	0.0061(0.21)	0.894
β_{c2}	0.00	0.0011(0.19)	0.858	0.0028(0.23)	0.849
α_{t1}	2.00	2.00(0.029)	0.954	2.00(0.036)	0.938
α_{t2}	1.00	1.00(0.028)	0.928	1.00(0.021)	0.919
β_{t1}	1.00	1.00(0.0051)	0.916	1.00(0.0043)	0.926
β_{t2}	0.50	0.50(0.0054)	0.946	0.50(0.0053)	0.931

[a]Standard Deviations of the point estimates are shown in the bracket.

Table 3.6: Simulation results based on 500 simulated datasets, sample size=500 in each group, Time point=3

Parameter	**Time Point=3**				
		Bayesian Approach		MLE Approach	
	True value	Estimates	95% Coverage	Estimates	95% Coverage
α_{c1}	2.00	2.00(0.050)	0.926	2.00(0.047)	0.913
α_{c2}	1.00	1.00(0.050)	0.942	1.00(0.054)	0.939
β_{c1}	0.00	-0.0052(0.18)	0.872	0.0074(0.14)	0.864
β_{c2}	0.00	-0.0022(0.18)	0.886	0.0045(0.16)	0.873
α_{t1}	2.00	2.00(0.029)	0.936	2.00(0.021)	0.941
α_{t2}	1.00	1.00(0.030)	0.928	1.00(0.023)	0.919
β_{t1}	1.00	1.00(0.0056)	0.952	1.00(0.0042)	0.942
β_{t2}	0.50	0.50(0.0055)	0.968	0.50(0.0059)	0.927

[a]Standard Deviations of the point estimates are shown in the bracket.

Chapter 4

MEASUREMENT ERROR IN GENERAL MULTIVARIATE LINEAR MODEL

4.1 Background

In medical studies, some variables of interest are difficult or expensive to obtain. Instead, surrogate variables which can be obtained relatively easy and inexpensive are recorded and used by researchers in real-life situation. However, these surrogate variables may contain measurement errors. Examples include blood pressure, cholesterol serum level, hormone level, and so on. Another example is radioimmunoassay. In radioimmunoassay, the variable of interest is the molecular level in the sample, but the observed value is the radioactive count.

Measurement error problems in predictor variables have recently received extensive attention by researchers. There are many papers about how to correct measurement errors in regressors in different applications (Carroll and Stefanski [1990]; Brown and Fuller [1990]; Byar and Gail [1989]; Fuller [1987]). However, there is little literature about the problem with measurement errors in the response variable. The main reason is that the problem can be handled with standard methodology if the measurement errors in response variable are additive errors in which the observed values = true value + error, where the error has mean 0. When the measurement errors are additive with constant variance, the errors can be ignored in regression analysis because they can be thought as an extra variance component. However, for many methods, the measurement error models are complex. The additive errors assumption may not be appropriate. If the measurement error in the response variable is not unbiased, ignoring the response measurement errors will lead to biased estimates of the regression parameters.

4.1.1 Measurement error in the response in the General linear model

Suppose we want to fit a linear regression model

$$Y = \beta_0 + \beta_x X + \sigma_\epsilon^2$$

The variance of Y is σ_ϵ^2. If Y were observed, we can easily estimate $\underline{\beta}$. Now suppose the true value Y is unobserved, and the observed value we have is $U = Y + V$, where V is additive with mean zero and constant variance σ_v^2. Then we can simply add variability, and then U has the same linear regression function as does Y, but the variance of U becomes $\sigma_\epsilon^2 + \sigma_v^2$.

From the above example, we can see that the response measurement error increases the variance of the fitted lines without causing bias in the linear regression if the response measurement error was unbiased and homoscedastic.

Carroll et al. [2006] made a stronger conclusion about unbiased homoscedastic response measurement error. He pointed out that, in linear or nonlinear regression that has homoscedastic errors about the true line, the only effects of unbiased homoscedastic response measurement error is increasing the variance of the model, and decreasing the power for detecting effects.

However, if the response measurement error is not unbiased, ignoring the response measurement error will lead to biased estimates of the regression parameters. For example, suppose we want to fit a linear regression model

$$Y = \beta_0 + \beta_x X + \sigma_\epsilon^2$$

The mean of true value Y should be $\beta_0 + \beta_x X$ and the variance of Y is σ_ϵ^2. If the true

value Y is unobserved, and we only have the observed value U, where U given (Y, X) follows a normal linear model with mean $\gamma_0 + \gamma_1 Y$ and constant variance σ_v^2. Then U is biased in this model, and U follows a normal linear regression model with mean $\gamma_0 + \beta_0\gamma_1 + \gamma_1\beta_x Y$, and the variance of U becomes $\sigma_v^2 + \gamma_1^2\sigma_\epsilon^2$. Therefore, ignoring measurement error in this case will get the estimates $\gamma_1\beta_x$ instead of β_x.

The straightforward solution to the biased homoscedastic response measurement error is adjusting U to unbiased. We can use $(U - \gamma_0)/\gamma_1$. The problem here is to obtain information about parameters of measurement error model (γ_0, γ_1). Buonaccorsi [1991, 1996] and Buonaccorsi and Tosteson [1993] proposed methods to adjust biased measurement error in general linear model. They use independent, external calibration data to estimate the parameters in the measurement error model, and develop the Pseudo-maximum likelihood estimators and their asymptotic properties under normality assumptions.

4.1.2 Non-linear response measurement error in linear model

However, the additive measurement error model is not appropriate in some situations. The measurement error models are more complicated in many measuring methods(Buonaccorsi [1989, 1990a,b, 1991], Buonaccorsi and Tosteson [1993], Buonaccorsi [1996], Carroll, Gail, and Lubin [1993], Pepe [1992], Rosner, Spiegelman, and Willett [1990], Tosteson, Stefanski, and Schafer [1989]). For example, a nonlinear measurement error model is used in radioimmunoassay. Buonaccorsi and Tosteson [1993] described pseudo-maximum likelihood estimators and their asymptotic properties when the nonlinear measurement error model was integrated in the general linear model.

The specific example Buonaccorsi and Tosteson [1993] described involved a comparison of serum neopterin levels between human immunodeficiency virus (HIV)-

positive and HIV-negative individuals. True neopterin levels can not be observed and are measured through a radioimmunoassay, and the observed value is a radioactive count. In this example, the measurement error model is nonlinear, a four-parameter logistic model

$$g(y,\gamma) = \gamma_1 + \frac{\gamma_2 - \gamma_1}{1 + (exp(Y)/\gamma_3)^{\gamma_4}}$$

was used to fit the calibration curves.

Buonaccorsi and Tosteson [1993] used likelihood-based methods in this nonlinear measurement error in response problem. To avoid the computational problems associated with full maximum likelihood estimation, pseudo-maximum likelihood was used by Buonaccorsi and Tosteson [1993]. Let $\hat{\underline{\theta}}$ denote the MLE of $\underline{\theta}$ obtained from the independent, external calibration data. The pseudo-MLE for $\underline{\xi}$, denoted $\hat{\underline{\xi}}$, maximizes the likelihood function $L(\underline{\xi}, \hat{\underline{\theta}}) = \prod_{i=1}^{n} f_{U_i}(u_i|\underline{x}_i; \underline{\xi}, \hat{\underline{\theta}})$ in $\underline{\xi}$. With this approach, EM algorithm (Dempster, Laird, and Rubin [1977]) can be used to estimate the pseudo-MLE for $\hat{\underline{\xi}}$. Buonaccorsi and Tosteson [1993] provided the detailed EM algorithm to get the pseudo-MLE if the true values Y follows a normal linear model.

Under suitable conditions, both the full MLE, $\hat{\xi}_F$, and the pseudo-MLE, $\hat{\underline{\xi}}$, are consistent and asymptotically normal. Therefore, the asymptotic covariance of the pseudo-MLE $\hat{\underline{\xi}}$ can be expressed as

$$A(\hat{\underline{\xi}}) = I_U(\underline{\xi})^{-1} + I_U(\underline{\xi})^{-1} I_U(\underline{\xi}, \underline{\theta}) I_C(\underline{\theta})^{-1} I_U(\underline{\theta}, \underline{\xi}) I_U(\underline{\xi})^{-1}$$

where $I_U(\underline{\xi})$ is the information matrix for $\underline{\xi}$ from the main data, $I_C(\underline{\theta})$ is the information matrix for $\underline{\theta}$ from the calibration data.

4.1.3 General likelihood methods for response measurement error

Let $f_{U|Y,X}(u|y,x,\gamma)$ denote the density or mass function for U given (Y,X). U is a surrogate response if its distribution depends only on the true response, that means $f_{U|Y,X}(u|y,x,\gamma) = f_{U|Y}(u|y,\gamma)$. All the models we considered in this dissertation are for surrogate responses only.

The likelihood function for the observed response variable U is

$$f_{U|X}(u|x,\underline{\beta},\gamma) = \int_y f_{Y|X}(y|x,\underline{\beta}) f_{U|Y}(u|y,\gamma)dy$$

The likelihood function shows that if there is no relationship between the true response Y and the predictors X, then neither is here a relationship between the observed response U and the predictors X.

4.1.4 General Multivariate Linear Model

Longitudinal data is obtained when subjects are followed over a period of time, and for each subject, some variables are measured at multiple time points. The primary goal of a longitudinal study is to characterize the change in response over time and the factors that affect the change. Because of the repeated measures on individuals, the researchers can capture within-individual change. In our motivation example, the PIP study, the white blood cell counts from the patients' blood samples were measured over a period of time. We assume that the true cell count values should follow a normal distribution since the cell counts range from several hundreds to several thousands. Therefore, longitudinal analysis should be used because of the correlation of repeated measurements in one individual. One approach for analyzing

longitudinal data is the general multivariate model.

In the longitudinal data analysis the data are assumed to be independent over individual units, but to be correlated over time for a given individual unit. Standard references for longitudinal linear model were introduced by Hsiao [1986], Diggie, Liang, and Zeger [1994], Baltagi [1995] and Fitzmaurice, Laird, and Ware [2004].

In the general multivariate linear model, the mean response vector is

$$E(Y_i) = X_i\beta,$$

where the response vector, Y, is assumed to follow a multivariate normal distribution with covariance matrix

$$Cov(Y_i) = \Sigma_i = \Sigma_i(\eta),$$

where η is a $q \times 1$ vector of covariance parameters.

The log-likelihood function is

$$L = -\frac{K}{2}log(2\pi) - \frac{1}{2}\sum_{i=1}^{N} log|\Sigma_i| - \frac{1}{2}\left[\sum_{i=1}^{N}(y_i - X_i\beta)'\Sigma_i^{-1}(y_i - X_i\beta)\right],$$

where $K = (\sum_{i=1}^{N} n_i)$ is the total number of observation.
The maximum likelihood (ML) estimate of β is

$$\hat{\beta} = \left[\sum_{i=1}^{N}(X_i'\hat{\Sigma}_i^{-1}X_i)\right]^{-1}\sum_{i=1}^{N}(X_i'\hat{\Sigma}_i^{-1}y_i),$$

where $\hat{\Sigma}_i$ is the maximum likelihood (ML) estimate of $\Sigma_i(\eta)$. The MLE of η is obtained by maximizing the log-likelihood with respect to η. In general, the equations to solve η are non-linear, and need an iterative technique, such as the Newton-Raphson algorithm, to obtain the ML estimate of η.

The general multivariate linear model accounts for all the potential sources of variability that have an impact on the covariance among repeated measurements on the same individual. That is, the model does not distinguish between-subject and within-subject sources of variability. The general multivariate linear model is appropriate for balanced longitudinal designs.

4.2 Modeling

Our aim is to extend the measurement error in response in general linear model by Buonaccorsi to the general multivariate linear measurement error model.

Now we consider the longitudinal data $y_{i1}, \ldots, y_{it}$ as the true values over time periods t. The true value variable $\underline{y}_i$ is a $t \times 1$ vector following a multivariate normal distribution with t dimensions, and $\underline{y}_i|\underline{x}_i \sim f(\underline{y}_i|\underline{x}_i; \underline{\xi}), i = 1, \ldots, n$. Our main interest is the multivariate linear model

$$\underline{Y}_i = \underline{x}_i\underline{\beta} + \epsilon_i$$

where the ϵ_i are the i.i.d t-dimensional normal with mean $\underline{0}$ and covariance Σ. Then we have

$$E(\underline{Y}_i) = \underline{X}_i\beta,$$

and

$$Cov(\underline{Y}_i) = \Sigma_i = \Sigma_i(\underline{\eta}),$$

where η is a $q \times 1$ vector of covariance parameters.

The log-likelihood function is

$$L = -\frac{K}{2}log(2\pi) - \frac{1}{2}\sum_{i=1}^{N} log|\Sigma_i| - \frac{1}{2}\left[\sum_{i=1}^{N}(\underline{y}_i - \underline{X}_i\beta)'\Sigma_i^{-1}(\underline{y}_i - \underline{X}_i\beta)\right],$$

where $K = (\sum_{i=1}^{N} n_i)$ is the total number of observation.

In this model, our main parameters of interest are $\underline{\xi} = (\underline{\beta}, \underline{\Sigma}(\eta))$. The goal of model inference is estimating $\underline{\xi}$ when the true values $\underline{Y}'s$ are not observable. The independent variables $\underline{x}_i$ are assumed fixed and observed without measurement error. Therefore, our model has measurement error in the dependent variable $\underline{Y}$ only.

The model is completed by adding the measurement error structure. Suppose the true value y_{it} is unobservable, and we have observed values u_{it} with measurement error, where u_{it} is the measured value of true value y_{it}. The measurement error model is defined as the conditional distribution of U given $Y = y$. We use independent, external calibration data to estimate the parameters of measurement error model. For the i^{th} individual unit at time point t, given $Y_{it} = y_{it}$, the observed value U_{it} has the density function $f_{U|Y}(u_{it}|y_{it};\underline{\theta})$. Here we assume the measurement error model is a normal nonlinear regression model with constant variance which has

$$U|Y = y \sim N(g(y,\underline{\gamma}), \tau^2) \tag{4.1}$$

The function $g(y,\gamma)$ is the calibration function with parameters $\underline{\theta} = (\underline{\gamma}, \tau)$. Then the marginal density function of observed value U is

$$f_{U_i}(\underline{u}_i|\underline{x}_i;\underline{\xi},\underline{\theta}) = \int_y f_{U|Y}(\underline{u}_i|\underline{y}_i;\underline{\theta}) f_Y(\underline{y}_i|\underline{x}_i;\underline{\xi})dy \tag{4.2}$$

If the $\underline{\theta}$ is unknown, the main parameters of interest $\underline{\xi}$ are not identifiable in

most settings. Therefore, we need either the information of $\underline{\theta}$ or the calibration data to estimate $\underline{\theta}$ to make $\underline{\xi}$ identifiable. Assume we have independent, external calibration data available. In the calibration data $(U_k^*, Y_k^*, k = 1, \ldots, m)$, the Y^* are fixed constants (observable and based on standards) and the U^* are observable random variable with U_k^* have the density function $f_{U|Y}(u_k|Y_k^*; \underline{\theta})$. The data in the independent, external calibration dataset are all at one time point, so we assume the measurement error model and its parameters are the same over time. By using the external calibration data, we make an assumption of transportability characteristic when transporting measurement error model to the main data (Carroll, Ruppert, A., and M. [2006]). Let $\hat{\underline{\theta}}$ be the estimator of $\underline{\theta}$ obtained from the calibration data, and we assume $\underline{\theta}$ is identifiable from the independent, external calibration data.

4.3 Model Inference

4.3.1 Point Estimation Procedure by EM Algorithm

For each subject i, let $\underline{Y}_i = Y_{i1}, \ldots, Y_{it}$ be the true values over time periods t. The true value variable $\underline{Y}_i$ is a $t \times 1$ vector following a multivariate normal distribution with t dimensions. $\underline{U}_i = U_{i1}, \ldots, U_{it}$ are the observed variable with measurement errors.

The full likelihood of the general multivariate linear measurement error model is

$$L(\underline{\xi}, \underline{\theta}) = L_1(\underline{\xi}, \underline{\theta}) \cdot L_2(\underline{\theta}) \tag{4.3}$$

where $L_1(\underline{\xi}, \underline{\theta}) = \prod_{i=1}^{n} f_{\underline{U}}(\underline{u}_i|\underline{x}_i; \underline{\xi}, \underline{\theta})$ is the likelihood from the main data $(\underline{U}_1, \ldots, \underline{U}_n)$ and $L_2(\underline{\theta}) = \prod_{j=1}^{J} f_{U|Y}(u_j^*|y_j^*; \underline{\theta})$ is the likelihood from the calibration data. The full maximum likelihood estimator (MLE) can be obtained by maximizing $L(\underline{\xi}, \underline{\theta})$ simul-

taneously in $\underline{\xi}$ and $\underline{\theta}$. However, if the measurement error model is non-linear, the full MLE is difficult to obtain. Standard methods such as Newton-Raphson will require the evaluation of many numerical integrations at each iteration. To avoid the computational problems, we can use the pseudo-maximum likelihood estimation instead of the full maximum likelihood estimation. In the pseudo-maximum likelihood method, we use the MLE of $\underline{\theta}$, $\hat{\underline{\theta}}$, obtained from the calibration data, and plug in the likelihood function of the main data, and obtain the pseudo-MLE of the interested parameters $\underline{\xi}$ by maximizing $L_1(\underline{\xi}, \hat{\underline{\theta}}) = \prod_{i=1}^{n} f_U(u_i|\underline{x}_i; \underline{\xi}, \hat{\underline{\theta}})$ in $\underline{\xi}$.

The EM algorithm (Dempster, Laird, and Rubin [1977]) can be used to estimate the pseudo-MLE for $\underline{\xi}$. In the EM-algorithm, the true values $\underline{Y}$'s are not observable and are treated as the missing values, and the complete data likelihood is

$$L_c(\underline{\xi}|\underline{u}, \underline{y}) = \prod_{i=1}^{n} f_Y(\underline{y}_i|\underline{x}_i; \underline{\xi}) f_{U|Y}(\underline{u}_i|\underline{y}_i; \hat{\underline{\theta}}) \tag{4.4}$$

where $\underline{y}_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{it} \end{pmatrix}$, is the true values of individual unit i at time 1 to t, and $f_Y(\underline{y}_i|\underline{x}_i; \underline{\xi})$ is the density function of multivariate normal distribution.

Let $Q(\underline{\xi}|\underline{\xi}^*) = E(logL_c(\underline{\xi}|\underline{u}, \underline{Y})|\underline{U} = \underline{u}, \underline{\xi}^*)$, and let $\hat{\underline{\xi}}_{(a)}$ denote the estimate of $\underline{\xi}$ after a steps. Then we need to find $Q(\underline{\xi}|\hat{\underline{\xi}}_{(a)})$ at the E-step and maximize $Q(\underline{\xi}|\hat{\underline{\xi}}_{(a)})$ at the M-step. Given the independence assumption, $Q(\underline{\xi}|\hat{\underline{\xi}}_{(a)}) = E(logL_c(\underline{\xi}|\underline{u}, \underline{Y})|\underline{U} = \underline{u}, \hat{\underline{\xi}}_{(a)})$ equals

$$\sum_{i=1}^{n} E(logf_Y(Y_i|\underline{x}_i; \underline{\xi})|U_i = u_i, \hat{\underline{\xi}}_{(a)}) + \sum_{i=1}^{n} E(logf_{U|Y}(u_i|Y_i; \hat{\underline{\theta}})|u_i, \hat{\underline{\xi}}_{(a)}) \tag{4.5}$$

Please note that the second term does not involve $\underline{\xi}$, and it will not be involved

in the M-step either. Thus there is no need to calculate it at the E-step. Let

$$\hat{f}_{(a)}(\underline{y}_i|\underline{x}_i,\underline{u}_i) = f_{U|Y}(\underline{u}_i|\underline{y}_i;\hat{\underline{\theta}})f_Y(\underline{y}_i|\underline{x}_i;\hat{\underline{\xi}}_{(a)})/f_U(\underline{u}_i|\underline{x}_i;\hat{\underline{\xi}}_{(a)},\hat{\underline{\theta}}) \tag{4.6}$$

denote the estimate of the conditional density of $\underline{Y}_i$ given $\underline{U}_i = \underline{u}_i$ at each a^{th} step. The estimates of $\underline{\xi}$ can be derived by the following EM algorithm.

1. Assign starting values to $\underline{\xi}$.

2. E-step:

   Conditional on $\underline{U}$ and the current values for $\hat{\underline{\xi}}$ calculate the expected value of $\underline{Y}$. We need to calculate $E(\underline{Y}_{i(a)}) = \int \underline{y}_i \hat{f}_{(a)}(\underline{y}_i|\underline{x}_i,\underline{u}_i)d\underline{y}_i$, which can be evaluated numerically. Here $\underline{y}_i$ is t-dimensional vector, and can be integrated out by multiple integration technique. Monte Carlo algorithm can be used for high dimensional integration.

   The CUBA library has been written by Thomas Hahn (Hahn [2005])in C. In this paper, we use the CUBA library to implement the Monte Carlo algorithm. The CUBA library is a library for multidimensional numerical integration using Monte Carlo methods. The R package "R2Cuba", which is an interface to R, was used to call the CUBA library into R.

3. M-step:

   Update $\hat{\underline{\xi}}$ by maximizing $Q(\underline{\xi}|\hat{\underline{\xi}}_{(a)})$ at given $E(\underline{Y}_{(a)})$.

   For the general multivariate linear model, the maximum likelihood (ML) estimate of β is

$$\hat{\beta} = \left[\sum_{i=1}^{N}(X_i'\hat{\Sigma}_i^{-1}X_i)\right]^{-1}\sum_{i=1}^{N}(X_i'\hat{\Sigma}_i^{-1}\hat{y}_{i(a)}),$$

   where $\hat{\Sigma}_i$ is the maximum likelihood (ML) estimate of $\Sigma_i(\eta)$. The ML estimate

of η is obtained by maximizing the log-likelihood with respect to η. In general, the equations to solve η are non-linear, and need an iterative technique, such as the Newton-Raphson algorithm, to obtain the ML estimate of η. Standard statistical software can be used to obtain $\hat{\underline{\xi}}$ at this step.

4. Repeat (2) and (3) until convergence.

4.3.2 Asymptotic Covariance

The EM algorithm is convenient for obtaining the estimates. However, unlike Newton-Raphson algorithm, EM algorithm does not provide a means of estimating the information matrix associated with the maximum likelihood estimates. We use the method of Louis (Louis [1982]) to deduce the information matrix of observed data.

Gong and Samaniego [1981] introduced the theory and applications of pseudo-MLE. In their paper, they jointly estimate both the nuisance and primary parameters from the data and determine the statistical properties of the primary parameters. The situation from this dissertation is different from the context of Gong and Samaniego [1981] because the nuisance parameters are estimated from independent calibration data. The problem can be viewed as one in which nuisance parameters (θ) are estimated from the independent calibration data and then are treated as known in calculating estimators for the primary interested parameters ($\underline{\xi}$). Therefore, the asymptotic results for pseudo-MLE, $\hat{\underline{\xi}}$ follow from Spall [1989].

Under suitable condition, the pseudo-MLE, $\hat{\xi}$, is consistent and asymptotically normal(Spall [1989]). For detail information, please see Appendix A.

Let $\underline{\omega}' = (\underline{\xi}', \underline{\theta}')$, the information matrix associated with the observed data $U =$

$(U_1, \ldots, U_n)$ is

$$\begin{aligned} I_U(\underline{\omega}) &= -E\left[\frac{\partial^2}{\partial^2 \underline{\omega}} log \prod_{i=1}^{n} f_{U_i}(U_i|\underline{x}_i;\underline{\omega})\right] \\ &= \begin{bmatrix} I_U(\underline{\xi}) & I_U(\underline{\xi},\underline{\theta}) \\ I_U(\underline{\xi},\underline{\theta})' & I_U(\underline{\theta}) \end{bmatrix} \end{aligned} \tag{4.7}$$

The asymptotic covariance of the pseudo-MLE $\hat{\underline{\xi}}$ can be obtained from Theorem 2, Spall [1989], Buonaccorsi and Tosteson [1993], Buonaccorsi [1996], which is

$$A(\hat{\underline{\xi}}) = I_U(\underline{\xi})^{-1} + I_U(\underline{\xi})^{-1} I_U(\underline{\xi},\underline{\theta}) I_C(\underline{\theta})^{-1} I_U(\underline{\theta},\underline{\xi}) I_U(\underline{\xi})^{-1} \tag{4.8}$$

where $I_U(\underline{\xi})$ is the information matrix for $\underline{\xi}$ from the main data, $I_C(\underline{\theta})$ is the information matrix for $\underline{\theta}$ from the calibration data. The second term of this formula is the contribution by the uncertainty in the calibration parameters $\hat{\underline{\theta}}$, since we use the MLE of the calibration parameters $\underline{\theta}$ instead of the true value of $\underline{\theta}$ in our estimation procedure.

When $g(y, \underline{\gamma})$ is linear in y, we can attempt to find a closed form expression for $I_U(\underline{\omega})$. Otherwise, a closed form expression for $I_U(\underline{\omega})$ does not exist. However it can be estimated by the observed information matrix

$$\hat{I}_U(\underline{\omega}) = -\frac{\partial^2}{\partial^2 \underline{\omega}} log \prod f_U(U_i|\underline{x}_i;\underline{\omega})|_{\underline{\xi}=\hat{\underline{\xi}},\underline{\theta}=\hat{\underline{\theta}}} \tag{4.9}$$

$\hat{I}_U(\underline{\omega})$ has to be calculated by numerical integration. Buonaccorsi and Tosteson [1993] used the pseudo-likelihood estimation and developed the asymptotic properties of pseudo-MLE. The asymptotic results for the pseudo-MLE follow Spall [1989] and some algebraic simplication. The asymptotic results consider the contribution

due to uncertainty of parameters in the independent, external calibration data. To calculate the observed information matrix, he obtained formulas directly via differentiation and simplication. He pointed out the the observed information matrix can also be obtained by using the approach of Louis method. The effort should be the same with either approach. In this dissertation, we used the Louis method(Louis [1982]) to calculate the observed information matrix $\hat{I}_U(\underline{\omega})$. Louis [1982] proposed a technique for computing the observed information within the EM framework. His method requires computation of the complete data gradient and second derivative matrix. Louis [1982] shows that the observed data information matrix

$$I(\hat{\underline{\xi}}; u) = \int I(\underline{\xi}; u, y) f(y|u, \hat{\underline{\xi}}) dy - \int \{S(\underline{\xi}; u, y) S(\underline{\xi}; u, y)^T\} f(y|u, \hat{\underline{\xi}}) dy$$

where $S(\underline{\xi}; u, y) = \partial l(\underline{\xi}; u, y)/\partial \underline{\xi}$, the complete-data score functions, and $I(\underline{\xi}; u, y) = -\partial S(\underline{\xi}; u, y)/\partial \underline{\xi}^T$, the complete-data information matrix. The first term of this equation is the conditional expected complete data observed information matrix, and the 2^{nd} term is the expected information for the conditional distribution of Y. The equation shows that the observed information is the difference between the complete information and the missing information to be adjusted for due to the missing data (Woodbury [1971]).

Then, by using a simplified notation, we have

$$I_U = I_Y - I_{Y|U}$$

I_U defined as the observed information, and this is a more appropriate measure of information than $E(B(U, \xi))$, where B is the negative of the 2^{nd} derivative matrix.

4.4 Simulation Study

A simulation study was conducted to investigate the performance of the proposed multivariate linear measurement error model. We assume the measurement error model is normal nonlinear regression model with constant variance which has

$$U|Y = y \sim N(g(y, \underline{\gamma}), \tau^2)$$

The function $g(y, \underline{\gamma})$ is the calibration function with parameters $\underline{\theta} = (\underline{\gamma}, \tau)$. We define function $g(y, \underline{\gamma})$ is a nonlinear, four-parameter logistic model

$$g(y, \gamma) = \gamma_1 + \frac{\gamma_2 - \gamma_1}{1 + (exp(Y)/\gamma_3)^{\gamma_4}} \tag{4.10}$$

We define the calibration parameters as $\gamma_1 = 0.05, \gamma_2 = 0.55, \gamma_3 = 10.0, \gamma_4 = 4.0$, and $\tau^2 = 0.0002$. The normal error model was used to generate the calibration data with the sample size $M = 9$. We define the standard values $exp(Y^*) = 0, 1.25, 2.5, 5, 10, 20, 40, 80$ and 160. The estimates of calibration parameters $\underline{\gamma}$ and τ^2 can be obtained by using nonlinear regression from standard statistical software.

To generate the main data, $Y_1, \ldots, Y_n$ were generated as i.i.d from multivariate normal distribution $MVN_t(\underline{\mu}, \Sigma)$, where μ is a $t \times 1$ mean vector, and Σ is the $t \times t$ variance-covariance matrix. The covariance matrix Σ was defined as $\begin{bmatrix} 0.02 & 0.01 & 0.01 \\ 0.01 & 0.02 & 0.01 \\ 0.01 & 0.01 & 0.02 \end{bmatrix}$, and the mean vector $\underline{\mu}$ was defined as $\begin{bmatrix} 2.303 \\ 2.485 \\ 2.708 \end{bmatrix}$.

We simulated data at sample sizes of 20, 100, and 250. At each sample size, 250

replicate datasets were generated.

A simple approach to correct measurement error is using adjusted/imputed values obtained from the fitted calibration curve. Assuming the inverse of the calibration function is defined, then the ij^{th} adjusted/imputed value of the true value $\hat{Y_{ij}} = g^{-1}(U_{ij}, \hat{\underline{\gamma}})$, where g is the calibration curve. The adjusted values are obtained by inverting the observed values using the fitted calibration curve. This approach ignores some aspects of the measurement error but may suffice if the measurement error variance is sufficiently small.

In the simulation study, we used the same calibration function g, a nonlinear, four-parameter logistic model. Then the adjusted value is defined as

$$\hat{Y_{ij}} = g^{-1}(U_{ij}, \hat{\underline{\gamma}}) = ln\hat{\gamma_3} + \frac{1}{\hat{\gamma_4}} ln(\frac{\hat{\gamma_2} - \hat{\gamma_1}}{U_{ij} - \hat{\gamma_1}} - 1) \tag{4.11}$$

In the simulation study, the estimators of pseudo-MLEs method and adjusted values methods were obtained. For the adjusted values methods, the estimators were obtained by using the sample mean and sample covariance of the adjusted values. Let Σ_A denotes the sample variance-covariance of the adjusted values. $E(\Sigma_A) \approx \Sigma + h^{'}\tau^2 Ih$ rather than Σ, where $h = \partial g^{-1}(U, \underline{\gamma})/\partial U$. This shows that Σ_A overestimates Σ. The analysis using adjusted values is not always based on 250 replicate datasets, because the adjusted values were not always defined for some observations in some cases. For the calibration curve we used, the fitted calibration curve is bounded between the high value of $\hat{\gamma_3}$ and the low value of $\hat{\gamma_1}$. Then if the U_{ij} falls out this range, the adjusted value does not exist. If this situation happened for any of the observations, the whole simulated dataset was deleted. The number of the simulated datasets used is showed in the Tables. The point estimates of the parameters are showed in Table 4.1 and 4.2. For the mean parameter $\underline{\mu}$, the differences between the

two approaches are relatively small. But for the variance-covariance matrix Σ, the adjusted value approach tends to overestimate the parameters.

A larger value of τ^2 is also used in the simulation study. Sometimes the adjusted values do not exist in the adjustd value approach. If this situation happened for any of the observation, the whole simulated dataset was deleted. When we used a larger value of τ^2, too many datasets were deleted because of the non-existence of the adjusted values, and no valid estimates can be obtained from the simulation study. Therefore, only the results from PML approach were showed.

To construct the 95% confidence region, we uses the mean of the adjusted values and the sample variance-covariance matrix for the adjusted value approach. Let $\underline{\bar{Y}}_A$ denotes the mean of the adjusted values, and S denotes the sample variance-covariance matrix of the adjusted values respectively. Then for $(n-t)$ is large, we have an approximate 95% confidence region for $\underline{\mu}$ is given by the set of all $\underline{\mu}$ such that

$$n(\underline{\bar{Y}}_A - \underline{\mu})' S^{-1} (\underline{\bar{Y}}_A - \underline{\mu}) \leq \chi^2_{t,0.05}$$

In the Adjusted values approach, we use the diagonal elements of S/n as the squared standard errors. To construct the 95% confidence intervals, we use the mean of the adjusted values with the standard errors of $[diag(S/n)]^{1/2}$.

For the pseudo-MLEs method, the approximate 95% confidence region for $\underline{\mu}$ is given by the set of all $\underline{\mu}$ such that

$$(\underline{\hat{\mu}} - \underline{\mu})' \hat{\Sigma}^{-1} (\underline{\hat{\mu}} - \underline{\mu}) \leq \chi^2_{t,0.05}$$

where $\underline{\hat{\mu}}$ is the pseudo-MLE of $\underline{\mu}$ and $\hat{\Sigma}$ is the estimators of variance-covariance of $\underline{\hat{\mu}}$ obtained by Louis method. The estimated coverage rates were showed in the Table

4.3. From the simulation results, we can see that the coverage rates of adjusted value approach are lower than the coverage rates of pseudo-MLE approach. The coverage rates of the adjusted value approach decrease when sample size n increases. Apparently, simply using the adjusted value approach to estimate the variance-covariance matrix and constructing the confidence regions is not appropriate.

4.5 Real-life Example

Our motivation study is the Protective Immunity Project (PIP) conducted at the Emory Transplant Center(Larsen and Ahmed [2005]). In the PIP study, the investigators enrolled 60 patients aged 18-59 years old who had renal transplantation at Emory University transplant center. This study began in 2005 and ended in 2011. The investigators enrolled 60 patients aged 18-59 years old who had renal transplantation at Emory University transplant center. They also enrolled 20 age-, sex- and race-matched healthy volunteers into control groups. All subjects enrolled in this study were followed for two years, and multiple blood samples were collected at baseline, 3 months, 6 months, 9 months, 12 months, 18 months and 24 months. The blood samples were analyzed with Flow cytometry. The total counts of white blood cells and the counts of subcategories of white blood cells were recorded based on cell surface markers. However, the cell counts obtained from the flow cytometer may contain measurement errors. Our goal is correcting the measurement errors in count data and obtaining the estimates of true cell counts. Right now we have data from 23 subjects. Because of the missing data issue in the later follow-up times, we only analyze the cell counts from baseline, 3 months and 6 months at this time. To correct the measurement errors in the cell counts obtained from the flow cytometer, we also need calibration data to estimate the parameters of measurement error model. In the calibration data, we have the leukocytes counts obtained from the flow cytometer. True leukocytes counts are obtained through a clinical approach which is treated as the gold standard approach. Figure 4.1 displays the calibration data and the fitted calibration line. The x-axis represents the log of the leukocytes counts from the clinical approach, which is treated as the gold standard method. The y-axis represents the log of the leukocytes counts from flow cytometer, which may contain measurement

errors. From the figure, we can see that a linear regression model can be used to fit the calibration data. Based on the calibration data, the measurement error model we used is

$$g(y, \underline{\beta}) = \beta_0 + \beta_1 y$$

where y is the log of leukocytes counts.

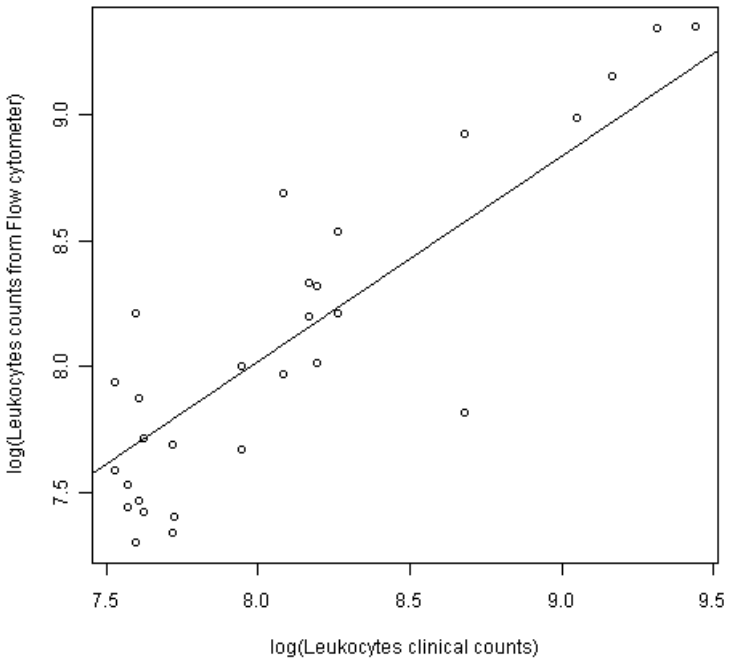


Figure 4.1: Calibration data and fitted calibration line

Now we consider the main data $y_{i1}, \ldots, y_{it}$ are the true values of the log of leukocytes counts over time periods t, t=3. The true value variable $\underline{y}_i$ is a $t \times 1$ vector following a multivariate normal distribution with t dimensions, and $\underline{y}_i | \sim f(\underline{y}_i; \underline{\xi}), i =$

$1, \ldots, n$. Our main interest is the multivariate linear model

$$\underline{Y}_i = \underline{\mu} + \epsilon_i$$

where the ϵ_i are the i.i.d t-dimensional normal with mean $\underline{0}$ and covariance Σ.

The measurement error model is defined as the conditional distribution of U given $Y = y$.

$$U|Y = y \sim N(g(y, \underline{\beta}), \tau^2)$$

The function $g(y, \underline{\beta})$ is the calibration function with parameters $\underline{\theta} = (\underline{\beta}, \tau)$. Table 4.4 shows the estimates of calibration parameters obtained from the calibration data. Table 4.5 shows the analysis from the pseudo-MLE approach and from the analysis of the adjusted values approach. The results show that the adjusted values approach tends to produce smaller standard errors than the pseudo-MLE approach.

4.6 Discussion

In this chapter, our objective is to propose likelihood based estimators for general multivariate linear model when non-linear measurement errors exist in the response variables. The observed response variables are related to the true values through a non-linear regression model, and the parameters in the measurement error model are estimated by using independent, external calibration data. The full MLE method maximize the full likelihood $L(\underline{\xi}, \underline{\theta})$. Obtaining the full MLE may be cumbersome. If we use the standard approaches such as Netwon-Raphson or scoring methods, we need to evaluate numerical integrals at each iteration. However, if we use the EM algorithm, the E-step does not have a closed form solution if the measurement error model is nonlinear. Under this situation, using Em algorithm to obtain pseudo-MLE is relative

easy in computation. To avoid computational issues, we used the pseudo-maximum likelihood estimators and described their asymptotic properties under normal assumptions. The expression of the asymptotic covariance matrix of the pseudo-MLE avoids calculating $\hat{I}_U(\underline{\theta})$, which can be tedious to obtain. If the standard approach, i.e. Newton-Raphson algorithm, was used for computing the full MLE, $\hat{I}_U(\underline{\theta})$ has to be calculated at every iteration. If the Newton-Raphson algorithm was used for computing the Pseudo-MLE, $\hat{I}_U(\underline{\xi})$ has to be calculated at every iteration. For the EM algorithm, $\hat{I}_U(\underline{\xi})$ is only needed to calculate at the last iteration.

A simulation study was conducted to evaluate the performance of the PML estimators and the adjusted value method, which simply analyzes the adjusted values obtained from the fitted calibration curve.

The naive analysis using the adjusted values is a common practice. If the adjusted value $\hat{Y} = Y + \varepsilon$, where the ε is i.i.d. with mean 0 and common variance, the inferences for $\underline{\beta}$ are correct. In this situation, the naive analysis provides a reasonable approximation. However, this situation does not hold exactly all the time. In general, $\hat{Y}$ is biased for Y. In some times, $E(\hat{Y})$ may not exist as we saw in the simulation study. Only if the measurement error model is a normal linear model, the exact results are available (Buonaccorsi [1991]). In that situation, the point estimates for $\underline{\beta}$ are consistent, but the confidence intervals of $\underline{\beta}$ are too small. However, if the measurement error model is non-linear, estimates of $\underline{\beta}$ obtained from the adjusted values are not only biased, but also are inconsistent with the asymptotic bias depending on the curvature of the response curve, and the measurement error variance (Buonaccorsi [1996]). In practice, the bias in estimated coefficients is often modest, then using adjusted values is attractive for its simplicity.

The simulation results show that simply using the adjusted value is not appropri-

ate when estimating the variance-covariance matrix and constructing the confidence regions. The coverage rates of the adjusted values method decrease as the sample size n increases.

Table 4.1: Simulation results of point estimates of mean vector based on 250 replicates, k is the number of replicates used for Adjusted Value approach.

			PML				Adjusted Value				
Parameters	n	True value	Estimate	S.D.	Mean Std error	95% coverage	Estimate	S.D.	Mean Std error	95% coverage	k
μ_1	20	2.303	2.306	0.033	0.037	0.932	2.300	0.032	0.032	0.943	246
μ_2		2.485	2.488	0.034	0.036	0.917	2.477	0.031	0.032	0.923	
μ_3		2.708	2.707	0.033	0.038	0.924	2.700	0.036	0.035	0.931	
μ_1	100	2.303	2.308	0.016	0.021	0.941	2.300	0.014	0.014	0.925	234
μ_2		2.485	2.489	0.015	0.025	0.937	2.478	0.015	0.014	0.902	
μ_3		2.708	2.709	0.015	0.020	0.929	2.700	0.017	0.016	0.883	
μ_1	250	2.303	2.303	0.010	0.017	0.952	2.303	0.009	0.009	0.883	214
μ_2		2.485	2.486	0.0099	0.015	0.942	2.479	0.008	0.010	0.767	
μ_3		2.708	2.709	0.0098	0.018	0.939	2.702	0.010	0.010	0.767	

Table 4.2: Simulation results of point estimates of covariance matrix based on 250 replicates, k is the number of replicates used for Adjusted Value approach.

			PML		Adjusted Value		
Parameters	n	True value	Estimate	S.D.	Estimate	S.D.	k
σ_{11}	20	0.02	0.021	0.007	0.021	0.007	246
σ_{12}		0.01	0.011	0.005	0.0098	0.005	
σ_{13}		0.01	0.010	0.003	0.010	0.006	
σ_{22}		0.02	0.022	0.008	0.021	0.008	
σ_{23}		0.01	0.011	0.006	0.011	0.006	
σ_{33}		0.02	0.021	0.006	0.025	0.013	
σ_{11}	100	0.02	0.020	0.006	0.021	0.003	234
σ_{12}		0.01	0.010	0.004	0.0098	0.002	
σ_{13}		0.01	0.0098	0.002	0.011	0.002	
σ_{22}		0.02	0.021	0.008	0.021	0.003	
σ_{23}		0.01	0.010	0.004	0.011	0.003	
σ_{33}		0.02	0.021	0.005	0.024	0.006	
σ_{11}	250	0.02	0.021	0.005	0.022	0.0019	214
σ_{12}		0.01	0.010	0.003	0.011	0.0014	
σ_{13}		0.01	0.0099	0.002	0.011	0.0011	
σ_{22}		0.02	0.021	0.008	0.021	0.0014	
σ_{23}		0.01	0.011	0.003	0.012	0.0013	
σ_{33}		0.02	0.021	0.003	0.025	0.0025	

Table 4.3: Estimated coverage rates of approximate 95 percent confidence regions for mean vector based on 250 replicates, k is the number of replicates used for Adjusted Value approach.

Parameters	n	PML	Adjusted Value	k
	20	0.9126	0.8821	246
$\underline{\mu}$	100	0.9204	0.8785	234
	250	0.9337	0.7333	214

Table 4.4: Simulation results of point estimates of mean vector based on 250 replicates with larger $\tau^2 = 0.02$

			PML		
Parameters	n	True value	Estimate	Mean Std error	95% coverage
μ_1	20	2.303	2.308	0.049	0.873
μ_2		2.485	2.491	0.053	0.896
μ_3		2.708	2.712	0.059	0.862
μ_1	100	2.303	2.309	0.036	0.907
μ_2		2.485	2.489	0.034	0.914
μ_3		2.708	2.709	0.041	0.895
μ_1	250	2.303	2.305	0.021	0.928
μ_2		2.485	2.483	0.027	0.917
μ_3		2.708	2.709	0.031	0.934

Table 4.5: Estimated calibration parameters with standard errors from the calibration data

Parameters	Estimates	Standard errors
$\hat{\beta}_0$	1.498	0.694
$\hat{\beta}_1$	0.816	0.0856
τ^2	0.0781	

Table 4.6: Analysis of real data using pseudo-MLE and adjusted values approaches

	PML		Adjusted Value	
Parameters	Estimate	S.E.	Estimate	S.E.
μ_1	8.678	0.0767	8.671	0.0731
μ_2	8.404	0.0880	8.401	0.0845
μ_3	8.441	0.0894	8.445	0.0886

Chapter 5

MEASUREMENT ERROR IN GENERAL MULTIVARIATE LINEAR MODEL-A BAYESIAN APPROACH

5.1 Background

5.1.1 Measurement Error in Response Variables

In medical studies, some variables of interest are difficult or expensive to obtain. Instead, surrogate variables which can be obtained relatively easy and inexpensive are recorded and used by researchers in real-life situation. However, these surrogate variables may contain measurement errors. The examples include blood pressure, cholesterol serum level, hormone level, and so on. A typical example can be found in radioimmunoassay. In radioimmunoassay, the variable of interest is the molecular level in the sample, but the observed value is the radioactive count. The molecular level in the sample can not be tested directly, and the observed radioactive counts may contain measurement errors.

Measurement error problems in predictor variables have recently received extensive attention by researchers. There are many papers about how to correct measurement errors in regressors in different applications (Carroll and Stefanski [1990]; Brown and Fuller [1990]; Byar and Gail [1989]; Fuller [1987]). However, there is little literature about the problem with measurement errors in the response variable. The main reason is that the problem can be handled with standard methodology if the measurement errors in response variable are additive errors in which the observed values = true value + error, where the error has mean 0. When the measurement errors are additive with constant variance, the errors can be ignored in regression analysis because they can be thought as an extra variance component. However, for many measuring methods, the measurement error models are complex. The additive errors assumption may not appropriate. If the measurement error in response variable is biased, ignoring the response measurement errors will lead to biased estimates of

the regression parameters. Buonaccorsi [1991, 1996] and Buonaccorsi and Tosteson [1993] proposed methods to adjust the nonlinear biased measurement errors in response variables in the general linear model. Buonaccorsi and Tosteson [1993] used the likelihood based method to adjust the nonlinear biased measurement error in the response variable. We propose a likelihood-based method to correct the measurement errors in response variables in the general multivariate linear model setting in Chapter 4. In this chapter, we propose a Bayesian method to adjust the nonlinear biased measurement errors in the response variables in the general multivariate linear model setting.

5.1.2 Bayesian methods for measurement errors

Bayesian methods have become a highly popular and powerful tool in statistical analysis over the past twenty years. We will give a brief introduction of Bayesian methods in the measurement error problems, and how to formulate measurement error models by using Bayesian methods.

There are five steps for measurement error problems in the Bayesian approach.

- Step 1: Select the likelihood model. We must specify a parametric model for every component of the data, and we assume the true unobservable values Y were observable in the model.

- Step 2: Select the measurement error model. We need to decide which measurement error model should be used. The error model could be a classical error model, or a Berkson model. We need to specify the distribution for the true unobservable values Y, if we choose a classical error model.

- Step 3: Form the likelihood function. At this step, we form the likelihood

function of all the data, including the observed data U with measurement error, and the true unobservable data Y. We form the complete likelihood function as if Y were available.

- Step 4: Select the priors. In the Bayesian analysis, we treat the parameters as random variables. Therefore, we need to establish prior distributions for these parameters. The true unobservable variables Y are treated as missing data, and can be imputed by drawing from the full conditional distribution of Y given all the other variables.

- Step 5: Compute full conditionals. The full conditionals are the distributions of the parameters, and the Y values, given everything else in the model. Given the prior distributions and all the observed data, we can get the posterior distribution of parameters. Markov Chain Monte Carlo (MCMC) methods can be used to compute Bayesian quantities. Gibbs sampler can be used to draw samples from the full conditionals. If we do not have conditional conjugancy, we can use Metropolis-Hasting algorithm.

5.2 Modeling

Our purpose in this chapter is proposing a Bayesian method to adjust the measurement errors in response variables in the general multivariate linear model.

Now we consider the longitudinal data $y_{i1}, \ldots, y_{it}$ as the true values over time periods t. The true value variable $\underline{y}_i$ is a $t \times 1$ vector following a multivariate normal distribution with t dimensions, and $\underline{y}_i| \sim f(\underline{y}_i; \underline{\xi}), i = 1, \ldots, n$. Our main interest is the multivariate linear model

$$\underline{Y}_i = \underline{\mu} + \epsilon_i$$

where the ϵ_i are i.i.d t-dimensional normal random variables with mean $\underline{0}$ and covariance Σ. Then the joint density function of $y_{i1}, y_{i2}, \ldots, y_{it}$ is

$$p(\underline{y}|\underline{\mu}, \Sigma) = \left(\frac{1}{2\pi}\right)^{\frac{t}{2}} |\Sigma|^{-\frac{1}{2}} exp\left[-\frac{1}{2}\left(\underline{y}-\underline{\mu}\right)^T \Sigma^{-1}\left(\underline{y}-\underline{\mu}\right)\right] \tag{5.1}$$

In this model, our main parameters of interest are $\underline{\xi} = (\underline{\mu}, \underline{\Sigma}(\eta))$. The goal of the model inference is estimating $\underline{\xi}$ while the true values $\underline{Y}'s$ are not observable. The independent variables $\underline{x_i}$ are assumed fixed and observed without measurement error. Therefore, our model has measurement error in the dependent variable $\underline{Y}$ only.

Suppose the true value y_{it} is unobservable, and we have observed values u_{it} with measurement error, where u_{it} is the measured value of true value y_{it}. The measurement error model is defined as the conditional distribution of U given $Y = y$. Independent calibration data is needed to estimate the parameters of measurement error model. For the i^{th} individual unit at time point t, Given $Y_{it} = y_{it}$, the observed value U_{it} has the density function $f_{U|Y}(u_{it}|y_{it}; \underline{\theta})$. Here we assume the measurement error model is a normal nonlinear regression model with constant variance which has

$$U|Y = y \sim N(g(y, \underline{\gamma}), \tau^2)$$

The function $g(y, \gamma)$ is the calibration function with parameters $\underline{\theta} = (\underline{\gamma}, \tau)$. We define function $g(y, \underline{\gamma})$ is a nonlinear, four-parameter logistic model

$$g(y, \gamma) = \gamma_1 + \frac{\gamma_2 - \gamma_1}{1 + (exp(Y)/\gamma_3)^{\gamma_4}} \tag{5.2}$$

The complete density function of Y and U is

$$f(\underline{Y}, \underline{U}|\underline{\mu}, \Sigma, \underline{\gamma}, \tau^2) = \left(\frac{1}{2\pi}\right)^{\frac{t}{2}} |\Sigma|^{-\frac{1}{2}} |\tau^2 I|^{-\frac{1}{2}}$$

$$\times exp\left[-\frac{1}{2}\left(\underline{y}-\underline{\mu}\right)^T\Sigma^{-1}\left(\underline{y}-\underline{\mu}\right)\right]$$
$$\times exp\left[-\frac{1}{2}\left(\underline{u}-g(\underline{y},\underline{\gamma})\right)^T(\tau^2 I)^{-1}\left(\underline{u}-g(\underline{y},\underline{\gamma})\right)\right] \quad (5.3)$$

Assume we have independent, external calibration data available. In the calibration data $(U_j^*, Y_j^*, j = 1, \ldots, m)$, the Y^* are fixed constants (observable and based on gold standards) and the U^* are observable random variables with U_j^* having the density function $f_{U|Y}(U_j^*|Y_j^*;\underline{\theta})$. The data in the independent, external calibration dataset are all at one time point, so we assume the measurement error model and its parameters are the same over time. By using the independent, external calibration data, we make an assumption of transportability characteristic when transporting measurement error model to the main data (Carroll, Ruppert, A., and M. [2006]). Let $\hat{\underline{\theta}}$ be the estimator of $\underline{\theta}$ obtained from the calibration data, and we assume $\underline{\theta}$ is identifiable from the independent, external calibration data. The density function of U^* is

$$f(U^*|Y^*,\underline{\gamma},\tau^2) = \left(\frac{1}{2\pi\tau^2}\right)^{\frac{1}{2}} exp\left[-\frac{1}{2\tau^2}\left(U^*-g(Y^*,\underline{\gamma})\right)^2\right] \quad (5.4)$$

Combining the main data and the independent calibration data, the complete likelihood is

$$\begin{aligned} f(\underline{Y},\underline{U},U^*,Y^*|\underline{\mu},\Sigma,\underline{\gamma},\tau^2) &= \left(\frac{1}{2\pi}\right)^{\frac{nt}{2}}|\Sigma|^{-\frac{n}{2}}|\tau^2 I|^{-\frac{n}{2}} \\ &\quad \times exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\underline{y_i}-\underline{\mu}\right)^T\Sigma^{-1}\left(\underline{y_i}-\underline{\mu}\right)\right] \\ &\quad \times exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\underline{u_i}-g(\underline{y_i},\underline{\gamma})\right)^T(\tau^2 I)^{-1}\left(\underline{u_i}-g(\underline{y_i},\underline{\gamma})\right)\right] \\ &\quad \times \left(\frac{1}{2\pi\tau^2}\right)^{\frac{m}{2}} exp\left[-\frac{1}{2\tau^2}\sum_{j=1}^{m}\left(U_j^*-g(Y_j^*,\underline{\gamma})\right)^2\right] \end{aligned}$$

(5.5)

5.3 Model Inference

We can use Bayesian method for making inference about unknown parameters and also unobservable true variables $\underline{Y}$.

To implement Bayesian approach, we need to specify the prior distributions for all the unknown parameters in the model. The unknowns in this model are $(\underline{\mu}, \Sigma)$, $(\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_t)$, and $(\underline{\gamma}, \tau)$. In the Bayesian approach, we treat the true unobservable variable $\underline{Y}$ as missing data, and impute $\underline{Y}$ multiple times by drawing from the conditional distribution of $\underline{Y}$ given all other variables. The priors we will use are

$$\begin{aligned} \underline{\mu} &\sim Normal_t(\underline{0}, \sigma_\mu^2 I_t) \\ \Sigma^{-1} &\sim Wishart(\Psi^{-1}, \rho) \\ \tau^2 &\sim IG(\alpha, \beta) \\ \gamma_1, \gamma_2, \gamma_3, \gamma_4 &\sim Normal(0, \sigma_\gamma^2) \end{aligned} \tag{5.6}$$

Here, IG is the inverse gamma density function, Ψ is a $t \times t$ positive definite matrix, and ρ denotes the degrees of freedom of the Wishart distribution. I_t is the $t \times t$ identity matrix. The hyperparameters σ_μ^2, σ_γ^2 can be chosen to be large, and the hyperparameters α, β can be chosen to be small, so that the priors are relatively noninformative.

The joint density function of all observed data and all unknown quantities (parameters and the true unobservable variable Y) is the product of the joint likelihood

and the joint priors.

To find the full conditional for $\underline{\mu}$, we isolate the terms depending on $\underline{\mu}$ in this joint density. Then we can write the full conditional of $\underline{\mu}$ given all the other parameters.

$$f(\underline{\mu}|\ldots) \propto exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\underline{y}_i-\underline{\mu}\right)^T\Sigma^{-1}\left(\underline{y}_i-\underline{\mu}\right)-\frac{1}{2}\left(\underline{\mu}^T(\sigma_\mu^2 I)^{-1}\underline{\mu}\right)\right] \tag{5.7}$$

where the first term in the exponent comes from the likelihood and the second term comes from the prior. After some rearrangements, we find that the full conditional distribution of $\underline{\mu}$ is a multivariate normal distribution with mean

$$\left[(\sigma_\mu^2 I_t)^{-1}+n\Sigma^{-1}\right]^{-1}\left[n\Sigma^{-1}\frac{1}{n}\sum_{i=1}^{n}(\underline{y}_i)\right]$$

and variance-covariance matrix

$$\left[(\sigma_\mu^2 I_t)^{-1}+n\Sigma^{-1}\right]^{-1}.$$

Similarly, we can find the full conditional of Σ^{-1} given all the other parameters.

$$f(\Sigma^{-1}|\ldots) \propto |\Sigma|^{-\frac{\rho}{2}}exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}(\underline{y}_i-\underline{\mu})^T\Sigma^{-1}(\underline{y}_i-\underline{\mu})+tr(\Psi\Sigma^{-1})\right]\right\} \tag{5.8}$$

After some rearrangements, we find that the full conditional distribution of the variance-covariance matrix Σ is an inverse Wishart distribution with parameter matrix $(V+\Psi)$, where

$$V=\left[\sum_{i=1}^{n}\underline{y}_i-\underline{\mu}\right]\left[\sum_{i=1}^{n}\underline{y}_i-\underline{\mu}\right]^T$$

and degrees of freedom is $n+\rho$.

The full conditional of τ^2 is

$$\begin{aligned} f(\tau^2|\ldots) &\propto |\tau^2 I_t|^{-\frac{n}{2}} exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\underline{u}_i - g(\underline{y},\underline{\gamma}))^T(\tau^2 I_t)^{-1}(\underline{u}_i - g(\underline{y},\underline{\gamma}))\right] \\ &\quad \times \left(\tau^2\right)^{-\frac{m}{2}} exp\left[-\frac{1}{2\tau^2}\sum_{j=1}^{m}\left(U_j^* - g(Y_j^*,\underline{\gamma})\right)^2\right](\tau^2)^{-(\alpha+1)}exp(\frac{\beta}{\tau^2}) \end{aligned} \tag{5.9}$$

which implies that the full conditional distribution of τ^2 is an inverse gamma distribution with the shape parameter is $\alpha + \frac{n*t+m}{2}$, and the scale parameter is $\beta + \frac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{T}[u_{i,t} - g(y_{i,t},\underline{\gamma})]^2 + \frac{1}{2}\sum_{j=1}^{m}[U_j^* - g(Y_j^*,\underline{\gamma})]^2$.

The full conditional for the true unobservable variable y_i is

$$\begin{aligned} f(\underline{y}_i|\ldots) &\propto exp\left[-\frac{1}{2}\left(\underline{y}_i - \underline{\mu}\right)^T\Sigma^{-1}\left(\underline{y}_i - \underline{\mu}\right)\right] \\ &\quad \times exp\left[-\frac{1}{2}\left(\underline{u}_i - g(\underline{y}_i,\underline{\gamma})\right)^T(\tau^2 I)^{-1}\left(\underline{u}_i - g(\underline{y}_i,\underline{\gamma})\right)\right] \end{aligned} \tag{5.10}$$

Since $g(\underline{y}_i,\underline{\gamma})$ is a nonlinear function in y, the full conditional of Y can not be expressed as a known family of distributions. Therefore, we can use Metropolis-Hasting (MH) algorithm to update Y. To update $\underline{y}_i$, we use a random walk MH step using $Normal(\underline{y}_i, B\Sigma)$ as the proposal density, where B is the dispersion factor. B should be chosen to make sure a reasonable acceptance rate.

The full conditional for the γ_1 is

$$\begin{aligned} f(\gamma_1|\ldots) &\propto exp\left[-\frac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(u_{it} - g(y_{it},\underline{\gamma})\right)^T(\tau^2)^{-1}\left(u_{it} - g(y_{it},\underline{\gamma})\right)\right] \\ &\quad \times exp\left[-\frac{1}{2\tau^2}\sum_{j=1}^{m}\left(U_j^* - g(Y_j^*,\underline{\gamma})\right)^2\right]exp\left[-\frac{\gamma_1^2}{2\sigma_\gamma^2}\right] \end{aligned}$$

(5.11)

Similarly, the full conditional for the γ_2, γ_3 and γ_4 are

$$\begin{aligned} f(\gamma_2|\ldots) \quad \propto \quad & exp\left[-\frac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(u_{it}-g(y_{it},\underline{\gamma})\right)^T(\tau^2)^{-1}\left(u_{it}-g(y_{it},\underline{\gamma})\right)\right] \\ & \times exp\left[-\frac{1}{2\tau^2}\sum_{j=1}^{m}\left(U_j^*-g(Y_j^*,\underline{\gamma})\right)^2\right] exp\left[-\frac{\gamma_2^2}{2\sigma_\gamma^2}\right] \end{aligned} \tag{5.12}$$

$$\begin{aligned} f(\gamma_3|\ldots) \quad \propto \quad & exp\left[-\frac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(u_{it}-g(y_{it},\underline{\gamma})\right)^T(\tau^2)^{-1}\left(u_{it}-g(y_{it},\underline{\gamma})\right)\right] \\ & \times exp\left[-\frac{1}{2\tau^2}\sum_{j=1}^{m}\left(U_j^*-g(Y_j^*,\underline{\gamma})\right)^2\right] exp\left[-\frac{\gamma_3^2}{2\sigma_\gamma^2}\right] \end{aligned} \tag{5.13}$$

$$\begin{aligned} f(\gamma_4|\ldots) \quad \propto \quad & exp\left[-\frac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(u_{it}-g(y_{it},\underline{\gamma})\right)^T(\tau^2)^{-1}\left(u_{it}-g(y_{it},\underline{\gamma})\right)\right] \\ & \times exp\left[-\frac{1}{2\tau^2}\sum_{j=1}^{m}\left(U_j^*-g(Y_j^*,\underline{\gamma})\right)^2\right] exp\left[-\frac{\gamma_4^2}{2\sigma_\gamma^2}\right] \end{aligned} \tag{5.14}$$

The full conditionals of γ_1, γ_2, γ_3 and γ_4 can not be expressed as a known family of distributions. Therefore, we can use Metropolis-Hasting algorithm to update γ's. To update γ's, we use a random walk MH step using $Normal(\gamma_i, B\sigma_\gamma^2), i = 1, 2, 3, 4$ as the proposal density, where B is the dispersion factor. B should be chosen to ensure

a reasonable acceptance rate.

We now summarize the Gibbs sampler and Metropolis algorithm for this model. Given the current state $(\underline{Y}_{(k)}, \underline{\mu}_{(k)}, \Sigma_{(k)}, \tau^2_{(k)}, \gamma_{1-4,(k)})$ of the Markov chain, we execute the following:

- 1. Sample the unobserved true $\underline{Y}_{k+1}$ from above mentioned posterior distribution;
- 2. Sample $\underline{\mu}_{(k+1)}$ from the posterior distribution;
- 3. Sample $\Sigma_{(k+1)}$ from the posterior distribution;
- 4. Sample $\tau^2_{(k+1)}$ from the posterior distribution;
- 5. Sample $\gamma_{1-4,(k+1)}$ from their respective posterior distribution.
- 6. Take $(\underline{Y}_{(k+1)}, \underline{\mu}_{(k+1)}, \Sigma_{(k+1)}, \tau^2_{(k+1)}, \gamma_{1-4,(k+1)})$ as the current state and return to step 1.

The sampling procedure requires the starting values for the unknown parameters and unobservable true values $\underline{Y}$'s. For γ_{1-4}, τ^2, we can use estimates from the regression of U^* on Y^* based on the calibration data. For the starting values of $\underline{Y}$, we can use the adjusted/imputed value of $\underline{Y}$. The adjusted values are obtained by inverting the observed values using the fitted calibration curve. Assuming the inverse of the calibration function is defined, then the ij^{th} adjusted/imputed value of the true value $\hat{Y_{ij}} = g^{-1}(U_{ij}, \underline{\hat{\gamma}})$, where g is the calibration curve. Although this approach ignores some aspects of the measurement error, it may suffice since $\underline{\hat{Y}}$'s are being used only as starting values. $\underline{\mu}$ and Σ can be started at the sample mean and sample covariance of the starting values of the Y's.

5.4 Simulation Study

A simulation study was conducted to investigate the performance of the proposed Bayesian approach multivariate linear measurement error model. We assume the measurement error model is a normal nonlinear regression model with constant variance which has

$$U|Y = y \sim N(g(y, \underline{\gamma}), \tau^2)$$

The function $g(y, \underline{\gamma})$ is the calibration function with parameters $\underline{\theta} = (\underline{\gamma}, \tau)$. We define function $g(y, \underline{\gamma})$ is a nonlinear, four-parameter logistic model

$$g(y, \gamma) = \gamma_1 + \frac{\gamma_2 - \gamma_1}{1 + (exp(Y)/\gamma_3)^{\gamma_4}} \tag{5.15}$$

We define the calibration parameters as $\gamma_1 = 0.05, \gamma_2 = 0.55, \gamma_3 = 10.0, \gamma_4 = 4.0$, and $\tau^2 = 0.0002$. The normal error model was used to generate the calibration data with the sample size $M = 9$. We define the standard values $exp(Y^*) = 0, 1.25, 2.5, 5, 10, 20, 40, 80$ and 160. The estimates of calibration parameters $\underline{\gamma}$ and τ^2 can be obtained by using nonlinear regression from standard statistical software.

To generate the main data, $Y_1, \ldots, Y_n$ were generated as i.i.d from multivariate normal distribution $MVN_t(\underline{\mu}, \Sigma)$, where μ is a $t \times 1$ mean vector, and Σ is the $t \times t$ variance-covariance matrix. The covariance matrix Σ was defined as $\begin{bmatrix} 0.02 & 0.01 & 0.01 \\ 0.01 & 0.02 & 0.01 \\ 0.01 & 0.01 & 0.02 \end{bmatrix}$, and the mean vector $\underline{\mu}$ was defined as $\begin{bmatrix} 2.303 \\ 2.485 \\ 2.708 \end{bmatrix}$.

We simulated data at sample sizes of 20, 100, and 250. At each sample size, 250

replicate datasets were generated.

A simple approach to correct measurement error is using adjusted/imputed values obtained from the fitted calibration curve. Assuming the inverse of the calibration function is defined, then the ij^{th} adjusted/imputed value of the true value $\hat{Y_{ij}} = g^{-1}(U_{ij}, \underline{\hat{\gamma}})$, where g is the calibration curve. The adjusted values are obtained by inverting the observed values using the fitted calibration curve. This approach ignores some aspects of the measurement error but may suffice if the measurement error variance is sufficiently small.

In the simulation study, we used the same calibration function g, a nonlinear, four-parameter logistic model. Then the adjusted value is defined as

$$\hat{Y_{ij}} = g^{-1}(U_{ij}, \underline{\hat{\gamma}}) = ln\hat{\gamma_3} + \frac{1}{\hat{\gamma_4}} ln(\frac{\hat{\gamma_2} - \hat{\gamma_1}}{U_{ij} - \hat{\gamma_1}} - 1) \tag{5.16}$$

The point estimates of the parameters are showed in Table 5.1 and 5.2. For the mean parameter $\underline{\mu}$, the differences between the two approaches are relatively small. But for the variance-covariance matrix Σ, the estimates from the adjusted value approach tend to overestimate the parameters.

The estimated coverage rates were showed in the Table 5.3. From the simulation results, we can see that the coverage rates of adjusted value approach are lower than the coverage rates of Bayesian approach. The coverage rates of the adjusted value approach decrease when sample size n increases. Apparently, simply using the adjusted value approach to estimate the variance-covariance matrix and constructing the confidence regions is not appropriate.

5.5 Real-Life Example

Our motivation study is the Protective Immunity Project (PIP) conducted at the Emory Transplant Center(Larsen and Ahmed [2005]). In the PIP study, the investigators enrolled 60 patients aged 18-59 years old who had renal transplantation at Emory University transplant center. This study began in 2005 and ended in 2011. The investigators enrolled 60 patients aged 18-59 years old who had renal transplantation at Emory University transplant center. They also enrolled 20 age-, sex- and race-matched healthy volunteers into control groups. All subjects enrolled in this study are followed for two years, and multiple blood samples were collected at baseline, 3 months, 6 months, 9 months, 12 months, 18 months and 24 months. The blood samples were analyzed with Flow cytometry. The total counts of white blood cells and the counts of subcategories of white blood cells were recorded based on the cell surface markers. However, the cell counts obtained from the flow cytometer may contain measurement errors. Our goal is correcting the measurement errors in count data and obtaining the estimates of true cell counts. Right now we have data from 23 subjects. Because of the missing data issue in the later follow-up times, we only analyze the cell counts from baseline, 3 months and 6 months at this time. To correct the measurement errors in the cell counts obtained from the flow cytometer, we also

need the calibration data to estimate the parameters of measurement error model. In the calibration data, we have the leukocytes counts obtained from the flow cytometer. True leukocytes counts are obtained clinically, and are treated as the gold standard. Based on the calibration data, the measurement error model we used is

$$g(y, \underline{\beta}) = \beta_0 + \beta_1 y$$

where y is the log of leukocytes counts.

Now we consider the main data $y_{i1}, \ldots, y_{it}$ as the true values of the log of leukocytes counts over time periods t, t=3. The true value variable $\underline{y}_i$ is a $t \times 1$ vector following a multivariate normal distribution with t dimensions, and $\underline{y}_i| \sim f(\underline{y}_i; \underline{\xi}), i = 1, \ldots, n$. Our main interest is the multivariate linear model

$$\underline{Y}_i = \underline{\mu} + \epsilon_i$$

where the ϵ_i are the i.i.d t-dimensional normal with mean $\underline{0}$ and covariance Σ.

The measurement error model is defined as the conditional distribution of U given $Y = y$.

$$U|Y = y \sim N(g(y, \underline{\beta}), \tau^2)$$

The function $g(y, \underline{\beta})$ is the calibration function with parameters $\underline{\theta} = (\underline{\beta}, \tau)$.

Combining the main data and the calibration data, the complete likelihood is

$$\begin{aligned} f(\underline{Y}, \underline{U}, U^*, Y^*|\underline{\mu}, \Sigma, \underline{\gamma}, \tau^2) &= \left(\frac{1}{2\pi}\right)^{\frac{nt}{2}} |\Sigma|^{-\frac{n}{2}} |\tau^2 I|^{-\frac{n}{2}} \\ &\quad \times exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\underline{y_i} - \underline{\mu}\right)^T \Sigma^{-1}\left(\underline{y_i} - \underline{\mu}\right)\right] \\ &\quad \times exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\underline{u_i} - g(\underline{y}_i, \underline{\beta})\right)^T (\tau^2 I)^{-1}\left(\underline{u_i} - g(\underline{y}_i, \underline{\beta})\right)\right] \end{aligned}$$

$$\times \left(\frac{1}{2\pi\tau^2}\right)^{\frac{m}{2}} exp\left[-\frac{1}{2\tau^2}\sum_{j=1}^{m}\left(U_j^* - g(Y_j^*, \underline{\beta})\right)^2\right] \tag{5.17}$$

where U, Y are from the main data, and U^*, Y^* are from the calibration data.

To implement the proposed Bayesian approach, we need to specify the prior distribution for all the unknown parameters. The priors we will use are

$$\begin{aligned} \underline{\mu} &\sim Normal_t(\underline{0}, \sigma_\mu^2 I_t) \\ \Sigma^{-1} &\sim Wishart(\Psi^{-1}, \rho) \\ \tau^2 &\sim IG(\alpha, \beta) \\ \beta_0, \beta_1 &\sim Normal(0, \sigma_\beta^2) \end{aligned} \tag{5.18}$$

Here, IG is the inverse gamma density function, Ψ is a $t \times t$ positive definite matrix, and ρ denotes the degrees of freedom of the Wishart distribution. I_t is the $t \times t$ identity matrix. The hyperparameters σ_μ^2, σ_β^2 can be chosen to be large, and the hyperparameters α, β can be chosen to be small, so that the priors are relatively noninformative.

The full conditional of $\underline{\mu}$ given all the other parameters is

$$f(\underline{\mu}|\ldots) \propto exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\underline{y}_i - \underline{\mu}\right)^T \Sigma^{-1}\left(\underline{y}_i - \underline{\mu}\right) - \frac{1}{2}\left(\underline{\mu}^T(\sigma_\mu^2 I)^{-1}\underline{\mu}\right)\right] \tag{5.19}$$

where the first term in the exponent comes from the likelihood and the second term comes from the prior. After some rearrangement, we find that the full conditional

distribution of $\underline{\mu}$ is a multivariate normal distribution with mean

$$\left[(\sigma_\mu^2 I_t)^{-1} + n\Sigma^{-1}\right]^{-1}\left[n\Sigma^{-1}\frac{1}{n}\sum_{i=1}^{n}(\underline{y}_i)\right]$$

and variance-covariance matrix

$$\left[(\sigma_\mu^2 I_t)^{-1} + n\Sigma^{-1}\right]^{-1}.$$

Similarly, we can find the full conditional of Σ^{-1} given all the other parameters.

$$f(\Sigma^{-1}|\ldots) \propto |\Sigma|^{-\frac{\rho}{2}} exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}(\underline{y}_i - \underline{\mu})^T\Sigma^{-1}(\underline{y}_i - \underline{\mu}) + tr(\Psi\Sigma^{-1})\right]\right\} \tag{5.20}$$

After some rearrangement, we find that the full conditional distribution of the variance-covariance matrix Σ is an inverse Wishart distribution with parameter matrix $(V+\Psi)$, where

$$V = \left[\sum_{i=1}^{n}\underline{y}_i - \underline{\mu}\right]\left[\sum_{i=1}^{n}\underline{y}_i - \underline{\mu}\right]^T$$

and degrees of freedom is $n + \rho$.

The full conditional of τ^2 is

$$\begin{aligned} f(\tau^2|\ldots) \quad \propto \quad & |\tau^2 I_t|^{-\frac{n}{2}} exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\underline{u}_i - g(\underline{y},\underline{\beta}))^T(\tau^2 I_t)^{-1}(\underline{u}_i - g(\underline{y},\underline{\beta}))\right] \\ & \times\left(\tau^2\right)^{-\frac{m}{2}} exp\left[-\frac{1}{2\tau^2}\sum_{j=1}^{m}\left(U_j^* - g(Y_j^*,\underline{\beta})\right)^2\right](\tau^2)^{-(\alpha+1)}exp(\frac{\beta}{\tau^2}) \end{aligned} \tag{5.21}$$

which implies that the full conditional distribution of τ^2 is an inverse gamma distribution with the shape parameter is $\alpha + \frac{n*t+m}{2}$, and the scale parameter is $\beta + \frac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{T}[u_{i,t} - g(y_{i,t},\underline{\beta})]^2 + \frac{1}{2}\sum_{j=1}^{m}[U_j^* - g(Y_j^*,\underline{\beta})]^2$.

The full conditional for the true unobservable variable $\underline{y}_i$ is

$$\begin{aligned} f(\underline{y}_i|\ldots) \quad &\propto \quad exp\left[-\frac{1}{2}\left(\underline{y}_i-\underline{\mu}\right)^T\Sigma^{-1}\left(\underline{y}_i-\underline{\mu}\right)\right] \\ &\quad \times exp\left[-\frac{1}{2}\left(\underline{u}_i-g(\underline{y}_i,\underline{\beta})\right)^T(\tau^2 I)^{-1}\left(\underline{u}_i-g(\underline{y}_i,\underline{\beta})\right)\right] \end{aligned} \tag{5.22}$$

The full conditional for the β_0 is

$$\begin{aligned} f(\beta_0|\ldots) \quad &\propto \quad exp\left[-\frac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(u_{it}-\beta_0-\beta_1 y_{it})\right)^T(\tau^2)^{-1}\left(u_{it}-\beta_0-\beta_1 y_{it})\right)\right] \\ &\quad \times exp\left[-\frac{1}{2\tau^2}\sum_{j=1}^{m}\left(U_j^*-\beta_0-\beta_1 Y_j^*)\right)^2\right] exp\left[-\frac{\beta_0^2}{2\sigma_\beta^2}\right] \end{aligned} \tag{5.23}$$

Similarly, the full conditional for the β_1 is

$$\begin{aligned} f(\beta_1|\ldots) \quad &\propto \quad exp\left[-\frac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(u_{it}-\beta_0-\beta_1 y_{it})\right)^T(\tau^2)^{-1}\left(u_{it}-\beta_0-\beta_1 y_{it})\right)\right] \\ &\quad \times exp\left[-\frac{1}{2\tau^2}\sum_{j=1}^{m}\left(U_j^*-\beta_0-\beta_1 Y_j^*)\right)^2\right] exp\left[-\frac{\beta_1^2}{2\sigma_\beta^2}\right] \end{aligned} \tag{5.24}$$

After some rearrangement, we find that the full conditional distributions of $\underline{y}_i$, β_0 and β_1 follow multivariate normal distributions, and can be updated by Gibbs sampling method.

The MCMC method is employed for model inference. Estimations of the parameter values are made over 12,000 MCMC iterations, in which the first 2,000 iterations were "burn-in" phase, and were disregarded. The point estimates of the parameters and their 95% credible regions are presented in the table 5.4.

We use the posterior predictive p-value to evaluate the fit of the posterior distribution of our Bayesian model. We use the method of Gelman, Meng, and Stern [1996] to calculate the posterior predictive p-value. The discrepancy function we used in this analysis is

$$D(u, \hat{\theta}) = \sum_{i=1}^{n} \frac{[u_i - E(u_i|\hat{\theta})]^2}{E(u_i|\hat{\theta})} \tag{5.25}$$

Based on 10,000 simulation iterations, the posterior predictive p-value, p_B is 0.84, which is the probability that the replicated data (u^{rep}) could be more extreme than the observed data (u^{obs}), as measured by the discrepancy function $D(u, \hat{\theta})$. Figure 5.1 shows the scatterplot of replicated vs. observed discrepancies (D(rep, θ) vs. D(obs, θ)) under the joint posterior distribution; the p-value is calculated as the proportion of points in the upper-left half of the plot. Figure 5.2 shows the histogram of 10,000 simulations from the difference of the replicated discrepancy (D(rep, θ)) and the observed discrepancy (D(rep, θ) - D(obs, θ)). If the model is reasonable, the histogram should include 0. Figure 5.3 shows the scatterplot of 10,000 simulations from the difference of the replicated discrepancy (D(rep, θ)) and the observed discrepancy (D(rep, θ) - D(obs, θ)). Based on the result of posterior predictive p-value, p_B, and the histogram and scatterplots showed, we conclude that there are no systematic differences between the replicated data generated under the model and the observed data. Therefore, our model fits data well.

5.6 Discussion

In this chapter, we proposed a Bayesian approach for correcting the measurement error in the general multivariate linear model when the non-linear measurement errors exist in the response variables. The observed response variables are related to the true

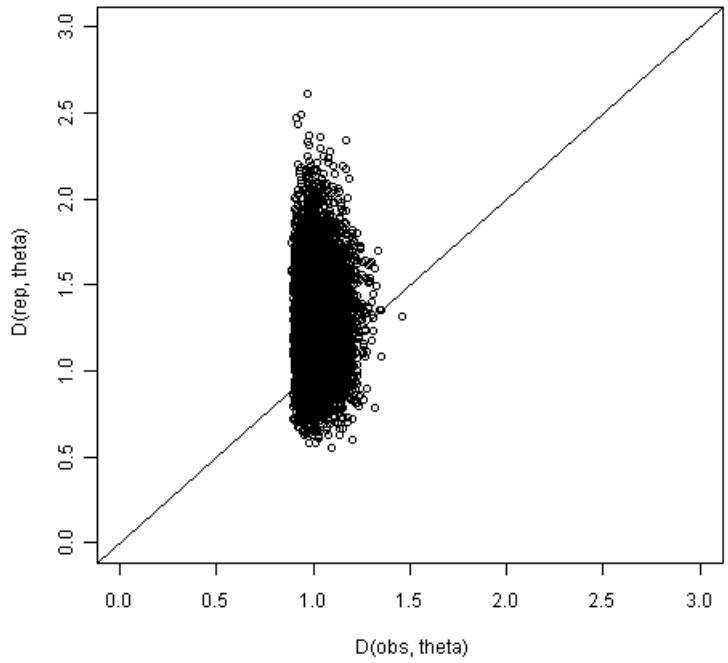


Figure 5.1: Scatterplot of replicated vs. observed discrepancies (D(rep, θ) vs. D(obs, θ)) under the joint posterior distribution; the p-value is estimated by the proportion of points above the $45^{\emptyset}$ line.

values through a non-linear regression model, and the parameters in the measurement error model are estimated by using the independent, external calibration data. We have outlined how the estimations of the parameters of interest can be carried out in a Bayesian framework using Gibbs sampling and the Metropolis Algorithm. In the Bayesian approach, we impute the values of the unobservable variable Y by sampling from their conditional distribution given all the observed data and the other parameters. Therefore, the Bayesian approach can avoid numerical integrations which may be tedious and extensive.

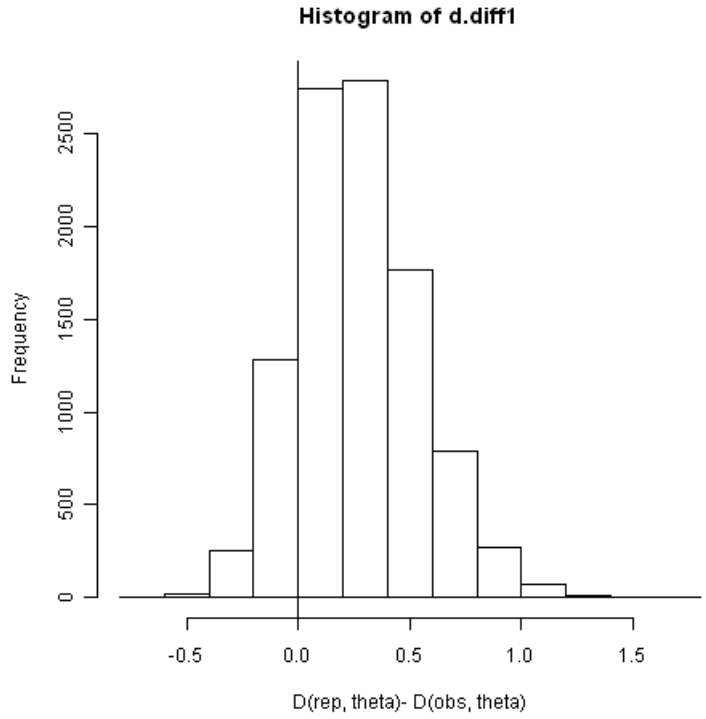


Figure 5.2: Histogram of 10,000 simulations from the difference of the replicated discrepancy (D(rep, θ)) and the observed discrepancy (D(rep, θ) - D(obs, θ)). Under the model, the histogram should include 0.

Table 5.1: Simulation results of point estimates of mean vector based on 250 replicates, k is the number of replicates used for Adjusted Value approach.

			Bayesian Approach				Adjusted Value				
Parameters	n	True value	Estimate	S.D.	Mean Std error	95% coverage	Estimate	S.D.	Mean Std error	95% coverage	k
μ_1	20	2.303	2.309	0.038	0.039	0.928	2.300	0.032	0.032	0.943	246
μ_2		2.485	2.490	0.036	0.033	0.909	2.477	0.031	0.032	0.923	
μ_3		2.708	2.711	0.035	0.035	0.917	2.700	0.036	0.035	0.931	
μ_1	100	2.303	2.302	0.019	0.027	0.936	2.300	0.014	0.014	0.925	234
μ_2		2.485	2.488	0.017	0.023	0.932	2.478	0.015	0.014	0.902	
μ_3		2.708	2.710	0.014	0.026	0.919	2.700	0.017	0.016	0.883	
μ_1	250	2.303	2.304	0.011	0.016	0.942	2.303	0.009	0.009	0.883	214
μ_2		2.485	2.487	0.010	0.015	0.927	2.479	0.008	0.010	0.767	
μ_3		2.708	2.708	0.010	0.017	0.934	2.702	0.010	0.010	0.767	

Table 5.2: Simulation results of point estimates of covariance matrix based on 250 replicates, k is the number of replicates used for Adjusted Value approach.

			Bayesian Approach		Adjusted Value		
Parameters	n	True value	Estimate	S.D.	Estimate	S.D.	k
σ_{11}	20	0.02	0.022	0.006	0.021	0.007	246
σ_{12}		0.01	0.011	0.004	0.0098	0.005	
σ_{13}		0.01	0.011	0.005	0.010	0.006	
σ_{22}		0.02	0.020	0.008	0.021	0.008	
σ_{23}		0.01	0.012	0.007	0.011	0.006	
σ_{33}		0.02	0.021	0.009	0.025	0.013	
σ_{11}	100	0.02	0.021	0.006	0.021	0.003	234
σ_{12}		0.01	0.0099	0.005	0.0098	0.002	
σ_{13}		0.01	0.001	0.004	0.011	0.002	
σ_{22}		0.02	0.021	0.009	0.021	0.003	
σ_{23}		0.01	0.011	0.005	0.011	0.003	
σ_{33}		0.02	0.020	0.009	0.024	0.006	
σ_{11}	250	0.02	0.020	0.006	0.022	0.0019	214
σ_{12}		0.01	0.011	0.004	0.011	0.0014	
σ_{13}		0.01	0.001	0.003	0.011	0.0011	
σ_{22}		0.02	0.022	0.008	0.021	0.0014	
σ_{23}		0.01	0.010	0.004	0.012	0.0013	
σ_{33}		0.02	0.021	0.006	0.025	0.0025	

Table 5.3: Estimated coverage rates of approximate 95 percent confidence regions for mean vector based on 250 replicates, k is the number of replicates used for Adjusted Value approach.

Parameters	n	Bayesian Approach	Adjusted Value	k
	20	0.9031	0.8821	246
$\underline{\mu}$	100	0.9102	0.8785	234
	250	0.9238	0.7333	214

Table 5.4: Simulation results of point estimates of mean vector based on 250 replicates with larger $\tau^2 = 0.02$

			Bayesian Approach		
Parameters	n	True value	Estimate	Mean Std error	95% coverage
μ_1	20	2.303	2.301	0.053	0.874
μ_2		2.485	2.492	0.058	0.868
μ_3		2.708	2.702	0.051	0.882
μ_1	100	2.303	2.306	0.032	0.891
μ_2		2.485	2.491	0.037	0.907
μ_3		2.708	2.710	0.041	0.885
μ_1	250	2.303	2.304	0.024	0.901
μ_2		2.485	2.488	0.021	0.924
μ_3		2.708	2.709	0.029	0.921

Table 5.5: Analysis of real data using Bayesian approach and adjusted values approach

	Bayesian Approach		Adjusted Value	
Parameters	Point estimate	95% Credible Region	Point estimate	95% CI
μ_1	8.65	(8.50, 8.85)	8.67	(8.53, 8.82)
μ_2	8.40	(8.23, 8.58)	8.40	(8.24, 8.57)
μ_3	8.44	(8.23, 8.65)	8.45	(8.27, 8.62)

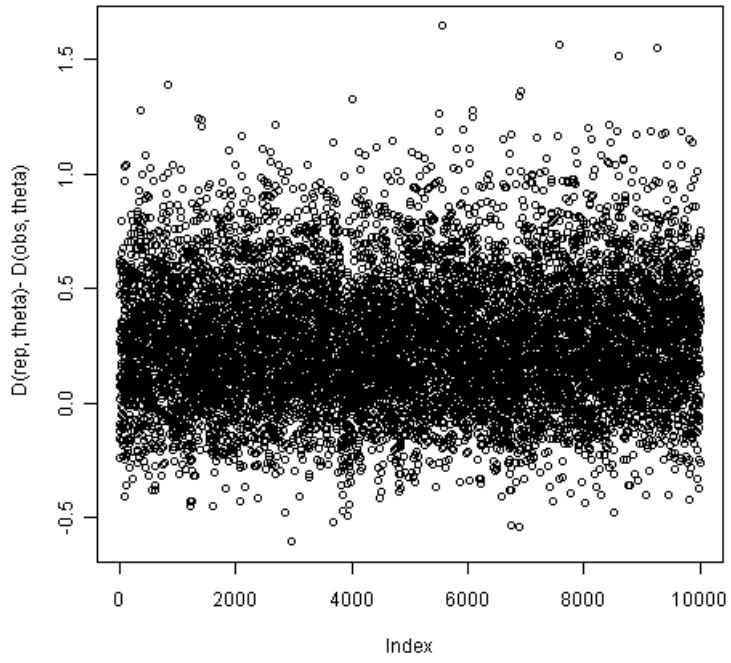


Figure 5.3: Scatterplot of 10,000 simulations from the difference of the replicated discrepancy (D(rep, θ)) and the observed discrepancy (D(rep, θ) - D(obs, θ)).

Chapter 6

SUMMARY AND FUTURE WORK

6.1 Summary

This dissertation focused on two research questions motivated by the Protective Immunity Project (PIP) conducted at the Emory Transplant Center (Larsen and Ahmed [2005]). The first research question is how to model the white cell compositions over time. The data obtained from the PIP study is compositional data with repeated measurements. Since the repeat measurements made on the same individual typically may be correlated, special attention is needed when we model the repeated compositional data. In this dissertation, we proposed a statistical model to analyze the compositional data with repeated measures. Our model extends Billheimer, Guttorp, and Fagan [2001]'s model to the longitudinal setting. The MCMC approach will be used for model inference. By using the algebra for compositions developed by Aichison and Billheimer, we can interpret the model parameter estimates and credible regions in terms of compositions. Since the proportions are the natural scale of measurement for composition data, interpretation in this way may help researchers have a better understanding of the statistical modeling results. For Aitchison's logistic-normal model in which interpretation of parameter estimates on the multivariate log-odds scale is difficult, our approach can overcome this problem. We develop a Bayesian approach for the analysis of the repeat-measured compositional data. Our results have been demonstrated that the Bayesian methodology can be used to analyzed repeat-measured compositional data. We use a Markov Chain Monte-Carlo method for model inference and show that the method is practical in high dimensional problems.

Another research question motivated in part from the PIP study is how to get the correct estimates when the measurement errors exist on the count data. In the medical studies, some variables of interest are difficult to obtain, and surrogate vari-

ables are recorded and used instead. However, these surrogate variables may contain measurement errors. In the PIP study, whole blood was passed through the flow cytometer to determine blood composition. The counts and percentage of subcategories of lymphocytes were recorded based on the cell surface markers. The variables of interest are the true counts of subcategories of lymphocytes. However, the recorded data obtained from flow cytometer may contain measurement errors. In a series of papers, Buonaccorsi [1991, 1996] and Buonaccorsi and Tosteson [1993] discussed how to correct measurement errors in response variables in the general linear model. In this dissertation, we extended Buonaccorsi's methods to the longitudinal/repeat-measures settings. We proposed the likelihood-based estimators for general multivariate linear model when the non-linear measurement errors exist in the response variables. The observed response variables are related to the true values through non-linear regression model, and the parameters in the measurement error model are estimated by using the independent, external calibration data. The pseudo-maximum likelihood estimation is used for model inference to avoid computational problems. Our proposed models provide a tool to correct for measurement errors in response variables in longitudinal data.

Finally, we proposed a Bayesian approach for correcting the measurement error in the general multivariate linear model when non-linear measurement errors exist in the response variables. The observed response variables are related to the true values through a non-linear regression model, and the parameters in the measurement error model are estimated by using independent, external calibration data. We have outlined how the estimation of the parameters of interest can be carried out in a Bayesian framework using Gibbs sampling and the Metropolis Algorithm. In this Bayesian approach, we impute the values of the unobservable variable Y by sampling

from their conditional distribution given all the observed data and the other parameters. Therefore, using a Bayesian approach can avoid numerical integrations which may be tedious and extensive.

6.2 Future Research

In chapter 3, we proposed a Bayesian model for repeated compositional data. By using the additive logratio transformation (ALR), we transformed the repeated compositional data into multivariate normal distribution data. The general multivariate linear model accounts for all the potential sources of variability that have an impact on the covariance among repeated measurements on the same individual. That is, the model does not distinguish between-subject and within-subject sources of variability. Therefore, we need assume only that the variance-covariance matrix Ω is an arbitrary positive definite matrix. We use a Bayesian approach for model inference, and it is straightforward to define the prior distribution of the unstructured variance-covariance marix Ω followed an inverse Wishart distribution, the most common prior for a covariance matrix. However, this prior is restrictive and lacks flexibility. For a k-part composition with t time points, we have a $t*(k-1) \times t*(k-1)$ variance-covariance matrix Ω. For an unstructured Ω, the number of parameters of Ω is $(t*(k-1) \times (t*(k-1)+1))/2$. When the number of parameters becomes large, estimation of the parameters in the covariance matrix becomes computationally burdensome. To overcome this problem, some dimension-reduction technics can be investigated in future research.

In chapter 4 and 5, we proposed the likelihood based method and bayesian approach for general multivariate linear model when a non-linear measurement errors exist in the response variables. The observed response variables are related to the

true variables through non-linear regression model, and the parameters in the measurement error model are estimated by using the independent, external calibration data. Both approaches assume that the measurement errors have constant variance. However, there are many situations that the variances of measurement errors is some function of the true values. The measurement errors with variances depending on true values should be considered in the future research. In this dissertation, we used independent, external calibration data to estimate the parameters of measurement error model. We can also consider the internal validation data in future research, where the true Y can be measured on some subjects of the main study data.

Appendix A

ASYMPTOTIC RESULTS

Gong and Samaniego [1981] introduced the theory and applications of pseudo-MLE. In their paper, they jointly estimate both the nuisance and primary parameters from the data and determine the statistical properties of the primary parameters. The situation from this proposal is different from the context of Gong and Samaniego [1981] because the nuisance parameters are estimated from independent calibration data. The problem can be viewed as one in which nuisance parameters (θ) are estimated from the independent calibration data and then are treated as known in calculating estimators for the primary interested parameters (ξ). Asymptotic results for the pseudo-MLE follow from Spall [1989].

Spall [1989] proves the consistency and asymptotic normality of the pseudo-MLE. Spall [1989] also provides the asymptotic variance of the pseudo-MLE. (Spall [1989], Theorem 2, page 225 in the paper). The main conditions needed are:

1. $n \to \infty, M \to \infty, n/M \to \rho, 0 < \rho < \infty$. where M is the number of entries from the calibration data.

2. The sequence $Y_1^*, \ldots, Y_M^*$ from calibration data is such that $\hat{\theta}$ is asymptotically

normal $(\theta, V(\hat{\theta}))$; see (Seber and Wild [1989], Chapter 12).

3. Let $\hat{\xi}(\theta)$ denote the estimator at the true θ, and let $\hat{\xi}$ denote the estimator at $\theta = \hat{\theta}$. At fixed θ, $\hat{\xi}$ is asymptotically normal. This can be treated, for example, via Bradley and Gart [1962].

Additional regularity conditions are easily met in the normal/normal and other continuous settings.

Bibliography

P. Abbitt and F. J. Breidt. A hierarchical model for estimating distribution profiles of soil texture. *Case Studies in Bayesian Statistics*, 5:263–278, 2001.

J. Aitchison. The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society, Ser. B*, 44:139–177, 1982.

J. Aitchison. *The Statistical Analysis of Compositional Data*. New York: Chapman & Hall, 1986.

J. Aitchison and S. M. Shen. Logistic-normal distributions: some properties and uses. *Biometrika*, 67:261–272, 1980.

B. H. Baltagi. *Econometric Analysis of Panel Data*. UK, John Wiley, 1995.

J. Barnard, R. McCulloch, and X. Meng. A natural strategy for modeling covariance matrices with application to shrinkage. *Statist. Sinica*, 10:1281–1311, 2000.

O. Barndorff-Nielsen and B. Jogensen. Some parametric models on the simplex. *Journal of Multivariate Analysis*, 39:106–116, 1991.

J. Besag, P. Green, D. Higdon, and K. Mengersen. Bayesian computation and stochastic systems (with discussion). *Statistical Science*, 10:1–66, 1995.

D. Billheimer and P. Guttorp. Spatial statistical models for discrete compositional data. *Technical Report, University of Washington, Seattle, Department of Statistics*, 1995.

D. Billheimer and P. Guttorp. Natural variablity in benthis species composition in the delaware bay. *Environmental and Ecological Statistics*, 4:95–115, 1997.

D. Billheimer, P. Guttorp, and W. F. Fagan. Statistical interpretation of species composition. *Journal of the American Statistical Association*, 96(456):1205–1214, 2001.

R. A. Bradley and J. J. Gart. The asymptotic of ml estimators when sampling from associated populations. *Biometrika*, 49:205–214, 1962.

P. J. Brown and W. Fuller. *Statistical Analysis of Measurement Error Models and Applications*. Providence, RI: American Mathematics Society, 1990.

T. M. Brunsdon and T. Smith. The time series analysis of compositonal data. *Journal of Official Statistics*, 14:237–253, 1998.

J. P. Buonaccorsi. Errors in variables with systematic biases. *Communications in Statistics, Part A-Theory and Methods*, 18:1001–1021, 1989.

J. P. Buonaccorsi. Double sampling for exact values in some multivariate measurement error problems. *Journal of the American Statistical Association*, 85:1075–1082, 1990a.

J. P. Buonaccorsi. Double sampling for exact values in the normal discriminant model with applications to binary regression. *Communications in Statistics, Part A-Theory and Methods*, 19:4569–4586, 1990b.

J. P. Buonaccorsi. Measurement error, linear calibration, and inferences for means. *Computational Statistics and Data Analysis*, 11:239–257, 1991.

J. P. Buonaccorsi. Measurement error in the response in the general linear model. *Journal of American Statistical Association*, 91:633–642, 1996.

J. P. Buonaccorsi and T. Tosteson. Correcting for nonlinear measurement error in the dependent variable in the general linear model. *Theory & Methods*, 22:2687–2702, 1993.

D. Byar and M. Gail. Workshop on errors in variables. *Statistics in Medicine*, 8: 1020–1179, 1989.

B. P. Carlin and T. A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis, 2nd ed.* London & New York: Chapman & Hall., 2000.

R. Carroll and L. A. Stefanski. Approximate quasi-likelihood estimation on models with surrogate predictors. *Journal of the American Statistical Association*, 85: 652–663, 1990.

R. J. Carroll, M. H. Gail, and J. H. Lubin. Case-control studies with errors in covariates. *Journal of the American Statistical Association*, 88:185–199, 1993.

R. J. Carroll, D. Ruppert, Stefanski L. A., and Crainiceanu C. M. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman & Hall /CRC, second edition edition, 2006.

M. J. Daniels and R. Kass. Nonconjugate bayesian estimation of covariance matrices and its use inhierarchical models. *Journal of the American Statistical Association*, 94:1254–1263, 1999.

M. J. Daniels and M. Pourahmadi. Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika*, 89:553–566, 2002.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood estimation from incomplete data via the em algorithm. *Journal of Royal Statistical Society, Ser. B*, 39:1–38, 1977.

P. J. Diggie, K-Y. Liang, and S. L. Zeger. *Analysis of Longitudinal Data*. Oxford, Clarendon Press, 1994.

P. Everson and C. Morris. Inference for multivariate normal hierarchical models. *Journal of the Royal Statistical Society B*, 62:399–412, 2000.

G. M. Fitzmaurice, N. M. Laird, and J. H. Ware. *Applied Longitudinal Analysis*. John Wiley & Sons, Inc., 2004.

J. M. Fry, T. R. L. Fry, and K. R. McLaren. Compositional data analysis and zeros in micro data. *Appl. Eocn.*, 32:953–959, 2000.

W. A. Fuller. *Measurement Error Models*. New York: John Wiley, 1987.

A. E. Gelfand, S. E. Hills, A. Racine-Poon, and A. F. M Smith. Illustration of bayesian inference in normal data models using gibbs sampling. *Journal of the American Statistical Association*, 85:972–985, 1990.

A.E. Gelfand and A.F.M Smith. Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

A. Gelman, X.L. Meng, and H.S. Stern. Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, 6:733–807, 1996.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall /CRC, second edition edition, 2004.

S. Geman and D. Geman. Stochastic relaxation, gibbs distribution and bayesian restoration of images. *IEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice.* London & New York: Chapman & Hall., 1996.

H. Goldstein. *The Design and Analysis of Longitudinal Studies*. New York: Academic Press, 1979.

G. Gong and F. J. Samaniego. Pseudo maximum likelihood estimation: Theory and applications. *The Annals of Statistics*, 9:861–869, 1981.

U. Grenander. Tutorial in pattern theory. *Division of Applied Mathematics, Brown University*, 1983.

R. Gupta and D. Richards. The history of the dirichlet and liouville distributions. *International Statistical Review*, 69:433–446, 2001.

T. Hahn. Cuba-a library for multidimensional numerical integration. *Computer Physics Communications*, 168:78–95, 2005.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109, 1970.

C. Hsiao. *Analysis of Panel Data*. UK, Cambridge University Press, 1986.

M. Iyengar and D. K. Dey. Box-fox transformations in bayesian analysis of compositional data. *Environmetrics*, 9:657–671, 1998.

M. Iyenger and D. K. Dey. A semiparametric model for compositional data analysis in the presence of covariates on the simplex. *Test*, 11:303–315, 2002.

N. M. Laird and J. H. Ware. Random effects models for longitudinal data. *Biometrics*, 38:963–974, 1982.

C. P. Larsen and R Ahmed. Immune function and biodefense in children, elderly, and immunocompromised populations. Technical report, Emory University, Department of Health and Human Services, Public Health Service, National Institutes of Health, NIH NO1-AI-50025, 2005.

T. Leonard and J. S. Hsu. Bayesian inference for a covariance matrix. *Journal of the American Statistical Association*, 20:1669–1696, 1992.

T. Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society B (Methodology)*, 44:226–233, 1982.

J. A. Martin-Fernandez, C. Barcelo-Vidal, and V. Pawlowsky-Glahn. *Zero replacement in compositional data sets*. Studies in classification, data analysis, and knowledge organization: Springer-Verlag, Berlin, 2000.

N. Metropolis and S. Ulam. The monte carlo method. *Journal of the American Statistical Association*, 44:335–341, 1949.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

J. R. Nesselroade and P. B. Baltes. *Longitudinal Research in the Study of Behavior and Development*. New York: Academic Press, 1979.

Michael G. Ormerod. *Flow Cytometry - A practical approach. 3rd edition.* Oxford University Press, Oxford, UK., 2000.

Michael G. Ormerod. *Flow Cytometry - A Basic Introduction.* De Novo Software, 2008.

V. Pawlowski and H. Burger. Spatial structure analysis of regionalized compositions. *Mathematical Geology*, 24:675–691, 1992.

M. S. Pepe. Inference using surrogate outcome data and a validation sample. *Biometrika*, 79:355–365, 1992.

W. S. Rayens and C. Srinivasan. Boxcox transformations in the analysis of compositional data. *Journal of Chemometrics*, 5:227–239, 1991a.

W. S. Rayens and C. Srinivasan. Estimation in compositional data analysis. *Journal of Chemometrics*, 5:361–374, 1991b.

B. Rosner, D. Spiegelman, and W. C. Willett. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: The case of multiple covariates measured with error. *American Journal of Epidemiology*, 132:734–745, 1990.

G. A. Seber and C. J. Wild. *Nonlinear Regression.* John Wiley: New York, 1989.

B. Smith and W. Rayens. Conditional generalized liouville distributions on the simplex. *Statistics*, 36:185–194, 2002.

J. C. Spall. Effect of imprecisely known nuisance parameters on estimates of primary parameters. *Communications in Statistics-Theory and Methods*, 18:219–237, 1989.

M. Stephens. Use of the von mises distribution to analyze continuous proportions. *Biometrika*, 69:197–203, 1982.

L. Tierney. Markov chains for exploring posterior distributions. *Ann. Statist*, 22: 1701–1762, 1994.

H. Tjelmeland and K. Lund. Baysian modeling of spatial compositional data. *Journal of Applied Statistics*, 30:87–100, 2003.

T. D. Tosteson, L. A. Stefanski, and D. W. Schafer. A measurement error model for binary and ordinal regression. *Statistics in Medicine*, 8:1139–1148, 1989.

M. A. Woodbury. Discussion of paper by hartley and hocking. *Biometrics*, 27:808–813, 1971.

R. Yang and J. O. Berger. Estimation of a covariance matrix using the reference prior. *Annals of Statistics*, 22:1195–1211, 1994.