UNIVERSITY OF CALIFORNIA
RIVERSIDE

Bayesian and Non-parametric Approaches to Missing Data Analysis

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

by

Yao Yu

September 2012

Dissertation Committee:

Dr. Jun Li, Co-Chairperson
Dr. Yaming Yu, Co-Chairperson
Dr. Subir Ghosh
Dr. Gregory J. Palardy

The Dissertation of Yao Yu is approved:

Co-Chairperson

Co-Chairperson

University of California, Riverside

Acknowledgments

I would like to take this opportunity to express my deepest appreciation for my advisors, Dr. Jun Li and Dr. Yaming Yu, who have inspired me with the thoughtful advice, encouraged me to take challenges, and supported me in all ways throughout my graduate study. I thank them for sharing with me their talented expertise, insightful suggestions and showing me the path towards success. I not only learn the knowledge and skills for my future work, but also the right attitude to face problems and the wisdom of life.

I am sincerely grateful to Dr. Subir Ghosh and Dr. Gregory J. Palardy for kindly agreeing to serve as my committee and for their precious time and valuable suggestions. I also thanks to Dr. Barry C. Arnold, Dr. Daniel Jeske, Dr. Keh-Shin Lii, Dr. James Flegal and Dr. Changxuan Mao for their profound knowledge and excellent guidance on my study of statistics. My thanks also go to Dr. Xinping Cui, Dr. Aman Ullah and Dr. Rollanda O'Connor for their delicate work as my oral qualifying committee.

My special thanks to my family and my friends, for always being there, supporting me ceaselessly and unconditionally.

To my parents, Xiaohe Liu and Weimin Yu for all the support.

ABSTRACT OF THE DISSERTATION

Bayesian and Non-parametric Approaches to Missing Data Analysis

by

Yao Yu

Doctor of Philosophy, Graduate Program in Applied Statistics
University of California, Riverside, September 2012
Dr. Jun Li, Co-Chairperson
Dr. Yaming Yu, Co-Chairperson

Missing data occur frequently in surveys, clinical trials as well as other real data studies. In the analysis of incomplete data, one needs to correctly identify the missing mechanism and then adopt appropriate statistical procedures. Recently, the analysis of missing data has gained more and more attention. People start to investigate the missing data analysis in several different areas. This dissertation concerns two projects. First, we propose a Bayesian solution to data analysis with non-ignorable missingness. The other one is the non-parametric test of missing mechanism for incomplete multivariate data.

First, Bayesian methods are proposed to detect non-ignorable missing and eliminate potential bias in estimators when non-ignorable missing presents. Two hierarchical linear models, pattern mixture model and selection model, are applied to a real data example: the National Assessment of Education Progress (NAEP) education survey data. The results show that the Bayesian methods can correctly recognize the missingness mechanism and provide model-based estimators which can eliminate the possible bias due to non-ignorable missing. We also evaluate the goodness-of-fit of these two proposed models using two methods: the comparison of the real data with the predictive

posterior distribution and the residual analysis by cross validation. A simulation study compares the performance of the two proposed Bayesian methods with the traditional design-based methods under different missing mechanisms and show the good properties of the Bayesian methods. Further, we discuss the three commonly used model selection criteria: the Bayes factor, the deviance information criterion and the minimum posterior predictive loss approach. Due to the complicated calculation of the Bayes factor and the uncertainty of the DIC, we conduct the last approach, which fails to correctly detect the real model structure for the hierarchical linear model.

Second, as an alternative to the fully specified model-based Bayesian method, a novel non-parametric test is proposed to detect the missing mechanism for multivariate missing data. The proposed test does not need any distributional assumptions and is proven to be consistent. A simulation study demonstrates that it has well controlled type I error and satisfactory power against a variety of alternative hypotheses.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In Statistics, missingness refers to the phenomenon of lack of data values for some variables in observations in the data collection step. In general the investigators should try their best to achieve the completeness of the data. But sometimes, we have to deal with the missing data issue either because the occurrence of missing data is unavoidable or the cost of the effort to avoid missingness is too expensive.

Missingness may occur in many real world studies. In social sciences, missing data are also called non-responses and often arise when either some questions or even the whole questionnaire is left as blank. In longitudinal studies, missingness can be intermittent missing or loss to follow-up during the long term study which may be caused by occasionally forgetting to respond or unwilling to participate.

The existence of missing values may have significant influence on the analysis of the data and therefore on the conclusion of the data analysis. When missing data are present, we may have the following issues:

1. Power and variability.

With more missing data, we will have smaller sample size, which means we will have

less statistical power for the analysis. And often since the extreme cases are more likely to be missing, we will have loss of data variability and the confidence interval will be forced to be narrower.

2. Bias.

For some circumstances, such as the situation where the participated interviewees in a survey are not a random sample of the population of interest, the bias issue exists. Bias is one of the worst effects that missingness brings. It also brings the issue of comparability of different groups and representativeness of the observed sample to the target population, as in some retrospective studies or observational studies.

Due to the above possible significant effects on the conclusion drawn from the missing data, there is a large amount of research on dealing with missing data in the literature. For example, Little and Rubin (2002) had a thorough book-length discussion about the treatment of missing data; Liang and Zeger (1987) proposed generalized estimating equations for incomplete longitudinal data; Rubin (1976) and Rubin (1987) elaborated the multiple imputation method, etc.

In modern Statistics, according to Rubin (1976) and Little and Rubin (2002), three major types of missing data mechanisms are generally accepted and used: missing completely at random (denoted as MCAR), missing at random (denoted as MAR) and not missing at random (denoted as NMAR).

Let $\mathbf{y} = (y_1, \cdots, y_n)'$ denote the complete set of the outcome variables, and $\mathbf{r} = (r_1, \cdots, r_n)'$ be the vector of missing data indicators such that $r_i = 1$ if y_i is observed, and $r_i = 0$ if y_i is missing. We note that each y_i and the corresponding r_i can also be vectors. Let $\mathbf{y}_{obs}$ and $\mathbf{y}_{mis}$ denote the observed and missing components of $\mathbf{y}$, respectively. With the above notation, the missing data mechanisms are characterized by the conditional distribution of $\mathbf{r}$ given $\mathbf{y}$, say $f(\mathbf{r}|\mathbf{y}, \boldsymbol{\phi})$, where $\boldsymbol{\phi}$ denotes some un-

known parameters.

(a)**Missing completely at random (MCAR)** denotes the mechanism that missingness does not depend on the values of the data $\mathbf{y}$, missing or observed.

$$f(\mathbf{r}|\mathbf{y}, \boldsymbol{\phi}) = f(\mathbf{r}|\boldsymbol{\phi}) \text{ for all } \mathbf{y}, \boldsymbol{\phi}$$

(b)**Missing at random (MAR)** denotes the mechanism that missingness only depends on the components $\mathbf{y}_{obs}$ of $\mathbf{y}$ that are observed, and not on the components that are missing.

$$f(\mathbf{r}|\mathbf{y}, \boldsymbol{\phi}) = f(\mathbf{r}|\mathbf{y}_{obs}, \boldsymbol{\phi}) \text{ for all } \mathbf{y}_{mis}, \boldsymbol{\phi}$$

(c)**Not missing at random (NMAR)** denotes the one that the distribution of $\mathbf{r}$ does depend on the missing values in the data $\mathbf{y}$.

The types of the missing mechanism will affect the choice of the appropriate statistical analysis procedures. If the missing mechanism is MCAR, then the observed data can be treated as an unbiased sample of the whole population, therefore the only loss is the sample size. If the missing mechanism is MAR, by some simple adjustment, we can easily get unbiased results. For these two missing mechanisms, where the missingness does not depend on the missing values, they are usually called ignorable missing. On the other hand, NMAR is also called non-ignorable missing. The analysis of data with missing not at random, where the probability of missing depends on the missing value itself, should be handled with extra caution. It is advised to use methods that are robust to missingness. The commonly used observed mean usually will be biased under NMAR. And the missing outcomes may have different values from the observed data under

NMAR. In addition, when the data structures are complicated, such as hierarchical data or multivariate data, with the presence of missingness, the identification of the missing mechanism itself is a non-trivial job, not to mention the data analysis.

In this dissertation, we consider two projects in the context of missing data analysis. In particular, in Chapter 2, we first propose model-based Bayesian estimates for a multi-level survey with non-ignorable non-responses. The two proposed Bayesian methods are applied to a real data set, the National Assessment of Education Progress (NAEP) education survey. In addition, two model adequacy check tools of the NAEP data are conducted to verify the models eligibility. Also, the simulation study shows that the performance of these two Bayesian methods can eliminate the bias, if any, compared with some design-based estimates under different missing mechanisms. Furthermore, we compared several commonly used model selection criteria in the simulation study. In Chapter 3, we propose a non-parametric test of missing completely at random for multivariate incomplete data. The test does not need any distributional assumptions and is proved to be consistent. The simulation study shows that the non-parametric test has the type I error well controlled at the nominal level and good power against a variety of alternative hypotheses.

Chapter 2

Model-based Bayesian Inferences For Multi-level Data with Non-ignorable Non-response

With the fast development in storage capability and computing power, many demographic and social researchers have the ability to conduct surveys on a nationwide or even on the worldwide level. The design of these surveys usually contain several levels. For example, the United States Census is organized to have four census regions and the four census regions contain nine divisions, which are further divided into sub-levels: states, counties, cities and families. These broader-domain surveys bring the challenge of handling incomplete multilevel data.

In this chapter we first give a brief introduction of the motivating data, the NAEP survey data and discuss the drawbacks of the existing design-based methods. Aimed to provide unbiased estimates by adjusting the possible impact of missing data, we propose to use hierarchical linear models to describe the hierarchical data structure

and to fit the model by Bayesian methods. The proposed methods are applied to the motivating data.

2.1 Motivating Data

NAEP, which stands for National Assessment of Education Progress, aims to provide the public with an objective and fair assessment of student performance. The organization has regularly collected, analyzed and reported information about how knowledgeable those students are and what they can do in various fields. For example, they evaluate the student performance in reading and mathematics, etc., by sets of questionnaires which contain related questions to acquire quantified scores. They have conducted surveys from samples of fourth-,eighth-, and twelfth- grade students for more than thirty years. (In this dissertation, we mainly focus on handling one year data.) A critical task is to estimate the average performance of students in the population based on the sampled students. A state-of-the-art multilevel sampling design is employed to ensure that the sample of students for each assessment is representative. Every participating school and every participating student is supposed to represent a portion of the population of interest. Although in different years, the sampling designs are different, the fundamental structure remains similar. Among the nationwide collection of data, we are particularly interested in the performance of the California data. Therefore, in this dissertation, we focus on one state data, the California data. The analysis of the nationwide data essentially can be treated as the combination of the analyses for each state.

Basically speaking, the NAEP sampling design for the California data contains two levels. At the first level, schools are selected with some pre-determined probability.

At the second level, a sample of students is selected from each selected school. The student samples are drawn and allocated to sessions using a computer-based system.

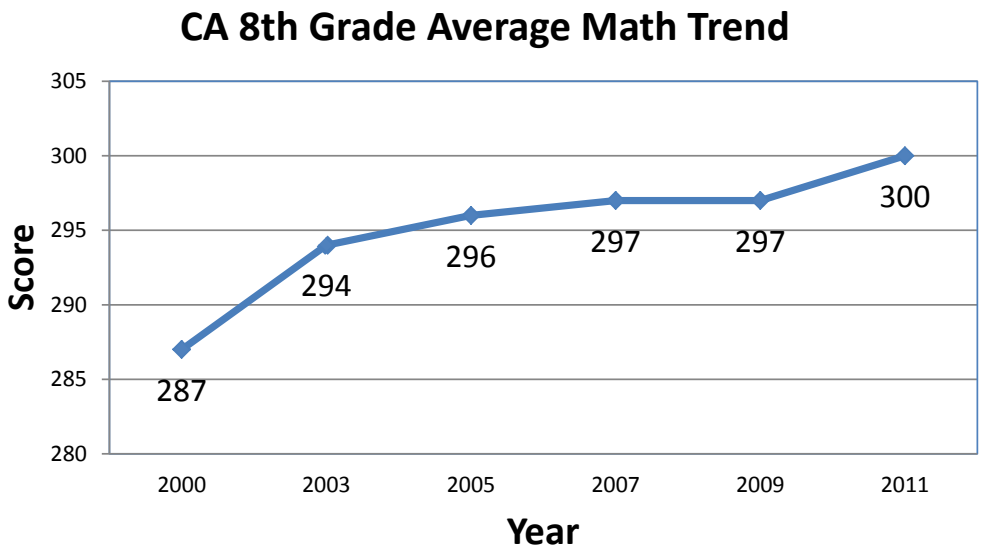

Figure 2.1: Observed trend in California 8th-grade NAEP mathematics average scores. Data from NAEP, Grade 8 Mathematics Scores

Figure 2.1 shows the observed trend in 8th-grade mathematics average scores for California students from year 2000 to year 2011. At the second level of the above multilevel sampling design for the California NAEP data, some of the sampled students within the participating schools may fail to show up for the assessment, and thus are considered as non-responses at the student level. We believe that, those students with low performance may have high probability to refuse to participate. For example, if a large proportion of students with low performance does not participate in the survey in 2003, the observed mean will be inflated and will not reflect the true trend of the data. Therefore, a method which can eliminate the bias is of great interest.

2.2 Classical Method to Deal With Non-ignorable Missing: Horvitz-Thompson Estimates

The Horvitz-Thompson (HT) (Horvitz and Thompson, 1952) estimator, which is a very popular method to measure the superpopulation mean in a stratified sample, is widely used in multi-level sampling designs. The basic idea is to weigh the outcomes by the inverse of their probabilities of selection.

Suppose the population which we are interested in consists of B schools with the ith school having N_i students, and the total number of students in the population is $N = \sum_{i=1}^{B} N_i$. Let y_{ij} denote the assessment score of student j in school i, and let

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{B} \sum_{j=1}^{N_i} y_{ij}$$

denote the population mean.

Suppose in the survey, at the first level, a sample of b of the B schools is selected; at the second level, n_i students are selected from the N_i students in the selected ith school. Each selection is performed with some pre-determined probabilities. For such a two-level sample with complete responses, the HT estimator is defined as

$$\hat{y}_{HT} = \frac{\sum_{i=1}^{b} \sum_{j=1}^{n_i} w_{ij} y_{ij}}{\sum_{i=1}^{b} \sum_{j=1}^{n_i} w_{ij}}$$

with

$$w_{ij} = 1/\pi_{ij} \qquad \text{and} \qquad \pi_{ij} = \pi_i \pi_{j|i}$$

where

π_{ij} is the selection probability of student j of school i;

π_i is the selection probability of school i; and

$\pi_{j|i}$ is the selection probability of student j when school i is selected.

In the NAEP survey data, the schools are selected with probabilities which are proportional to the school sizes and the students within selected schools are selected with the probabilities which are inversely proportional to the school size, so that overall the students across different schools are equally likely to be selected. In this case, we have $\pi_i = \frac{N_i}{N}$ and $\pi_{j|i} = \frac{n_i}{N_i}$, if the ith school is selected.

In reality, a survey or some other research study without non-response hardly occurs or may cost too much. For our data, some students in the selected school may refuse to answer the questionnaire. In order to provide an unbiased estimate, we need to correctly account for those missingness.

Suppose only m_i students respond ($m_i \leq n_i$) in the selected ith school. The student non-response adjustment is obtained by multiplying the student selection probability, $\pi_{j|i}$, by the observed student response rate in school i, $i = 1, \cdots, b$, which is given by

$$\pi^*_{j|i} = \frac{m_i}{n_i}\pi_{j|i}$$

where $\frac{m_i}{n_i}$ is the observed student response rate in the selected ith school.

Therefore, the non-response adjusted HT estimator is given by

$$\hat{y}^*_{HT} = \frac{\sum_{i=1}^{b}\sum_{j=1}^{m_i} w^*_{ij} y_{ij}}{\sum_{i=1}^{b}\sum_{j=1}^{m_i} w^*_{ij}}$$

with

$$w_{ij}^* = 1/\pi_{ij}^* \text{ and} \qquad\qquad \pi_{ij}^* = \pi_i \pi_{j|i}^*$$

where $\pi_{j|i}^*$ is defined as above.

The HT estimate approach is simple and easy to implement. When the data are complete or the response probabilities are constant among all the schools, the HT estimator is unbiased. But there are certain drawbacks for the HT estimator. First, when response rates are highly variable, the HT estimator will have a large variance. In the extreme case, in a school, none of the selected students participate in the survey, then the HT estimator is even undefined. Ways to solve this problem are to compute HT estimator by using only schools with respondents, or to combine schools with no respondents with other schools with similar background, but both approaches may lead to a biased result. Secondly, the HT estimator may be biased when non-ignorable non-responses occur. For the NAEP data, the non-responses at student level may depend on the value of the outcome variable. Those students who have low assessment scores may be unwilling to take the survey. Therefore, the missing values are not missing randomly from the population. HT estimators are unbiased only when the non-response mechanism at the student level is ignorable (see Little and Rubin, 2002). More specifically, if the response probability of student j in school i depends on the value of y_{ij} or some other student level characteristic variable(s), then the HT estimator is biased. This is a common drawback of design-based estimators. In the next section we will describe model-based methods to address these limitations.

2.3 Hierarchical Linear Models

Due to the complexity of the data structure, hierarchical linear modeling (HLM) (Bryk and Raudenbush, 1992) is often used in multilevel data analysis. Compared with the above design based estimators, it has the following advantages:

1.The feature of HLM provides the model with the flexibility to handle more sophisticated cases. One model can be built for each level. What's more, the assessment score of a student in a particular school can be predicted.

2.Additional information can be included, such as the student missing indicators. With proper modeling accounting for the missingness indicators, the estimator will be less biased.

3.If the data is complete and the model is correctly specified, the estimator from HLM is consistent and BLUE (best linear unbiased estimator) (Scott and Smith, 1969)

4.When sample size is large and a non-informative prior is used, results are comparable to design-based estimators.

2.3.1 Hierarchical Linear Models for NAEP Data

For the NAEP data, the basic model is built as follows.

$$\text{Level 1(student level): } y_{ij} = \beta_{0i} + \sum_{l \in L_2} \beta_{l\ i} X_{l\ ij} + \epsilon_{ij}$$

$$\text{Level 2(school level): } \beta_{0i} = \beta_0 + \sum_{l \in L_1} \beta_l Z_{l\ i} + \epsilon_i$$

where

y_{ij} is the assessment score for student j in school i;

β_{0i} is the mean score for school i;

$X_{l\ ij}$ is the l-th student characteristic for student j in school i, $l \in L_2$, where L_2 is the index set of $X_{l\ ij}$, which is thought to be related with y_{ij};

$\beta_{l\ i}$ is the regression coefficient associated with each $X_{l\ ij}$, $l \in L_2$;

ϵ_{ij} is the random error at the student level, and is assumed to be independently and normally distributed with mean 0 and a common variance σ_2^2 for all students in the population;

β_0 is the grand mean, adjusted for the school level covariates $Z_{l\ i}$;

$Z_{l\ i}$ is the l-th school characteristic for school i, $l \in L_1$, which is thought to be related with β_{0i};

β_l is the regression coefficient associated with each $Z_{l\ i}$, $l \in L_1$;

ϵ_i is the random error at the school level, and is assumed to be independently and normally distributed with mean 0 and a common variance σ_1^2 for all schools in the population. Furthermore, ϵ_i and ϵ_{ij} are independent of each other.

When the data is complete, i.e. no non-response in the data, the estimator of HLM is consistent and also the best linear unbiased predictor of the population mean once the model is correctly specified (Scott and Smith, 1969). For the ignorable non-response case, the estimator of HLM will hold similar properties. However, once the non-response is non-ignorable, an estimator without taking the missing mechanism into consideration will fail to be unbiased. (see Little and Rubin, 2002).

2.3.2 Pattern Mixture Model and Selection Model

The joint distribution of the outcome y_{ij} and its corresponding missing indicator r_{ij} is of great importance, since the missing mechanism is defined by the conditional distribution of r_{ij} given y_{ij}. The fact that there are two ways to factorize the joint distribution leads to two different model structures. One way is as the product of the conditional distribution of y_{ij} given r_{ij} and the marginal distribution of r_{ij}, while the other way is as the product of r_{ij} given y_{ij} and the marginal distribution of y_{ij}. They correspond to two kinds of hierarchical model structures, Pattern Mixture model and Selection model namely. Denote all the y_{ij} as $\mathbf{Y}$ and all the corresponding missing indicator r_{ij} as $\mathbf{R}$. The pattern mixture model (Glynn, Laird, and Rubin, 1986; Little, 1993) has the following form:

$$f(r_{ij}, y_{ij}|\mathbf{X}_{ij}, \boldsymbol{\gamma}, \boldsymbol{\phi}) = f_{Y|R}(y_{ij}|r_{ij}, \mathbf{X}_{ij}, \boldsymbol{\gamma}) \times f_R(r_{ij}|\boldsymbol{\phi})$$

where $\boldsymbol{\gamma}$ are parameters that y_{ij} may depend on, and $\boldsymbol{\phi}$ are parameters that r_{ij} may depend on.

The Selection model, which has been used extensively in the literature (Little and Rubin, 2002), has the following form:

$$f(r_{ij}, y_{ij}|\mathbf{X}_{ij}, \boldsymbol{\upsilon}, \boldsymbol{\delta}) = f_Y(y_{ij}|\mathbf{X}_{ij}, \boldsymbol{\upsilon}) \times f_{R|Y}(r_{ij}|y_{ij}, \boldsymbol{\delta})$$

where $\boldsymbol{\upsilon}$ and $\boldsymbol{\delta}$ are corresponding parameters for y_{ij} and r_{ij} respectively.

We illustrate how to use the pattern mixture model and the selection model to fit the data in the following two sections.

2.4 The Proposed Pattern Mixture Model

2.4.1 Pattern Mixture Model Implemented by Using Missing Latent Variable

Before we illustrate how to fit the data by using pattern mixture model, we first demonstrate how the latent variable works in the pattern mixture model.

Latent variables are "hidden variables". They are not directly observed, but rather are inferred from some other observed variables. For example, let r take value 1 if someone buys a house, 0 if not. We can imagine a continuous variable u which reflects the desire someone wants to buy a house. If the desire is "high enough", u is greater than some threshold c, say 0, someone buys a house.

$$u \geq 0 \Rightarrow r = 1$$

Otherwise, not buy.

$$u < 0 \Rightarrow r = 0$$

u is called the latent variable in this example. It may depend on some explanatory variable such as income.

$$u = \beta_0 + \beta_1 \text{Income} + \epsilon$$

The latent variable technique is a natural way to describe the relationship of the data. Originally, the patten mixture model can mix a limited number of models with different statuses of a categorical outcome. Such as in the above case, we can have two models, one model for the status of $r = 1$ (buy a house) and the other model for the status $r = 0$ (not buy a house). But the decision making process is a stochastic process.

If we can repeat the house buying process countless times, by assuming the desire of buying a house within a range of -100 and 100, the person with that desire equal to 60 will buy a house less times than a person with that desire equal to 80. Crudely dividing the data into two subsets will make the above difference being ignored. With the help of latent variables, we can treat the categorical outcome as a random realization instead of a fixed result, so that we can add the uncertainty into the model. Then the extended pattern mixture model is capable of describing the data structure by the continuous latent variable u instead of by the discontinuous categorical outcome r.

To handle the student-level non-ignorable non-responses, inspired by the Tobit model (Amemiya, 1984), we introduce latent variables to describe the relationship between the data of interest and their missing mechanism. Intuitively, people would like to use the missing indicator as a covariate in the model to describe the relationship between the dependent variable and the missing mechanism. Here, we use the continuous latent variable for the reason above.

For notation simplicity, we first assume that all of the students share a common slope of the latent variables and no explaining covariates at both student and school levels. This model can be easily modified to suit the more complicated situation. For a more general model, the strategy to solve the model is similar.

$$\text{Level 1(student level): } y_{ij} = \beta_{0i} + \alpha u_{ij} + \epsilon_{ij}$$

$$u_{ij} = \chi_i + e_{ij}$$

$$\chi_i = \chi + \zeta_i$$

$$r_{ij} = \begin{cases} 1 & \text{if } u_{ij} > 0, \\ 0 & \text{otherwise.} \end{cases}$$

where

y_{ij}, β_{0i} and ϵ_{ij} are the same as in the basic HLM;

u_{ij} is the latent variable for student j of school i, and is assumed to follow a normal distribution with mean χ_i and variance 1, when χ_i is given;

α is the regression coefficient associated with the latent variable u_{ij} for student j in school i;

ζ_i is the random error for χ_i, and is assumed to be independently and normally distributed with mean 0 and a common variance ω^2 for all the schools;

χ_i is the mean of u_{ij} for school i, and is assumed to follow a normal distribution with mean χ and variance ω^2, when χ is given;

χ is the grand mean of χ_i;

r_{ij} is the student response indicator with $r_{ij} = 1$ for responding students and $r_{ij} = 0$ for non-responding students.

In the above model, the outcome variable y_{ij} is modeled as a function of the latent variable u_{ij}. Those u_{ij}s are used to characterize the non-response mechanism in the sense that when u_{ij} is positive, the student responds; otherwise, the student does not respond. Therefore, if α is positive, for the participating students, $u_{ij} > 0$, they have higher assessment scores and on the other hand, the students who do not respond have negative u_{ij} values and thus tend to have lower scores. Overall, the non-response mechanism r_{ij} is related to the outcome variable y_{ij}, and the non-response is non-ignorable. The analysis for $\alpha < 0$ is similar, where the low performance ones are more likely to answer the survey. Therefore, after fitting this hierarchical linear model, the test for whether the missingness is ignorable or not is equivalent to test whether α is 0 or not. If $\alpha = 0$, the outcome variable $\mathbf{Y}$ will be independent of the latent

variable $\boldsymbol{u}$, furthermore will be independent of the missing indicator $\mathbf{R}$, which means the meaningness is ignorable. Otherwise, the missingness is NMAR.

Then, at the school level, the model does not include non-response, it is the same as regular linear model:

$$\text{Level 2(school level): } \beta_{0i} = \beta_0 + \epsilon_i$$

where

β_0 is the grand mean for student assessment scores;

ϵ_i is the random error at the school level, and is assumed to be independently and normally distributed with mean 0 and a common variance σ^2 for all schools in the population. Furthermore, ϵ_{ij}, ζ_i and ϵ_i are independent of each other.

2.4.2 Gibbs Sampler for the Pattern Mixture Model

For illustration purpose, the complete hierarchical linear model we propose above can be written in the following way.

$$\text{For level 1(student level): } \left[y_{ij}|\beta_{0i}, \alpha, u_{ij}, \sigma_2^2\right] = N\left(\beta_{0i} + \alpha u_{ij}, \sigma_2^2\right)$$

$$[u_{ij}|\chi_i] = N\left(\chi_i, 1\right)$$

$$r_{ij} = \begin{cases} 1 & \text{if } u_{ij} > 0, \\ 0 & \text{otherwise.} \end{cases}$$

$$\left[\chi_i|\chi, \omega^2\right] = N\left(\chi, \omega^2\right)$$

$$\text{For level 2(school level): } \left[\beta_{0i}|\beta_0, \sigma_1^2\right] = N\left(\beta_0, \sigma_1^2\right)$$

This random effects HLM can be fitted by maximum likelihood. However, a more convenient way is to use Bayesian methods. The Bayesian methods make inference based on the posterior distribution, which is a combined knowledge of prior and likelihood. The prior may be the knowledge from previous studies or the knowledge from experts. In this study, we add non-informative priors to the parameters, and then simulate draws from the posterior distribution. This method will bring us asymptotically equivalent results to maximum likelihood (Gelman et al., 2004). The above HLM is implemented by Gibbs sampler (Gelfand et al., 1990 and Gilks et al., 1996), which will be described in details as follows.

We use non-informative priors for the "mean" parameters and diffuse inverse gamma priors for the errors. So we assume priors for the parameters are of the form

$$\left[\beta_0, \sigma_1^2\right] \propto IG(a_1, b_1) \propto \left(\sigma_1^2\right)^{-a_1} exp\left(-\frac{b_1}{\sigma_1^2}\right)$$
$$\left[\alpha, \sigma_2^2\right] \propto IG(a_2, b_2) \propto \left(\sigma_2^2\right)^{-a_2} exp\left(-\frac{b_2}{\sigma_2^2}\right)$$
$$\left[\chi, \omega^2\right] \propto IG(a_3, b_3) \propto \left(\omega^2\right)^{-a_3} exp\left(-\frac{b_3}{\omega^2}\right)$$

with $a_1 = b_1 = a_2 = b_2 = a_3 = b_3 = 0.1$. The reason we choose them all equal to 0.1, is that in the posterior distributions for the errors, the information in likelihood can be the dominant factor compared with the information in the prior.

The implementation of MCMC chain involving latent variables should be approached with extra caution to avoid an unidentifiability issue. By using the normally distributed latent variables, it is equivalent to model the missing indicators by a probit regression model. And with the help of latent variables, we notice that the difference

between probit regression model and logistic regression model is the distribution of the latent variables. If the latent variables are logit distributed, then it is equivalent to model them with logistic regression. Albert and Chib (1993) proposed to use the truncated normal sampling technique to implement the Gibbs sampler for a probit model for binary responses. Intuitively, the probit model is a regression where only the sign of the dependent variable u is observed.

Setting the censoring threshold at 0 is arbitrary (any non-zero threshold will be offset by a corresponding shift in the intercept). To overcome the scale invariance problem, we set the scale of u_{ij} to be 1 since we only care about the sign of u_{ij}. Furthermore, the introduction of latent variables may bring large auto-correlations in the Bayesian MCMC chain. We need to wait long enough to make sure the independence of the MCMC draws. Because the direct sampling from the posterior distribution of all the parameters at the same time is difficult, we use Gibbs sampler, which is a simple case of MCMC chain to fit the model. The Gibbs sampler will update the parameter by using the conditional distribution for each of the parameter one at a time. After the chain converges, the procedure will generate a set of parameters which can be treated as they are from the joint posterior distribution. Details of fitting these models using Gibbs sampler are given in the following.

Consider the mth iteration. Let $\mathbf{Y}_{obs}$ and $\mathbf{Y}_{mis}$ denote values of the survey outcome $\mathbf{Y}$ for respondents and non-respondents, let

$$\mathbf{u} = (u_{11}, \cdots, u_{1n_1}, \cdots, u_{ij}, \cdots, u_{bn_b})'$$

denote values of the student-level latent variable. The first step is called "data augmentation" (Tanner and Wong, 1987). All the derivations are given in Appendix A.

STEP 1:

(I) First, when $r_{ij} = 0$, which implies the jth student in the selected school i does not participate in the survey, the latent variable u_{ij} and the missing outcome y_{ij} can be augmented as follows.

$$\left[u_{ij}|\mathbf{Y}_{obs}, r_{ij} = 0, \beta_{0i}, \alpha, \sigma_2^2, \chi_i\right] \sim TN_{[u_{ij}<0]}\left(\chi_i, 1\right)$$

$$\left[y_{ij}|\mathbf{Y}_{obs}, u_{ij}, r_{ij} = 0, \beta_{0i}, \alpha, \sigma_2^2\right] \sim N\left(\beta_{0i} + \alpha u_{ij}, \sigma_2^2\right)$$

The u_{ij} for the missing values have to be updated before the missing outcome variables y_{ij}, since it is not completely conditional on all the parameters, but rather is marginalized over y_{ij}. This situation is also called collapsed Gibbs sampling (Liu, 1994), which has been shown to result in more stable MCMC chains.

Since the value of u_{ij} for the missing values only affect the distribution of the missing y_{ij} itself, the collapsed Gibbs sampling does not have any influence on the other parameter update steps.

(II) When $r_{ij} = 1$, i.e. y_{ij} is observed. $u_{ij}(> 0)$ is drawn from the following left-truncated normal distribution:

$$\begin{aligned}&\left[u_{ij}|\mathbf{Y}_{obs}, r_{ij} = 1, \beta_{0i}, \alpha, \sigma_2^2, \chi_i\right] \\ \propto &TN_{[u_{ij}>0]}\left(\chi_i + \frac{\alpha(y_{ij} - \beta_{0i} - \alpha\chi_i)}{\alpha^2 + \sigma_2^2}, \frac{\sigma_2^2}{\alpha^2 + \sigma_2^2}\right)\end{aligned} \tag{2.1}$$

STEP 2: For these school level means, β_{0i} and χ_i are drawn from

$$\begin{aligned}
&\left[\beta_{0i}|\mathbf{Y}_{obs},\mathbf{Y}_{mis},\alpha,\mathbf{u}_i,\sigma_2^2,\beta_0,\sigma_1^2\right] \\
\propto &N\left(\frac{n_i\sigma_1^2\left(\bar{y}_i-\alpha\bar{u}_i\right)+\sigma_2^2\beta_0}{n_{ij}\sigma_1^2+\sigma_2^2},\frac{\sigma_2^2\sigma_1^2}{n_i\sigma_1^2+\sigma_2^2}\right) \\
&\left[\chi_i|\mathbf{Y}_{obs},\mathbf{Y}_{mis},\mathbf{u}_i,\chi,\omega^2\right]\propto N\left(\frac{\omega^2 n_i\bar{u}_i+\chi_i}{\omega^2 n_i+1},\frac{\omega^2}{\omega^2 n_i+1}\right)
\end{aligned}$$

where $\mathbf{u}_i$ denotes the vector $(u_{i1},\cdots,u_{in_i})'$;

$\bar{\mathbf{y}}_i=\frac{\sum_{j=1}^{n_i}y_{ij}}{n_i}$; and $\bar{\mathbf{u}}_i=\frac{\sum_{j=1}^{n_i}u_{ij}}{n_i}$.

Now, with augmented complete data $(\mathbf{Y}_{obs},\mathbf{Y}_{mis},\mathbf{u},\boldsymbol{\beta}_0,)$, where $\boldsymbol{\beta}_0$ denotes the vector $(\beta_1,\cdots,\beta_b)'$, parameters are drawn alternately.

STEP 3: the slope for the missing latent variables and the random error of student level, $\left(\alpha,\sigma_2^2\right)$ can be drawn from

$$\begin{aligned}
&\left[\sigma_2^2|\mathbf{Y}_{obs},\mathbf{Y}_{mis},\boldsymbol{\beta},\mathbf{u},\right] \\
\propto &IG\left(a_2+\frac{1}{2}\left(\sum_{i=1}^{b}n_i-1\right),b_2+\frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{b_i}(y_{ij}-(\beta_{0i}+\hat{\alpha}u_{ij}))^2\right) \\
&\left[\alpha|\mathbf{Y}_{obs},\mathbf{Y}_{mis},\boldsymbol{\beta},\mathbf{u},\sigma_2^2\right]\propto N\left(\hat{\alpha},\frac{\sigma_2^2}{\sum_{i=1}^{b}\sum_{j=1}^{n_i}u_{ij}^2}\right)
\end{aligned}$$

where $\hat{\alpha}=\frac{\sum_{i=1}^{b}\sum_{j=1}^{n_i}u_{ij}(y_{ij}-\beta_{0i})}{\sum_{i=1}^{b}\sum_{j=1}^{n_i}u_{ij}^2}$;

$\boldsymbol{\beta}$ denote the vector of β_{0i}, for $i=1,\cdots,b$;

$\mathbf{u}$ denote the vector of u_{ij}, for $i=1,\cdots,b$, $j=1,\cdots,n_i$; and

$IG(\cdot)$ denotes an inverse gamma distribution.

STEP 4: (χ, ω^2) are drawn from

$$\left[\omega^2 | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\chi}\right] \propto IG\left(a_3 + \frac{b-1}{2}, b_3 + \frac{1}{2}\sum_{i=1}^{b}(\chi_i - \bar{\chi})^2\right)$$

$$\left[\chi | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\chi}, \omega^2\right] \propto N\left(\bar{\chi}, \frac{\omega^2}{b}\right)$$

where $\boldsymbol{\chi}$ denotes the vector of χ_i, for $i = 1, \cdots, b$;

$\bar{\chi}_0 = \frac{1}{b}\sum_{i=1}^{b}\chi_i$; and

$\left[\bar{\chi} | \chi, \omega^2\right] \sim N\left(\chi, \frac{\omega^2}{b}\right)$.

STEP 5: (σ_1^2, β_0) are drawn from the conditional distributions

$$\left[\sigma_1^2 | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\beta}\right]$$

$$\propto IG\left(a_1 + \frac{b-1}{2}, b_1 + \frac{1}{2}\sum_{i=1}^{b}\left(\beta_{0i} - \bar{\beta}_0\right)^2\right)$$

$$\left[\beta_0 | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\beta}, \sigma_1^2\right] \propto N\left(\bar{\beta}, \frac{\sigma_1^2}{b}\right)$$

where $\boldsymbol{\beta}$ denotes the vector of β_{0i}, for $i = 1, \cdots, b$; and

$\bar{\beta}_0 = \frac{1}{b}\sum_{i=1}^{b}\beta_{0i}$.

The above five steps are the complete procedure for one iteration. Starting from some appropriate initial values, we follow this procedure to update parameters one by one. We iterate this procedure several times until the chain converges.

The choice of the initial values is a critical issue in MCMC chain analysis. Some people propose to use the estimates from some traditional methods as the initial values. In the pattern mixture model, there is no reference to the parameter α. And we find that, for some initial values, the MCMC chain does not work correctly. It may be difficult to converge for some initial values. We demonstrate our point in the following

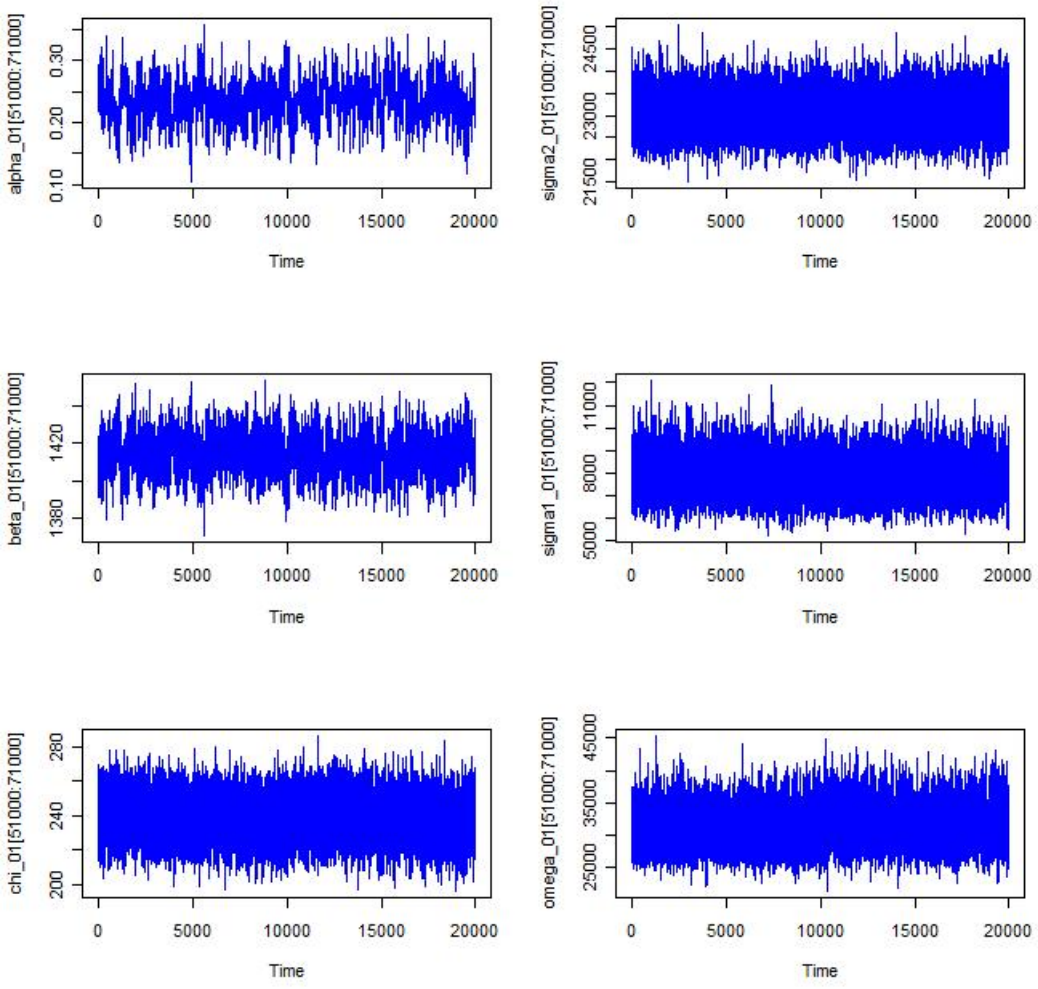

Figure 2.2: Ordinary Gibbs sampler for pattern mixture model with the initial value of α equal to 1

simulation study. Suppose we have a population with 30% of the schools contain missing values, and the overall missing percentage is 3%. The true value for α is 70. Figure 2.2 shows the trajectory plots of the parameters of interest for the ordinary Gibbs sampler for pattern mixture model with the initial value of α equal to 1. We can see that all the parameters converge perfectly, but to wrong values. For example, the trajectory plot for α converges to 0.24, which is far away from the true value 70. We further start a MCMC chain with the initial value of α equal to 70. The chain behavior is the same as in Figure 2.2, unless all the parameters (including the latent variables and the missing

values) starting from the true value. Figure 2.3, which shows the trajectory plots of the ordinary Gibbs sampler for pattern mixture model with the initial value of α equal to 0, has all the parameters converge to the right places, although the chain looks not as nicely converged as in Figure 2.2.

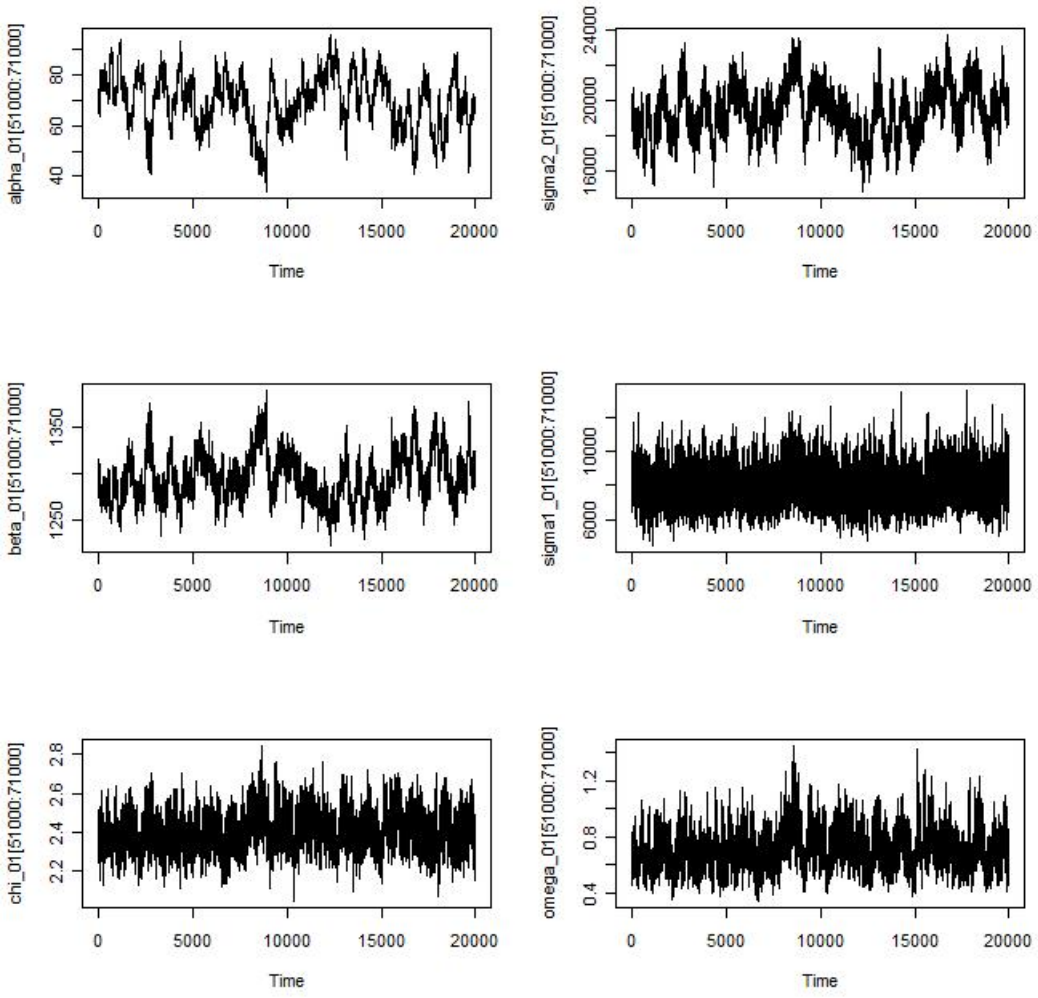

Figure 2.3: Ordinary Gibbs sampler for pattern mixture model with the initial value of α equal to 0

By a closer look at the chain, we find out that for non-zero starting values like the starting values in Figure 2.2, the latent missing variable for the observed students from the completely observed schools and the latent missing variable for the observed

students from the schools where not all the students are observed have different means. The difference between the two means are getting bigger until it reaches some point, which results in a complete separation in the data. We believe the reason is that in the mean part of the conditional posterior distribution in equation (2.1), a non-zero initial value of α will bring a cumulation part, $\frac{\alpha(y_{ij}-\beta_{0i}-\alpha\chi_i)}{\alpha^2+\sigma_2^2}$, which can never be offset in the future draws if all the students in this school are observed. In the schools where all the students participate in the survey, u_{ij}'s all go up to a huge number around 300, while in the schools some students refuse to participate, the u_{ij}'s for the observed students and the missing students are all around 0. This brings a complete separation between the schools that are fully observed and the schools where some of the students are missing. What we expect is that those observed students, no matter which schools they are from, their u_{ij} value should be comparable. In order to solve this problem, we introduced the empirical Bayesian method to put a threshold for the prior of u_{ij}.

When the total number of students is large enough, by the law of large numbers, the hierarchical structure of u_{ij} leads to the following two equations.

$$\frac{\sum_{i=1}^{b}\sum_{j=1}^{m_i} r_{ij}}{\sum_{i=1}^{b} m_i} = P(\tilde{\theta} > 0)$$
$$\frac{\sum_{i=1}^{b}\frac{\sum_{j=1}^{m_i} r_{ij}}{m_i}}{b} = P(\tilde{\chi} > 0)$$

where $\tilde{\theta}$ has identical distribution as u_{ij} and $\tilde{\chi}$ has identical distribution as χ_i. Marginalized over χ_i,

$$u_{ij} \sim N(\chi, 1+\omega^2) \tag{2.2}$$

Then

$$P(\tilde{\theta} > 0) = P(\frac{\tilde{\theta} - \chi}{\sqrt{1+\omega^2}} > \frac{0-\chi}{\sqrt{1+\omega^2}}) = 1 - \Phi\left(\frac{\chi}{\sqrt{1+\omega^2}}\right)$$
$$P(\tilde{\chi} > 0) = P(\frac{\tilde{\chi} - \chi}{\sqrt{\omega^2}} > \frac{0-\chi}{\sqrt{\omega^2}}) = 1 - \Phi\left(\frac{\chi}{\sqrt{\omega^2}}\right)$$

solving the above two equations for χ and ω^2, we can get

$$\hat{\chi} = \frac{|A|B}{\sqrt{B^2 - A^2}}$$
$$\hat{\omega^2} = \frac{A^2}{B^2 - A^2}$$

where

$$A = \Phi^{-1}\left(1 - \frac{\sum_{i=1}^{b}\sum_{j=1}^{m_i} rij}{\sum_{i=1}^{b} m_i}\right)$$
$$B = \Phi^{-1}\left(1 - \frac{\sum_{i=1}^{b}\frac{\sum_{j=1}^{m_i} r_{ij}}{m_{ij}}}{b}\right)$$

By using the estimate from the data, we could set the threshold to be within 3 standard deviations of the mean. (We note that the result is not sensitive for the choice of 3 standard deviations of the mean. We can also choose a wider boundary for the latent variables as long as there are a threshold for the latent variables.) Referring to equation (2.2), the thresholds of u_{ij} are $\hat{\chi} - 3\sqrt{1+\hat{\omega}^2}$ and $\hat{\chi} + 3\sqrt{1+\hat{\omega}^2}$. After setting up the lower and upper boundaries for the missing latent variables, the possible complete separation between the completely observed schools and the schools with missing value disappears. In Figure 2.4, we show the trajectory plots of the MCMC chain with the Bayesian empirical boundaries on the missing latent variables. The initial value of α is

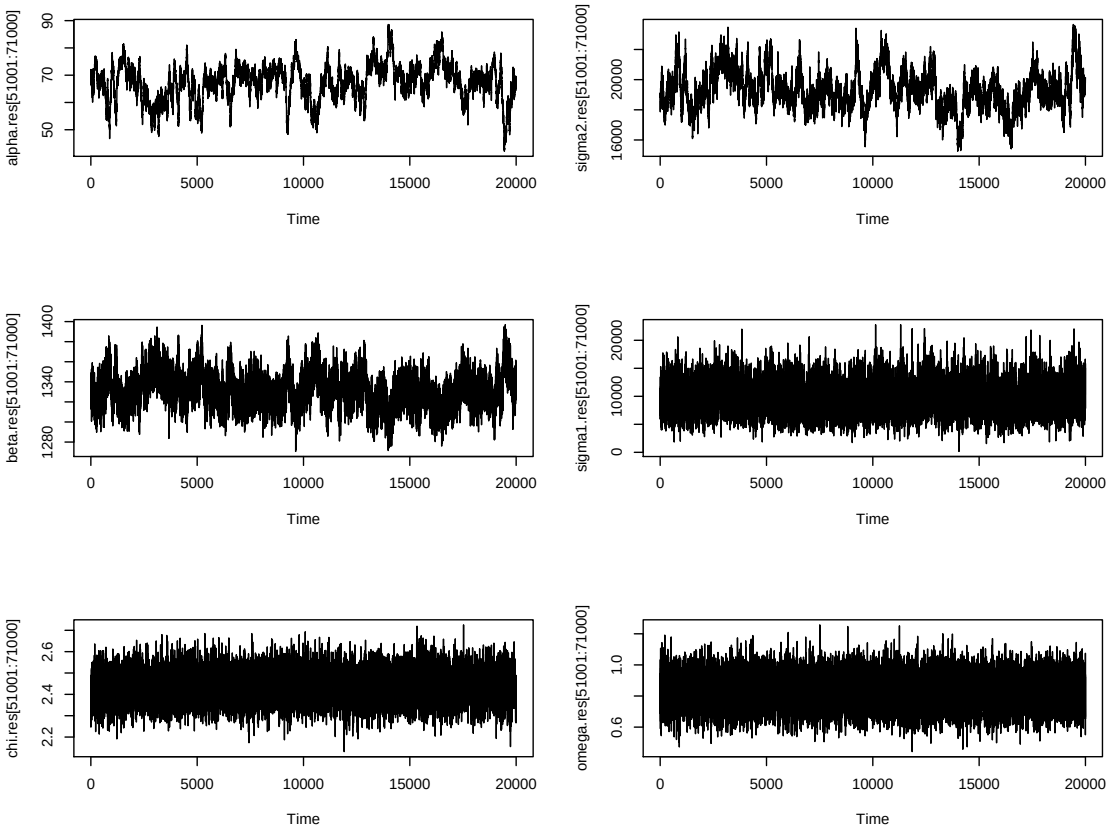

Figure 2.4: Ordinary Gibbs sampler for pattern mixture model with the initial value of α equal to 1 and u bounded

set to also be 1, but the chain converges to the true values.

In the above, we find that the ordinary Gibbs sampler solution for pattern mixture model has some limitations. With some initial values, the chain seems to "perfectly converge" to some wrong places. So here we propose to use the Gibbs sampler with empirical Bayesian method, which essentially uses the information from the data to put a constrain on the chain so that the limitation of using ordinary Gibbs sampler can be avoided.

2.5 The Proposed Selection Model

In this section, we illustrate how to use the selection model to fit the data. For simplicity, we assume all the students have the same slope in the logistic regression for missing indicators given the assessment score. Then the 2-level selection model is

$$\text{Level 1 (student level) } y_{ij} = \beta_{0i} + \epsilon_{ij}$$

$$r_{ij}|y_{ij}, \gamma_i, \lambda \sim Bernoulli\left\{\frac{exp\left\{\gamma_i + \lambda y_{ij}\right\}}{1 + exp\left\{\gamma_i + \lambda y_{ij}\right\}}\right\}$$

$$\gamma_i|\gamma, \omega^2 \sim N\left(\gamma, \omega^2\right)$$

where

y_{ij} is the assessment score of the jth student in ith school;

β_{0i} is the mean score of ith school;

ϵ_{ij} is the random error at student level for jth student in ith school, it is assumed to be independently and normally distributed with mean 0 and a common variance σ_2^2;

r_{ij} is the student response indicator, with $r_{ij} = 1$, if y_{ij} is observed and $r_{ij} = 0$, if y_{ij} is missing;

γ_i is the intercept in the logistic regression for ith school;

λ is the slope in the logistic regression;

γ is the grand mean for γ_i, $j = 1, \ldots, b$, respectively;

ω^2 is the variance for γ_i;

$$\text{Level 2 (school level) } \beta_{0i} = \beta_0 + \epsilon_i$$

where

β_0 is the grand mean assessment score;

ϵ_i is the random error at school level for ith school, it is assumed to be independently and normally distributed with mean 0 and a common variance σ_1^2;

For the proposed selection model, by introducing the logistic regression method, the full joint posterior distribution will no longer be in the exponential family. Furthermore, some of the conditional posterior densities do not belong to exponential family.

The full joint posterior distribution has the form

$$
\begin{aligned}
&P\left(\mathbf{Y}_{mis}, \sigma_2^2, \boldsymbol{\beta}_0, \beta_0, \sigma_1^2, \boldsymbol{\gamma}, \lambda, \omega^2 | \mathbf{Y}_{obs}\right) \propto \\
&\prod_{i=1}^{b}\left\{\prod_{j=1}^{m_i}\frac{1}{\sigma_2}exp\left\{-\frac{1}{2\sigma_2^2\left(y_{ij}-\beta_{0i}\right)^2}\right\}\frac{exp\left\{\gamma_i+\lambda y_{ij}\right\}}{1+exp\left\{\gamma_i+\lambda y_{ij}\right\}}\right\} \\
&\times\prod_{i=1}^{b}\left\{\prod_{j=m_i+1}^{n_i}\frac{1}{\sigma_2}exp\left\{-\frac{1}{2\sigma_2^2\left(y_{ij}-\beta_{0i}\right)^2}\right\}\frac{1}{1+exp\left\{\gamma_i+\lambda y_{ij}\right\}}\right\} \\
&\times\prod_{i=1}^{b}\frac{1}{\sigma_1}exp\left\{-\frac{1}{2\sigma_1^2}\left(\beta_{0i}-\beta_0\right)^2\right\}\times\prod_{l=1}^{2}\left(\sigma_l^2\right)^{-(a_l+1)}exp\left\{-\frac{b_l}{\sigma_l^2}\right\}\times\left(\omega^2\right)^{-(a_3+1)}exp\left\{-\frac{b_3}{\omega^2}\right\} \\
&\times\prod_{i=1}^{b}\frac{1}{\omega}exp\left\{-\frac{1}{2\omega^2}\left(\gamma_i-\gamma\right)^2\right\}
\end{aligned}
$$

where $\mathbf{Y}_{mis}$ and $\mathbf{Y}_{obs}$ denote the missing part and observed part of $\mathbf{Y}$ respectively;

$\boldsymbol{\beta}_0$ denotes the collection of β_{0i} for $i = 1, \ldots, b$;

$\boldsymbol{\gamma}$ denotes the collection of γ_i for $i = 1, \ldots, b$;

We show two methods, the Gibbs sampler directly based on logistic regression and the Gibbs sampler based on an approximation of the logistic regression, the robit regression, to fit the selection model in the following.

First, we derive the conditional posterior density function of the ordinary Gibbs sampler directly based on logistic regression.

STEP 1: We need to augment the missing data when $r_{ij} = 0$, which denotes the case that the jth student in the ith school did not participate in the survey.

$$\begin{aligned} & P\left(y_{ij}|r_{ij}=0,\beta_0,\sigma_2^2,\gamma_i,\lambda\right) \\ \propto & exp\left\{-\frac{1}{2\sigma_2^2}\left(y_{ij}-\beta_{0i}\right)^2\right\}\cdot\frac{1}{1+exp\left\{\gamma_i+\lambda y_{ij}\right\}} \end{aligned}$$

This conditional posterior density is not in the exponential family. Directly drawing a random variable from it is difficult. We utilize the Metroplois-Hasting algorithm (Chib and Greenberg, 1995) to approximate the original distribution by sampling from a proposal density with some acceptance rate. We construct a proposal density based on least squares estimate. Propose y_{ij}^{new} from

$$y_{ij}^{new} \sim N\left(\beta_{0i},\sigma_2^2\right)$$

Draw a random variable u from uniform distribution $U[0,1]$, then we accept y_{ij}^{new} if $log(u) < l\left(y_{ij}^{new}\right) - l\left(y_{ij}^{old}\right) - h\left(y_{ij}^{new}\right) + h\left(y_{ij}^{old}\right)$, where $l(\cdot) = logp\left(\cdot|\beta_{0i},\sigma_2^2,\gamma_i,\lambda\right)$ is the log-likelihood function and $h(\cdot)$ is the log-density of the proposal function. In this case, $l(y) - h(y) = -log\left(1+exp\left\{\gamma_i+\lambda y\right\}\right)$, so we accept y_{ij}^{new} if $log(u) \leq log\left(1+exp\left\{\gamma_i+\lambda y_{ij}^{old}\right\}\right) - log\left(1+exp\left\{\gamma_i+\lambda y_{ij}^{new}\right\}\right)$

STEP 2: We draw β_{0i} and σ_2^2.

$$
\begin{aligned}
&P\left(\beta_{0i}|\mathbf{Y}_{obs},\mathbf{Y}_{mis},\sigma_2^2,\sigma_1^2,\beta_0\right)\\
\propto&\prod_{j=1}^{n_i} exp\left\{-\frac{1}{2\sigma_2^2}\left(y_{ij}-\beta_{0i}\right)^2\right\}\times exp\left\{-\frac{1}{2\sigma_1^2}\left(\beta_{0i}-\beta_0\right)^2\right\}\\
\propto& exp\left\{-\frac{1}{2}\left(\frac{n_i\sigma_1^2+\sigma_2^2}{\sigma_1^2\sigma_2^2}\right)\left[\beta_{0i}-\frac{\sigma_1^2\sum_{j=1}^{n_i}y_{ij}+\sigma_2^2\beta_0}{n_i\sigma_1^2+\sigma_2^2}\right]^2\right\}\\
\propto& N\left(\frac{\sigma_1^2\sum_{j=1}^{n_i}y_{ij}+\sigma_2^2\beta_0}{n_i\sigma_1^2+\sigma_2^2},\frac{\sigma_1^2\sigma_2^2}{n_i\sigma_1^2+\sigma_2^2}\right)
\end{aligned}
$$

$$
\begin{aligned}
P\left(\sigma_2^2|\mathbf{Y}_{obs},\mathbf{Y}_{mis},\boldsymbol{\beta}\right)&\propto\left(\sigma_2^2\right)^{-(a_2+1)}exp\left\{-\frac{b_2}{\sigma_2^2}\right\}\\
&\times\left(\sigma_2^2\right)^{-\frac{1}{2}\sum_{i=1}^{b}n_i}exp\left\{-\frac{1}{2\sigma_2^2}\sum_{i=1}^{b}\sum_{j=1}^{n_i}\left(y_{ij}-\beta_{0i}\right)^2\right\}\\
&\propto IG\left(a_2+\frac{1}{2}\sum_{i=1}^{b}n_i,b_2+\frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{n_i}\left(y_{ij}-\beta_{0i}\right)^2\right)
\end{aligned}
$$

STEP 3: Update $\boldsymbol{\gamma}_0$ and λ.

$$
\begin{aligned}
&P\left(\begin{pmatrix}\boldsymbol{\gamma}_0\\ \\ \lambda\end{pmatrix}|\mathbf{Y}_{obs},\mathbf{Y}_{mis},\gamma,\omega^2\right)\propto\prod_{i=1}^{b}\prod_{j=1}^{n_i}\frac{exp\left\{r_{ij}\left(\gamma_i+\lambda y_{ij}\right)\right\}}{1+exp\left\{\gamma_i+\lambda y_{ij}\right\}}\\
&\times\prod_{i=1}^{b}exp\left\{\frac{1}{2\omega^2}\left(\gamma_i-\gamma\right)^2\right\}
\end{aligned}
$$

We utilize the Metropolis-Hasting algorithm here also. First, we use the estimators from the logistic regression, where school is the random effect, as the mode, $(\hat{\boldsymbol{\gamma}},\hat{\lambda})'$, and then

compute the Fisher information matrix. Let

$$
\begin{aligned}
&l(\boldsymbol{\gamma}_i, \lambda)^{'} = log(L((\boldsymbol{\gamma}_0, \lambda)^{'} | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \gamma, \omega^2)) \\
&= \sum_{i=1}^{b} \sum_{j=1}^{n_i} r_{ij} \left(\gamma_i + \lambda y_{ij}\right) - \sum_{i=1}^{b} \sum_{j=1}^{n_i} log\left(1 + exp\left\{\gamma_i + \lambda y_{ij}\right\}\right) \\
&- \sum_{i=1}^{b} \frac{1}{2\omega^2} \left(\gamma_i - \gamma\right)^2
\end{aligned}
$$

The first derivatives are

$$
\begin{aligned}
\frac{\partial l\left(\boldsymbol{\gamma}_0, \lambda\right)^{'}}{\partial \gamma_i} &= \sum_{j=1}^{n_i} r_{ij} - \sum_{j=1}^{n_i} \frac{exp\left\{\gamma_i + \lambda y_{ij}\right\}}{1 + exp\left\{\gamma_i + \lambda y_{ij}\right\}} - \frac{1}{\omega^2} \left(\gamma_i - \gamma\right) \\
\frac{\partial l\left(\boldsymbol{\gamma}_0, \lambda\right)^{'}}{\partial \lambda} &= \sum_{i=1}^{b} \sum_{j=1}^{n_i} r_{ij} y_{ij} - \sum_{i=1}^{b} \sum_{j=1}^{n_i} \frac{y_{ij} exp\left\{\gamma_i + \lambda y_{ij}\right\}}{1 + exp\left\{\gamma_i + \lambda y_{ij}\right\}}
\end{aligned}
$$

The second derivatives are

$$
\begin{aligned}
\frac{\partial^2 l\left(\boldsymbol{\gamma}_0, \lambda\right)^{'}}{\partial \gamma_i^2} &= -\sum_{j=1}^{n_i} \frac{exp\left\{\gamma_i + \lambda y_{ij}\right\}}{\left(1 + exp\left\{\gamma_i + \lambda y_{ij}\right\}\right)^2} - \frac{1}{\omega^2} \\
\frac{\partial^2 l\left(\boldsymbol{\gamma}_0, \lambda\right)^{'}}{\partial \gamma_i \partial \gamma_j} &= 0, \text{ for } i \neq j \\
\frac{\partial^2 l\left(\boldsymbol{\gamma}_0, \lambda\right)^{'}}{\partial \gamma_i \partial \lambda} &= -\sum_{j=1}^{n_i} \frac{y_{ij} exp\left\{\gamma_i + \lambda y_{ij}\right\}}{\left(1 + exp\left\{\gamma_i + \lambda y_{ij}\right\}\right)^2} \\
\frac{\partial^2 l\left(\boldsymbol{\gamma}_0, \lambda\right)^{'}}{\partial \lambda^2} &= -\sum_{i=1}^{b} \sum_{j=1}^{n_i} \frac{y_{ij}^2 exp\left\{\gamma_i + \lambda y_{ij}\right\}}{\left(1 + exp\left\{\gamma_i + \lambda y_{ij}\right\}\right)^2}
\end{aligned}
$$

Then $I = -\left.\frac{\partial^2 l((\boldsymbol{\gamma}_0, \lambda)^{'})}{\partial (\gamma_i, \lambda)^{'} \partial (\gamma_i, \lambda)}\right|_{(\boldsymbol{\gamma}_0, \lambda)^{'} = \left(\hat{\boldsymbol{\gamma}}_0, \hat{\lambda}\right)^{'}}$ We draw T_5 according to a p-variate standard t_5 distribution, and propose $(\boldsymbol{\gamma}_0^{new}, \lambda^{new})^{'} = \left(\hat{\boldsymbol{\gamma}}_0, \hat{\lambda}\right)^{'} + I^{-\frac{1}{2}} T_5$. Draw $u \sim U[0,1]$, and accept $(\boldsymbol{\gamma}_0^{new}, \lambda^{new})^{'}$ if $log(u) \leq l\left(\boldsymbol{\gamma}_0^{new}, \lambda^{new}\right)^{'} - l\left(\boldsymbol{\gamma}_0^{old}, \lambda^{old}\right)^{'} - h\left(\boldsymbol{\gamma}_0^{new}, \lambda^{new}\right)^{'} + h\left(\boldsymbol{\gamma}_0^{old}, \lambda^{old}\right)^{'}$, where $h(\cdot)$ is the log density of T_5 centered at $\left(\hat{\boldsymbol{\gamma}}_0, \hat{\lambda}\right)^{'}$ with scale $I^{-\frac{1}{2}}$

STEP 4: Update γ and ω^2 The joint distribution of γ and ω^2 is

$$\begin{aligned} & P\left(\gamma, \omega^2 | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\gamma}_0\right) \\ & \propto \prod_{i=1}^{b} \frac{1}{\omega} exp\left\{-\frac{1}{2\omega^2}\left(\gamma_i - \gamma\right)^2\right\} \times \left(\omega^2\right)^{a_3+1} exp\left\{\frac{b_3}{\omega^2}\right\} \end{aligned}$$

After integrating out γ, the marginal posterior distribution of ω^2 is

$$P\left(\omega^2 | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\gamma}_0\right) \propto IG\left(a_3 + \frac{b-1}{2}, b_3 + \frac{1}{2}\left(\gamma_i - \bar{\gamma}\right)^2\right)$$

and the conditional posterior distribution for γ is

$$P\left(\gamma | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\gamma}_0, \omega^2\right) \propto N\left(\bar{\gamma}, \frac{\omega^2}{b}\right)$$

where $\bar{\gamma} = \frac{1}{b}\sum_{i=1}^{b} \gamma_i$

STEP 5: Update β_0 and σ_1^2

$$P\left(\beta_0, \sigma_1^2 | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\beta}_0\right) \propto \prod_{i=1}^{b} \frac{1}{\sigma_1} exp\left\{-\frac{1}{2\sigma_1^2\left(\beta_{0i} - \beta_0\right)^2}\right\} \times \left(\sigma_1^2\right)^{-(a_1+1)} exp\left\{-\frac{b_1}{\sigma_1^2}\right\}$$

Integrate β_0 out to get the marginal posterior distribution of σ_1^2,

$$\begin{aligned} & P\left(\sigma_1^2 | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\beta}_0\right) \propto \left(\sigma_1^2\right)^{-\left(a_1 + \frac{b}{2} + 1\right)} exp\left\{-\frac{b_1}{\sigma_1^2}\right\} \int exp\left\{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{b}\left(\beta_0 - \beta_{0i}\right)^2\right\} d\beta \\ & \propto IG\left(a_1 + \frac{b-1}{2}, b_1 + \frac{1}{2}\sum_{i=1}^{b}\left(\beta_{0i} - \bar{\beta}\right)^2\right) \end{aligned}$$

$$P\left(\beta_0 | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\beta}, \sigma_1^2\right) = \frac{p\left(\beta_0, \sigma_1^2 | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\beta}\right)}{p\left(\sigma_1^2 | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\beta}_0\right)} \propto N\left(\bar{\beta}, \frac{\sigma_1^2}{b}\right)$$

where $\bar{\beta} = \frac{\sum_{i=1}^{b} \beta_{0i}}{b}$

As shown in the above five steps, the proposed selection model involves drawing random variables from a logistic distribution. Since logistic distribution is not in the exponential family, more specifically, it is not in a close form, we use the Metropolis-Hasting algorithm to approximate it by a t distribution with degree of freedom 5 with matching means and covariance matrices. One drawback of this approach is that, the coefficients for the logistic regression have to be updated at the same time, which may cause slow convergence. Furthermore, the choice of the initial values is critical for the selection model based on the logistic regression, especially the initial values for the coefficients in the logistic regression. Some people suggest to use the pattern mixture model to impute the data, and then use the augmented data to "estimate" the initial values in the selection model. Here, we will use an alternative method, the robit regression, to approximate the logistic regression. It will be illustrated as follows. First, the above selection model is equivalent to the following model:

$$
\begin{aligned}
f_Y: \quad & y_{ij} = \beta_{0i} + \epsilon_{ij} \\
& \beta_{0i} = \beta_0 + \epsilon_i \\
f_{R|Y}: \quad & z_{ij} = \gamma_i + \lambda y_{ij} + e_{ij} \\
& \gamma_i = \gamma + e_i
\end{aligned}
$$

Here, z_{ij} is the latent variable and the missing indicator can be determined by z_{ij} in the sense that

$$
r_{ij} = \begin{cases} 1 & \text{if } z_{ij} > 0, \\ 0 & \text{otherwise.} \end{cases}
$$

and in the model, ϵ_{ij} is the random error at student level and is assumed to be independently and normally distributed with mean 0 and a common variance σ_2^2 for all students;

ϵ_i is the random error at school level and is assumed to be independently and normally distributed with mean 0 and a common variance σ_1^2;

e_i is the random error for the latent variable at school level and is assumed to be independently and normally distributed with mean 0 and a common variance ω^2;

e_{ij} is the random error for the latent variable at student level and is assumed to be independently distributed with standard logistic distribution, whose cumulative distribution function (CDF) is logistic function, the inverse of logit function.

Based on the definition of logistic distribution, $P(e_{ij} < x) = logit^{-1}(x)$, since

$$\begin{aligned} P(r_{ij} = 1|y_{ij}) &= P(z_{ij} > 0|y_{ij}) = P(\gamma_i + \lambda y_{ij} + e_{ij} > 0) \\ &= P(e_{ij} > -(\gamma_i + \lambda y_{ij})) = P(e_{ij} < \gamma_i + \lambda y_{ij}) \\ &= logit^{-1}(\gamma_i + \lambda y_{ij}) = p_i \end{aligned}$$

the latent variable method is equivalent to the traditional logistic regression model.

We note that the logistic distribution is not in the exponential family, a more convenient way is to use t distribution with matched moments to approximate it. By matching the first four moments (the skewness of both distributions is always 0, the other three as shown in 2.3),

$$\begin{aligned} \mu &= 0 \\ \frac{n}{n-2}s^2 &= \frac{\pi^2}{3} \\ \frac{6}{n-4} &= \frac{6}{5} \end{aligned} \tag{2.3}$$

the t distribution with degree of freedom $n = 9$ and standard deviation $s = \sqrt{\frac{7}{9}\frac{\pi^2}{3}}$ is a good approximation as shown in Figure 2.5.

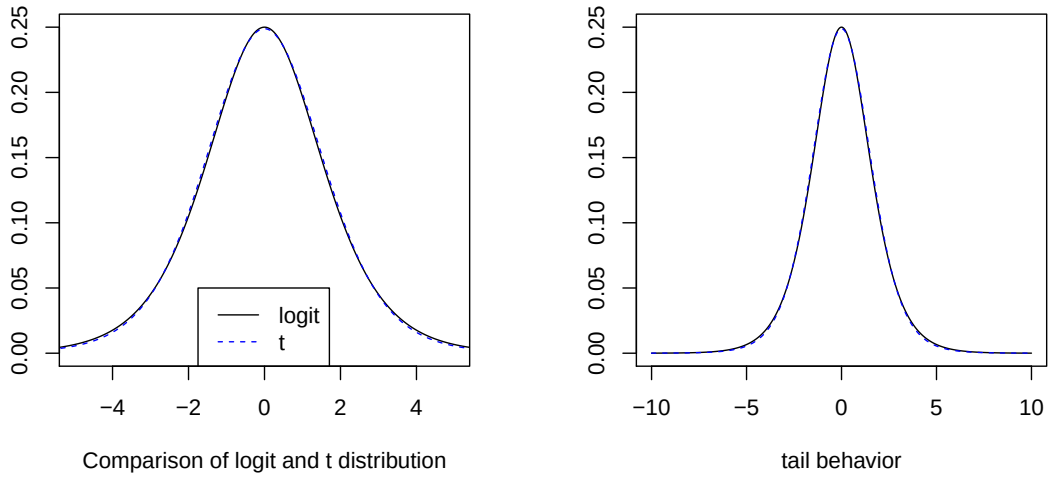

Figure 2.5: Comparison of logistic distribution and approximating t distribution

Furthermore, t distribution is the compound distribution of a normal distribution while its variance has an inverse gamma distribution. In this set up, the coefficient parameters γ_i and λ in the logistic regression can have conjugate posterior distribution with a normal prior. To be more specific,

$$t(x|\upsilon, \mu, \tau^2) \propto \int N(x|\mu, \tau^2\rho) \times IG(\rho|\frac{\upsilon}{2}, \frac{\upsilon}{2})d\rho$$

here $\upsilon = 9$ and $\tau = \sqrt{\frac{7}{9}\frac{\pi^2}{3}} \approx 1.59962$. So e_{ij} can be drawn from $N(0, \tau^2\rho)$ where ρ has distribution $IG(\frac{9}{2}, \frac{9}{2})$

STEP 1: The first step of the Gibbs sampler is to augment $\mathbf{Y}_{mis}$

$$
\begin{aligned}
&[y_{ij}|r_{ij}=0,\beta_{0i},\sigma_2^2,z_{ij},\gamma_i,\lambda,\rho] \\
&\propto exp\left\{-\frac{1}{2\sigma_2^2}\left(y_{ij}-\beta_{0i}\right)^2\right\}\times exp\left\{-\frac{1}{2\tau^2\rho}\left(z_{ij}-\gamma_i-\lambda y_{ij}\right)^2\right\} \\
&\propto N\left(\beta_{0i}+\frac{\sigma_2^2\lambda\left(z_{ij}-\gamma_i-\lambda\beta_{0i}\right)}{\lambda^2\sigma_2^2+\tau^2\rho},\frac{\sigma_2^2\tau^2\rho}{\lambda^2\sigma_2^2+\tau^2\rho}\right)
\end{aligned}
$$

STEP 2: β_{0i} and σ_2^2 can be drawn from the conditional distribution

$$
\begin{aligned}
&P\left(\beta_{0i}|\mathbf{Y}_{obs},\mathbf{Y}_{mis},\sigma_2^2,\sigma_1^2,\beta_0\right) \\
&\propto\prod_{j=1}^{n_i}exp\left\{-\frac{1}{2\sigma_2^2}\left(y_{ij}-\beta_{0i}\right)^2\right\}\times exp\left\{-\frac{1}{2\sigma_1^2}\left(\beta_{0i}-\beta_0\right)^2\right\} \\
&\propto exp\left\{-\frac{1}{2}\left(\frac{n_i\sigma_1^2+\sigma_2^2}{\sigma_1^2\sigma_2^2}\right)\left[\beta_{0i}-\frac{\sigma_1^2\sum_{j=1}^{n_i}y_{ij}+\sigma_2^2\beta_0}{n_i\sigma_1^2+\sigma_2^2}\right]^2\right\} \\
&\propto N\left(\frac{\sigma_1^2\sum_{j=1}^{n_i}y_{ij}+\sigma_2^2\beta_0}{n_i\sigma_1^2+\sigma_2^2},\frac{\sigma_1^2\sigma_2^2}{n_i\sigma_1^2+\sigma_2^2}\right)
\end{aligned}
$$

$$
\begin{aligned}
P\left(\sigma_2^2|\mathbf{Y}_{obs},\mathbf{Y}_{mis},\boldsymbol{\beta}\right)&\propto\left(\sigma_2^2\right)^{-(a_2+1)}exp\left\{-\frac{b_2}{\sigma_2^2}\right\} \\
&\times\left(\sigma_2^2\right)^{-\frac{1}{2}\sum_{i=1}^{b}n_i}exp\left\{-\frac{1}{2\sigma_2^2}\sum_{i=1}^{b}\sum_{j=1}^{n_i}\left(y_{ij}-\beta_{0i}\right)^2\right\} \\
&\propto IG\left(a_2+\frac{1}{2}\sum_{i=1}^{b}n_i,b_2+\frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{n_i}\left(y_{ij}-\beta_{0i}\right)^2\right)
\end{aligned}
$$

STEP 3: β_0 and σ_1^2 can be drawn from the conditional distribution

$$
\begin{aligned}
&P\left(\sigma_1^2|\mathbf{Y}_{obs},\mathbf{Y}_{mis},\boldsymbol{\beta}_0\right)\propto\left(\sigma_1^2\right)^{-\left(a_1+\frac{b}{2}+1\right)}exp\left\{-\frac{b_1}{\sigma_1^2}\right\}\int exp\left\{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{b}\left(\beta_0-\beta_{0i}\right)^2\right\}d\beta \\
&\propto IG\left(a_1+\frac{b-1}{2},b_1+\frac{1}{2}\sum_{i=1}^{b}\left(\beta_{0i}-\bar{\beta}\right)^2\right)
\end{aligned}
$$

$$P\left(\beta_0|\mathbf{Y}_{obs},\mathbf{Y}_{mis},\boldsymbol{\beta},\sigma_1^2\right)=\frac{p\left(\beta_0,\sigma_1^2|\mathbf{Y}_{obs},\mathbf{Y}_{mis},\boldsymbol{\beta}\right)}{p\left(\sigma_1^2|\mathbf{Y}_{obs},\mathbf{Y}_{mis},\boldsymbol{\beta}_0\right)}\propto N\left(\bar{\beta},\frac{\sigma_1^2}{b}\right)$$

where $\bar{\beta}=\frac{\sum_{i=1}^{b}\beta_{0i}}{b}$

STEP 4: For the $R|Y$ model part, we first augment latent variable z_{ij}. The original range of the missing probabilities is $(0,1)$, and we consider to fix it between $(0.001,0.999)$ for real data problems. Then the corresponding threshold for the values of z_{ij} is $(-7,7)$.

(I) For $r_{ij}=1$, generate z_{ij} from a truncated normal distribution so that $z_{ij}>0$.

$$[z_{ij}|r_{ij}=1,\mathbf{Y}_{obs},\mathbf{Y}_{mis},\gamma_i,\lambda,\rho]\propto TN_{z_{ij}>0}\left(\gamma_i+\lambda y_{ij},\tau^2\rho\right)$$

(II) For $r_{ij}=0$, generate z_{ij} from a truncated normal distribution so that $z_{ij}<0$.

$$[z_{ij}|r_{ij}=0,\mathbf{Y}_{obs},\mathbf{Y}_{mis},\gamma_i,\lambda,\rho]\propto TN_{z_{ij}<0}\left(\gamma_i+\lambda y_{ij},\tau^2\rho\right)$$

Let X be a $\left(\sum_{i=1}^{b}n_i\right)$-by-$(b+1)$ matrix, with the first b columns are

$$\begin{pmatrix}\mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1}\\ \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2}\\ \vdots & \vdots & \vdots & \vdots\\ \mathbf{0}_{n_b} & \mathbf{0}_{n_b} & \cdots & \mathbf{1}_{n_b}\end{pmatrix}_{\left(\sum_{i=1}^{b}n_i\right)\times b}$$

and the last column is $(y_{11},\cdots,y_{1n_1},\cdots,y_{bn_b})'$. And let $\Sigma_1=\tau^2\rho I_{\left(\sum_{i=1}^{b}n_i\right)\times\left(\sum_{i=1}^{b}n_i\right)}$, μ as a $(b+1)$-by-1 vector with the first b elements equal γ and the last element equal 0, and

$$\Sigma_2=\begin{pmatrix}\omega^2 I_{b\times b} & \mathbf{0}\\ \mathbf{0} & \infty\end{pmatrix}_{(b+1)\times(b+1)}$$

So the coefficient parameter γ_i and λ can be updated from a multivariate normal distribution.

$$
\begin{aligned}
&[(\gamma_1, \cdots, \gamma_b, \lambda)' | \mathbf{Y}_{mis}, \mathbf{Y}_{obs}, \mathbf{z}, \rho] \\
&\propto MVN\left(\left(X'\Sigma_1^{-1}X + \Sigma_2^{-1}\right)^{-1}\left(X'\Sigma_1^{-1}\mathbf{Z} + \Sigma_2^{-1}\boldsymbol{\mu}\right),\right. \\
&\left.\left(X'\Sigma_1^{-1}X + \Sigma_2^{-1}\right)^{-1}\right)
\end{aligned}
$$

where $\mathbf{Z}$ denotes the vector of $(z_{11}, \cdots, z_{1n_1}, \cdots, z_{bn_b})'$.

STEP 5: γ and ω^2 are drawn from

$$[\omega^2 | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\gamma}] \propto IG\left(a_3 + \frac{b-1}{2}, b_3 + \frac{1}{2}\sum_{i=1}^{b}(\gamma_i - \bar{\gamma})^2\right)$$

$$[\gamma | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\gamma}, \omega] \propto N\left(\bar{\gamma}, \frac{\omega^2}{b}\right)$$

where $\bar{\gamma} = \sum_{i=1}^{b}\gamma_i/b$

STEP 6: Then ρ is drawn from

$$[\rho | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\gamma}, \omega] \propto IG(a_4 + \frac{\sum_{i=1}^{b} n_i}{2}, b_4 + \frac{1}{2\tau^2}\sum_{i=1}^{b}\sum_{j=1}^{n_i}(z_{ij} - \gamma_i - \lambda y_{ij})^2)$$

Furthermore, we can standardize the outcome value y_{ij} in the $R|Y$ model, so that the parameter γ's and λ can be updated less correlated to each other. This extra step can help to improve the efficiency and is conducted before we run the MCMC chain.

2.6 Application to NAEP Data

In this section, we apply the above two proposed model structures, the pattern mixture model and the selection model, to the motivating data. In this dissertation, the primary interest is the 8th-grade math assessment scores of California students in year 2003. The research data involve two levels, the school and the student levels. The sample size is 6198 and there are 241 schools in total. The overall student response rate is 97.13%. By monitoring the MCMC chain behavior, we let the chain contain 61000 iterations and the first 1000 iterations are treated as the burn-in period.

2.6.1 Data Fitting Using Pattern Mixture Model

In this subsection, the observed data are fitted by the proposed pattern mixture model. We evaluate the MCMC chain behavior by a series of diagnosis plots of the parameters of interest: the trajectory plots, the auto-correlation plots and the empirical posterior density plots.

We show the diagnosis plots for the fitted pattern mixture model in Figure 2.6, Figure 2.7 and Figure 2.8. In Figure 2.6, we can see that all the parameters of interest converge to fixed values. Especially, α, the parameter which we use to test the missing mechanism, is above zero all the time, which suggests that the missingness is non-ignorable. Furthermore, the missing values tend to be the low score ones, which is equivalent to say that the low performance students are more likely to be missing. Figure 2.7 tells us that for all the parameters of interest, the auto-correlations die down eventually, which can guarantee that we can generate "independent draws" from the MCMC chain. Figure 2.8 shows the empirical posterior density functions of the parameters

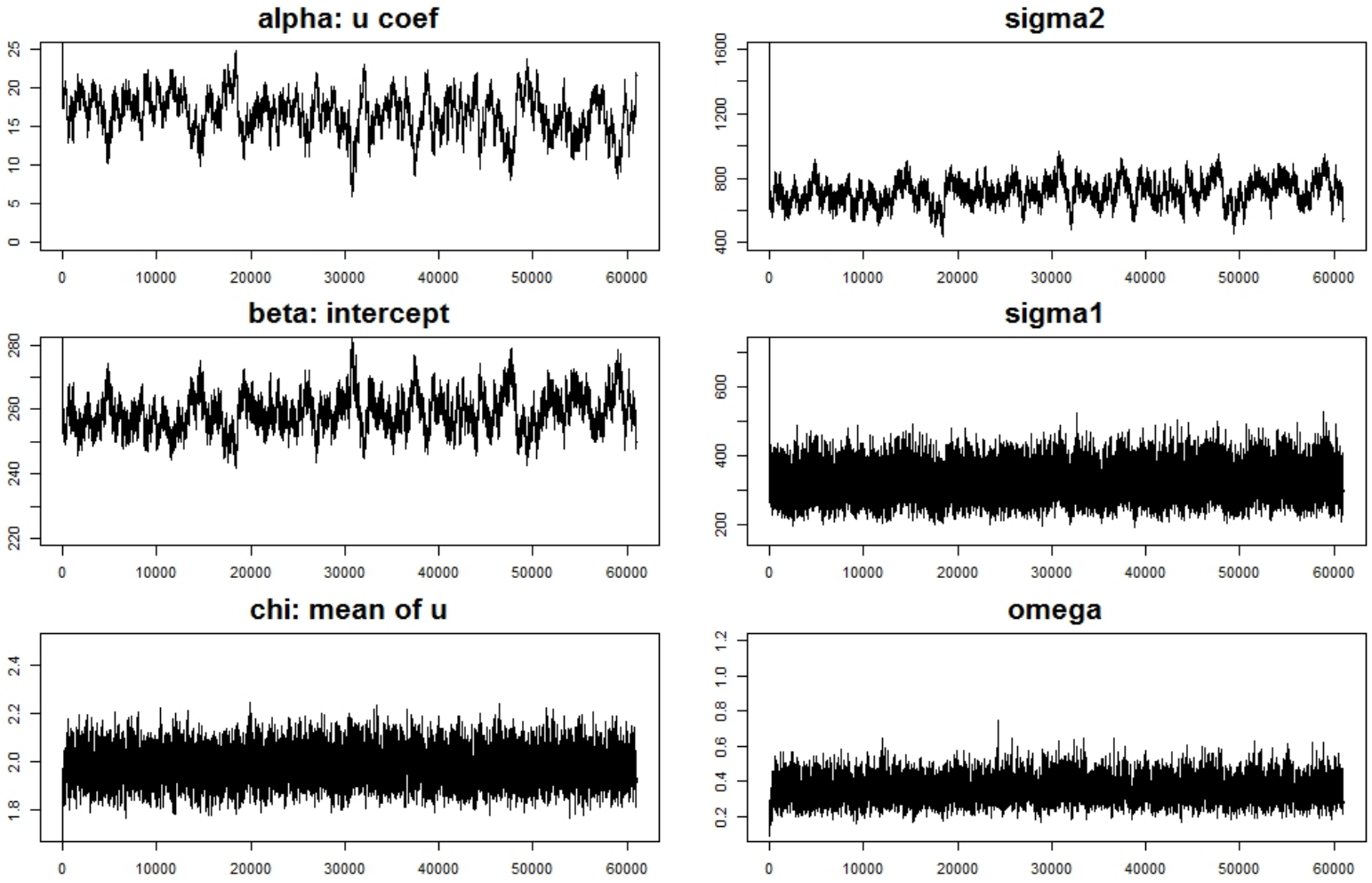

Figure 2.6: Trajectory plots for the parameter of interest in the pattern mixture model

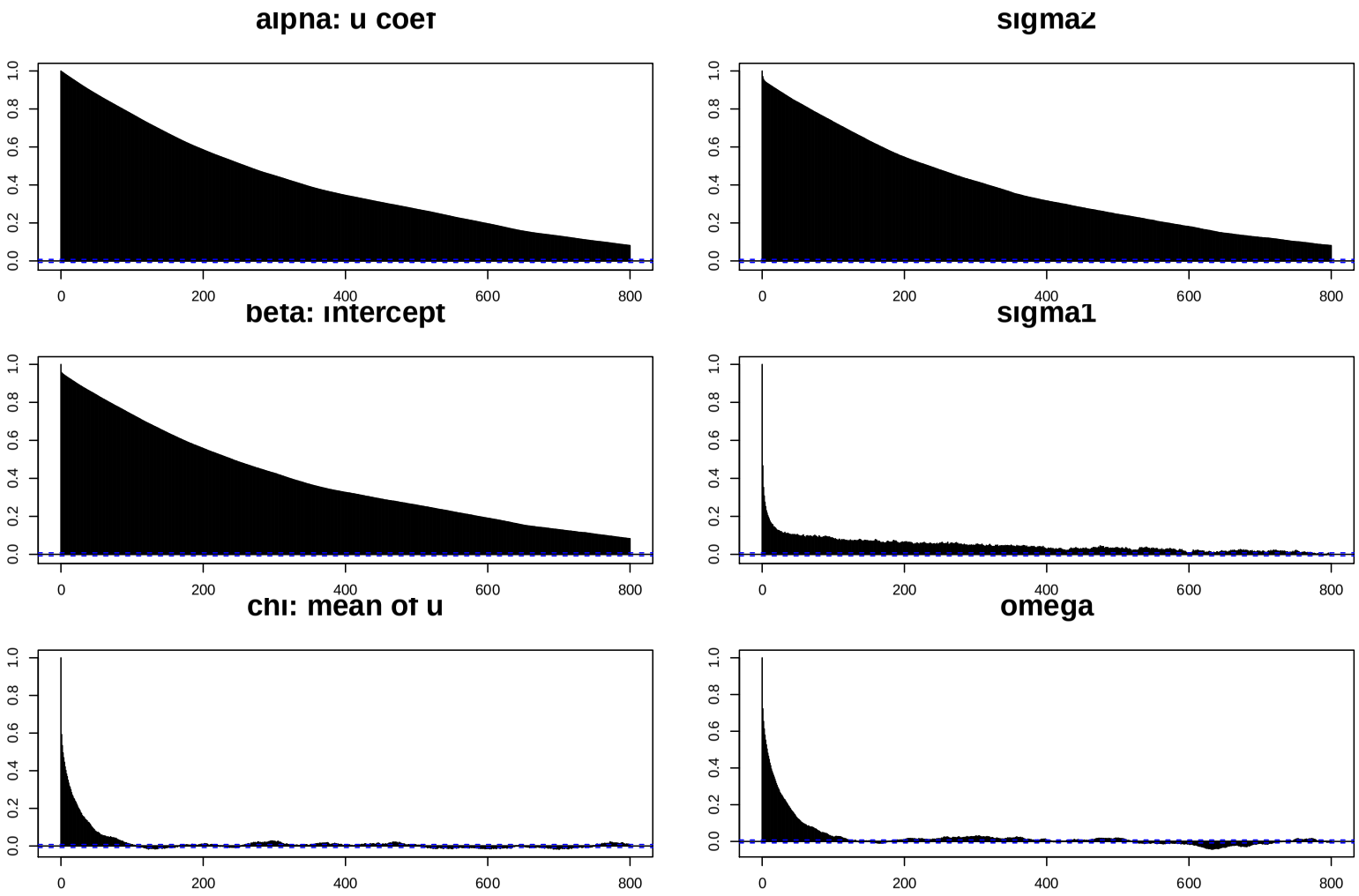

Figure 2.7: Autocorrelation plots for the parameter of interest in the pattern mixture model

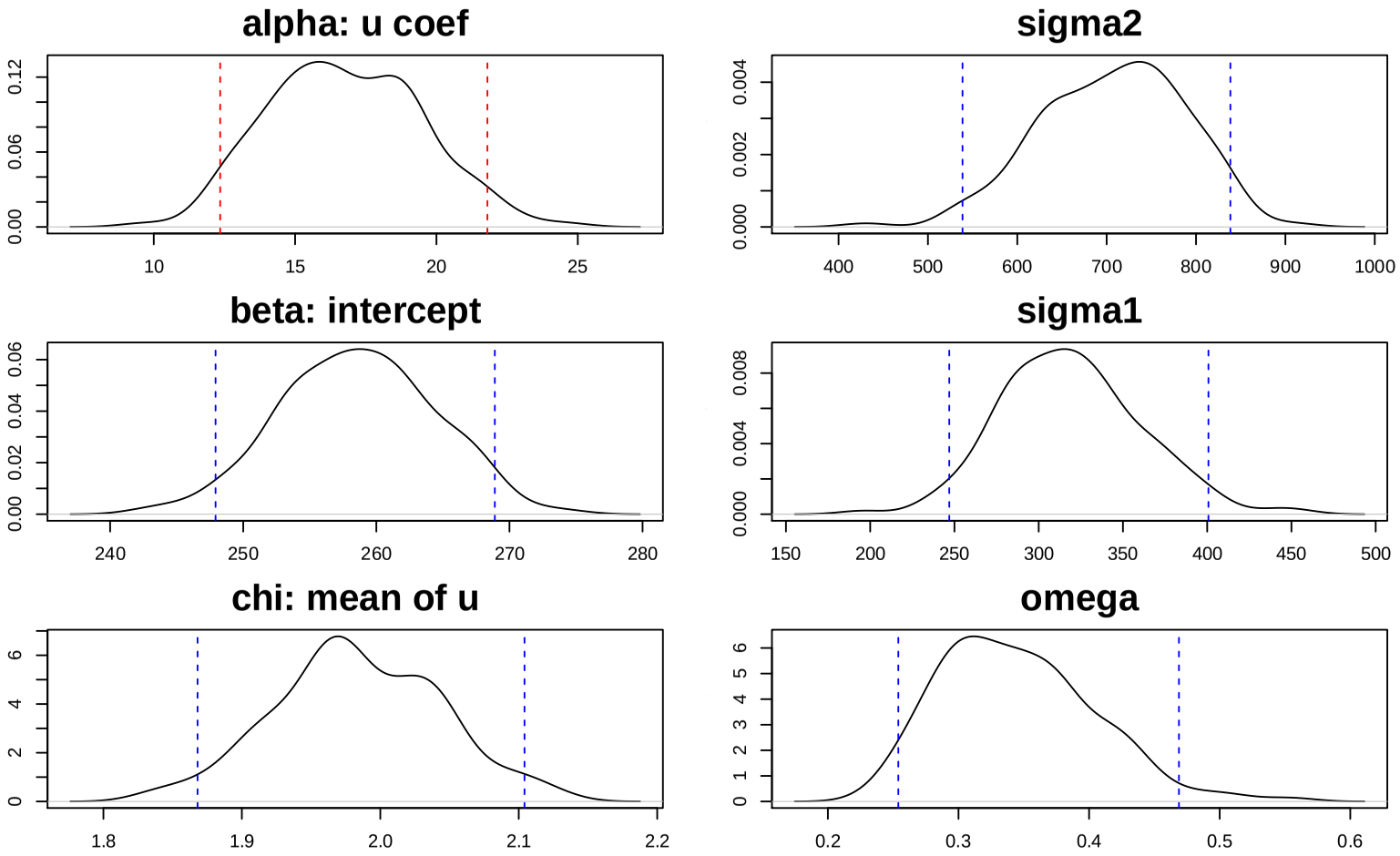

Figure 2.8: Empirical posterior distribution plots for the parameter of interest in the pattern mixture model

of interest. We can see that these densities do not violate any distributional properties and they have the same shape as the theoretical posterior densities which we derive in Section 2.4. For example, from the theoretical derivation, the posterior distribution of α is of the form of a normal distribution, while the empirical posterior distribution in Figure 2.8 also keeps the normal form. In each sub-figure, the two vertical dash lines also mark the 95% credible interval.

In Table 2.1, we show the summary information such as the mean, standard deviation, median, and the 95% credible interval for the independent draws we select from the pattern mixture model MCMC chain. The 95% credible interval for α, which we use to identify the missing mechanism, does not include 0 and is above 0 all the time. So the missing is not at random and the missing values have smaller values than that of the observed values. In other words, the low performance students are more likely to be missing. And with the help of the imputed missing values, we obtain that the average student performance score is 292.18 over all the selected students, observed or missing.

2.6.2 Data Fitting Using Selection Model

We also fit the motivating data by using the proposed selection model. And the MCMC chain is evaluated using the above series of plots. The trajectory plots (Figure 2.9) show that all the parameters of interest converge to a fixed value. Especially for the

Para	Mean	S.D.	2.50%	Median	97.50%
α	16.778	2.678	12.349	16.683	21.802
σ_2^2	706.794	80.552	538.685	713.512	838.410
β	258.610	5.765	247.921	258.524	268.901
σ_1^2	319.072	41.882	246.933	317.097	400.732
χ	1.984	0.060	1.868	1.979	2.104
ω^2	0.345	0.059	0.254	0.340	0.469

Table 2.1: Pattern Mixture Model Parameter Estimates for NAEP Data

parameter we used to test the missing mechanism, λ, the trajectory plot is above zero all the time, which gives us the consistent result as in the fitted pattern mixture model. The auto-correlation plots (Figure 2.10) and empirical posterior density plots (Figure 2.11) suggest that the auto-correlation dies down eventually and the subset draws are legitimate.

Our primary interest is the overall mean of the student performance score. Originally, the observed mean is 294.20, while the HT estimator, which is adjusted by response rate, is 293.62. For the Bayesian methods, we can "estimate" the overall mean based on the implemented data. The estimate is 292.18 by using the pattern mixture model while estimate is 292.44 by using the selection model. The model-based Bayesian methods provide us the estimates slightly less than the design based methods. From the above results, we can see that these two model not only identify whether the missingness is ignorable or non-ignorable, but also provide us a model-based estimator. The improvement is not quite big as what we expected it would be. The possible reason may be because of the small percentage of missingness, since our data only have less than 3% of missing values. In order to better show the performance of the model-based Bayesian methods, we artificially delete some of the observed data so that the modified

Para	Mean	S.D.	2.50%	Median	97.50%
β	292.067	1.301	289.871	292.210	294.755
σ_2^2	984.206	18.84	946.781	982.160	1019.586
λ	0.0103	0.0020	0.0067	0.0102	0.0141
σ_1^2	440.973	47.473	360.461	439.079	537.734
χ	-1.648	0.493	-2.447	-1.696	-0.743
ω^2	0.115	0.036	0.063	0.107	0.192
ρ	0.142	0.034	0.084	0.139	0.211

Table 2.2: Selection Model Parameter Estimates for NAEP Data

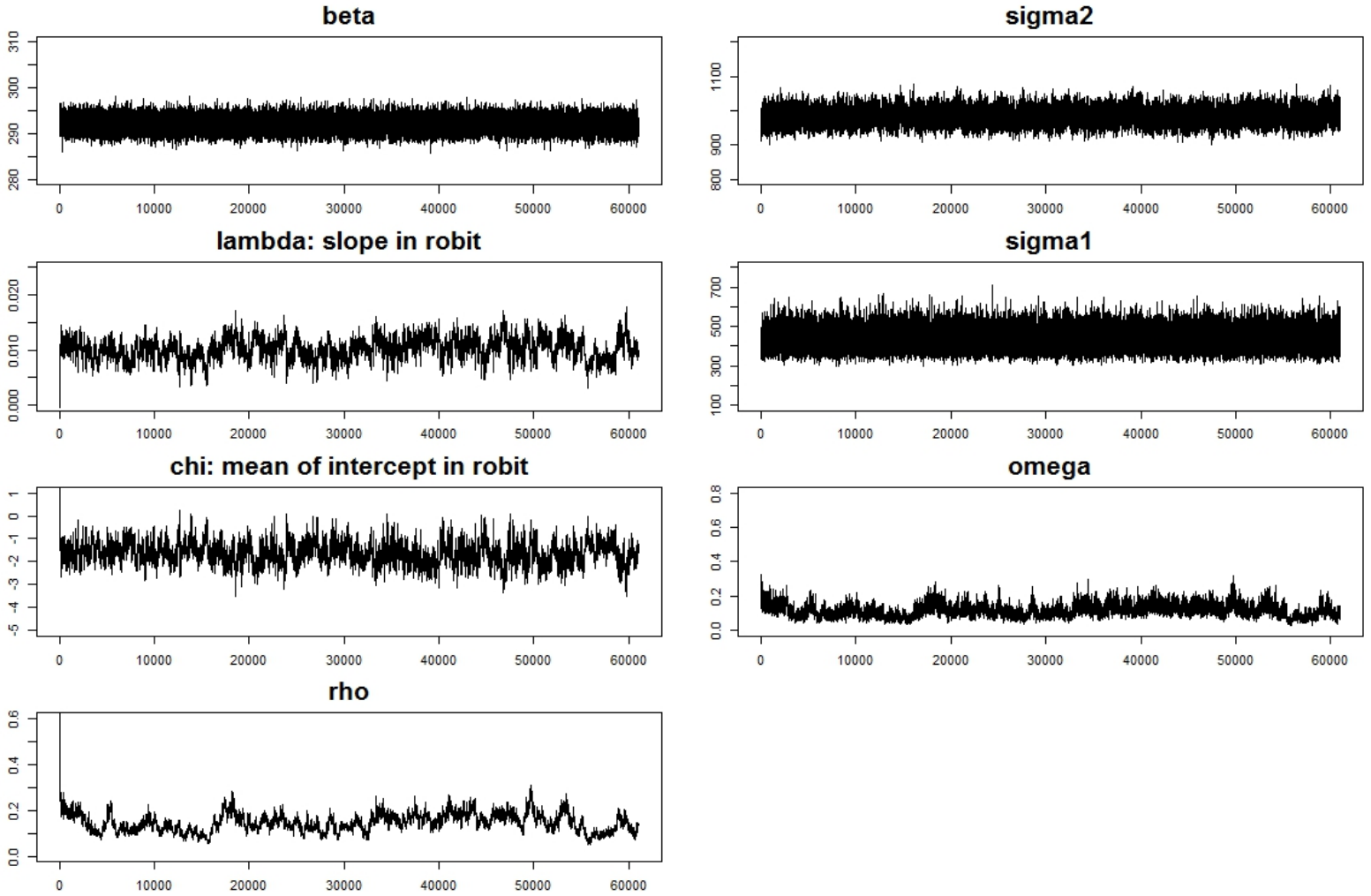

Figure 2.9: Trajectory plots for the parameter of interest in the selection model

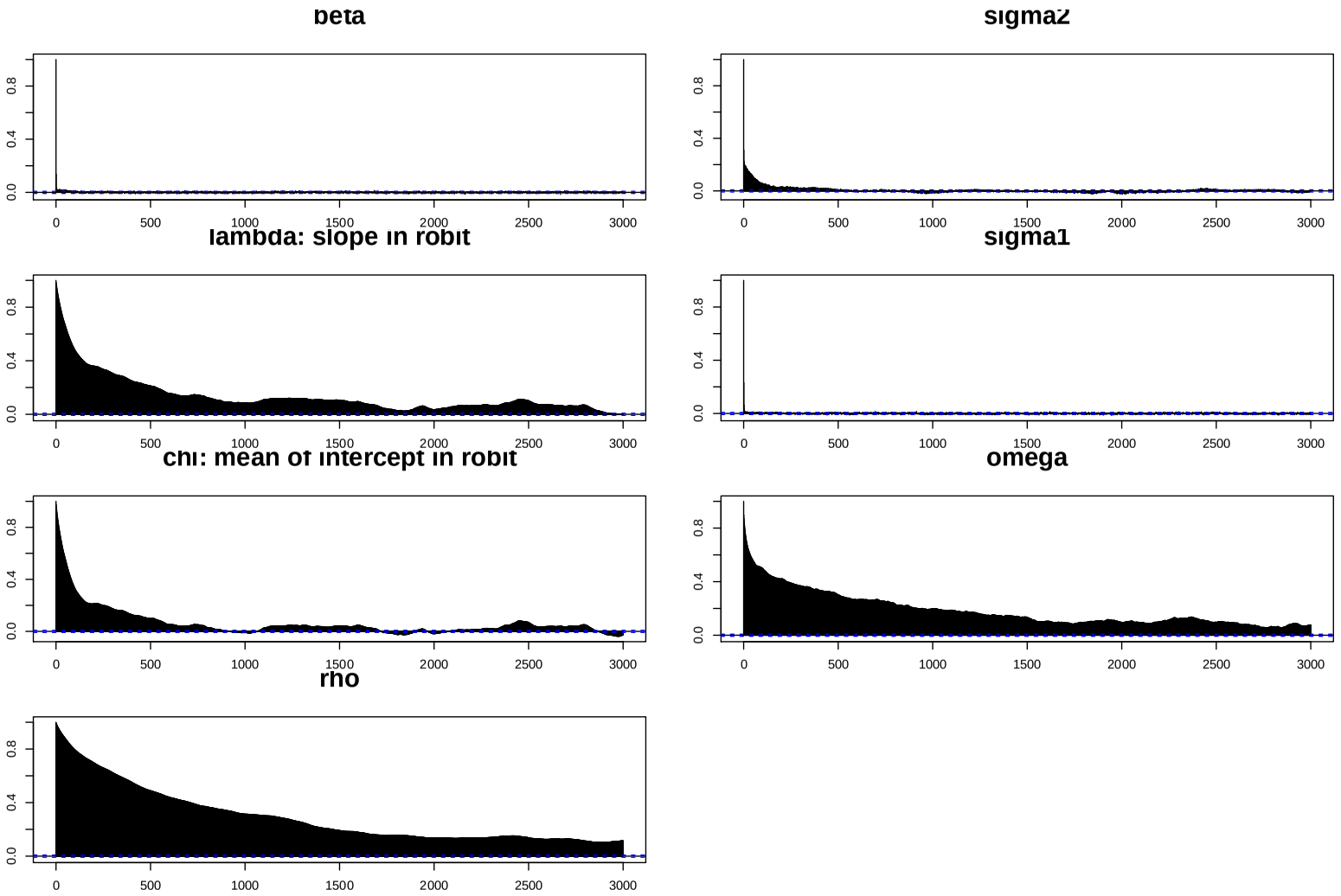

Figure 2.10: Autocorrelation plots for the parameter of interest in the selection model

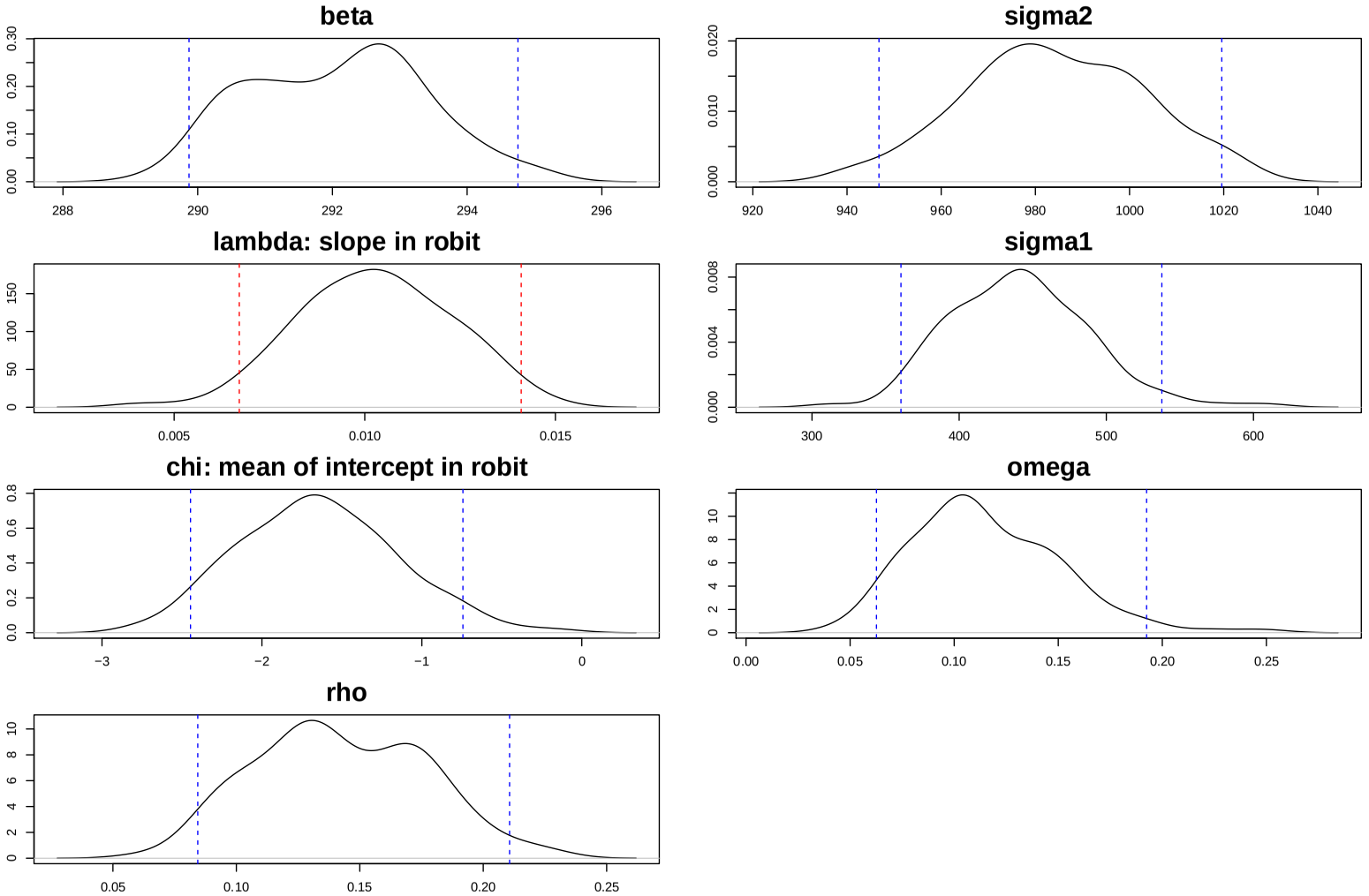

Figure 2.11: Empirical posterior distribution plots for the parameter of interest in the selection model

data set have 15% percent of the data are missing. For an originally observed datum, the probability that it will be deleted is proportional to its value. By deliberately deleting the low scores, the observed mean is 298.03 and the HT estimator is 296.32. Although the true overall mean is unknown, the observed mean and the HT estimator for the modified data definitely overestimate the truth since they are much higher than the results we get from the original data. Now, let us see the results of the proposed Bayesian methods.

The MCMC diagnosis plots are shown in Figure 2.12, Figure 2.13 and Figure 2.14. These plots verify that the MCMC chain for the modified data converges and we can generate independent draws from the MCMC chain. Compared with the MCMC chain for the original data, the auto-correlations are smaller for the modified data as we expected, since we have less information (the observed part) carried out from iteration to iteration. This makes the lag for the independent draws smaller than the original MCMC chain.

And in Table 2.3, we observe the following facts. First, the estimates of β, which is the overall mean adjusted by the missing latent variable, and estimates of α are different for the observed and the modified data. We believe that this is because the way we delete the outcomes may be different from the true missing mechanism.

Para	Mean	S.D.	2.50%	Median	97.50%
α	21.055	1.723	17.231	20.955	23.977
σ_2^2	636.613	48.020	553.458	638.420	738.957
β	267.199	2.805	262.484	267.361	273.672
σ_1^2	266.156	29.932	213.928	266.604	320.513
χ	1.107	0.036	1.038	1.106	1.179
ω^2	0.176	0.025	0.129	0.173	0.222

Table 2.3: Pattern Mixture Model Parameter Estimates for the modified NAEP Data

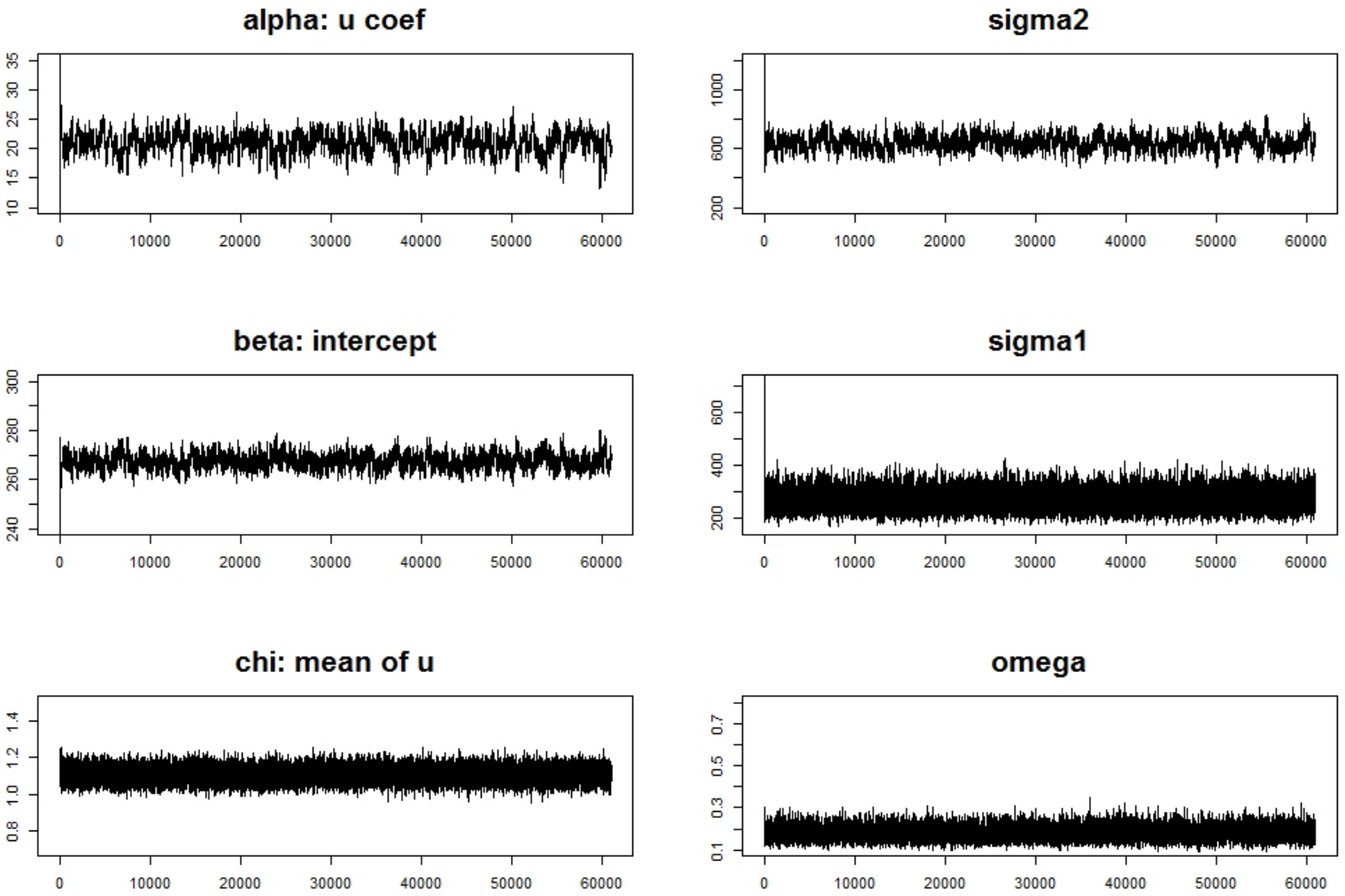

Figure 2.12: Trajectory plots for the parameter of interest in the pattern mixture model for the modified NAEP data

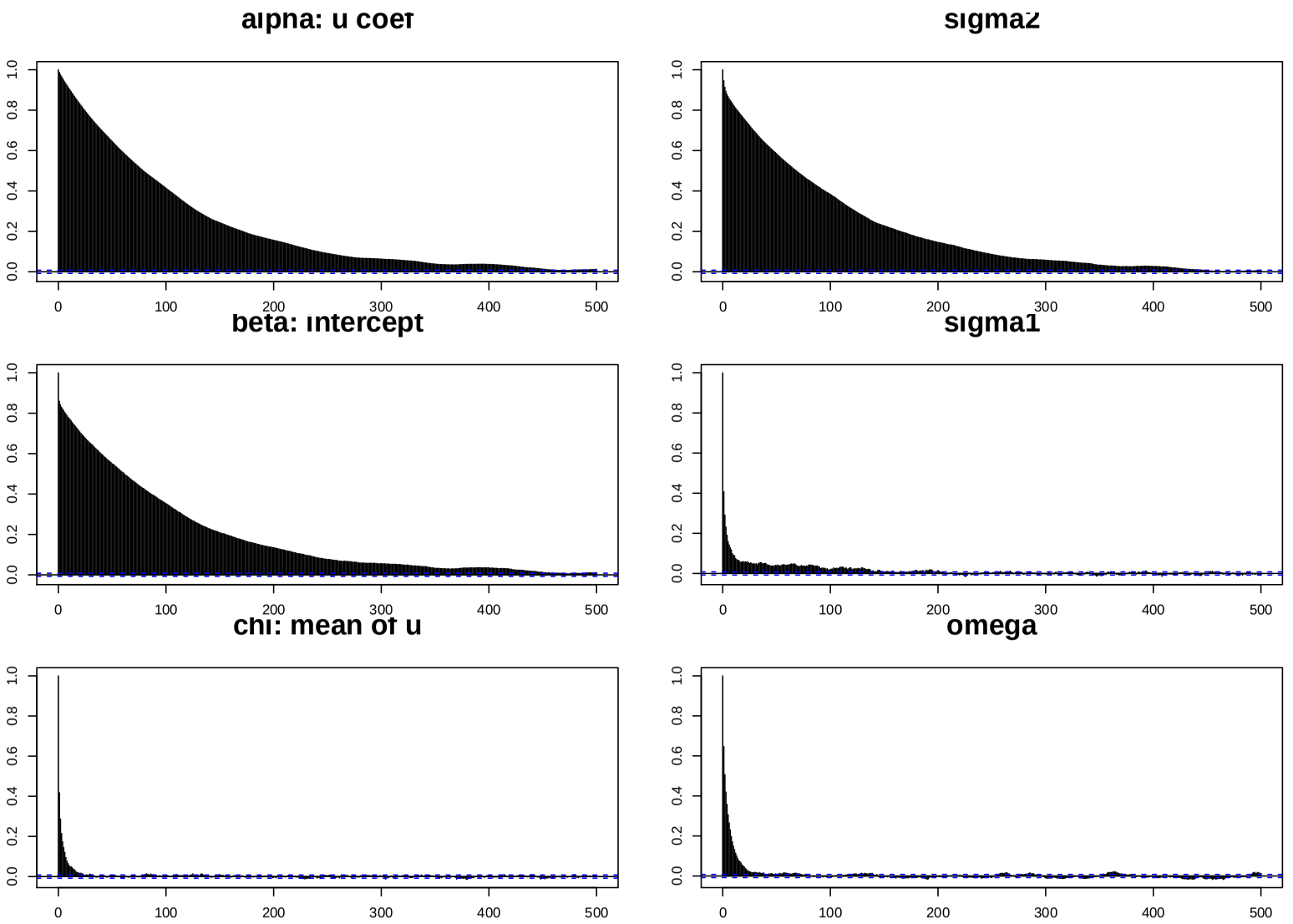

Figure 2.13: Autocorrelation plots for the parameter of interest in the pattern mixture model for the modified NAEP data

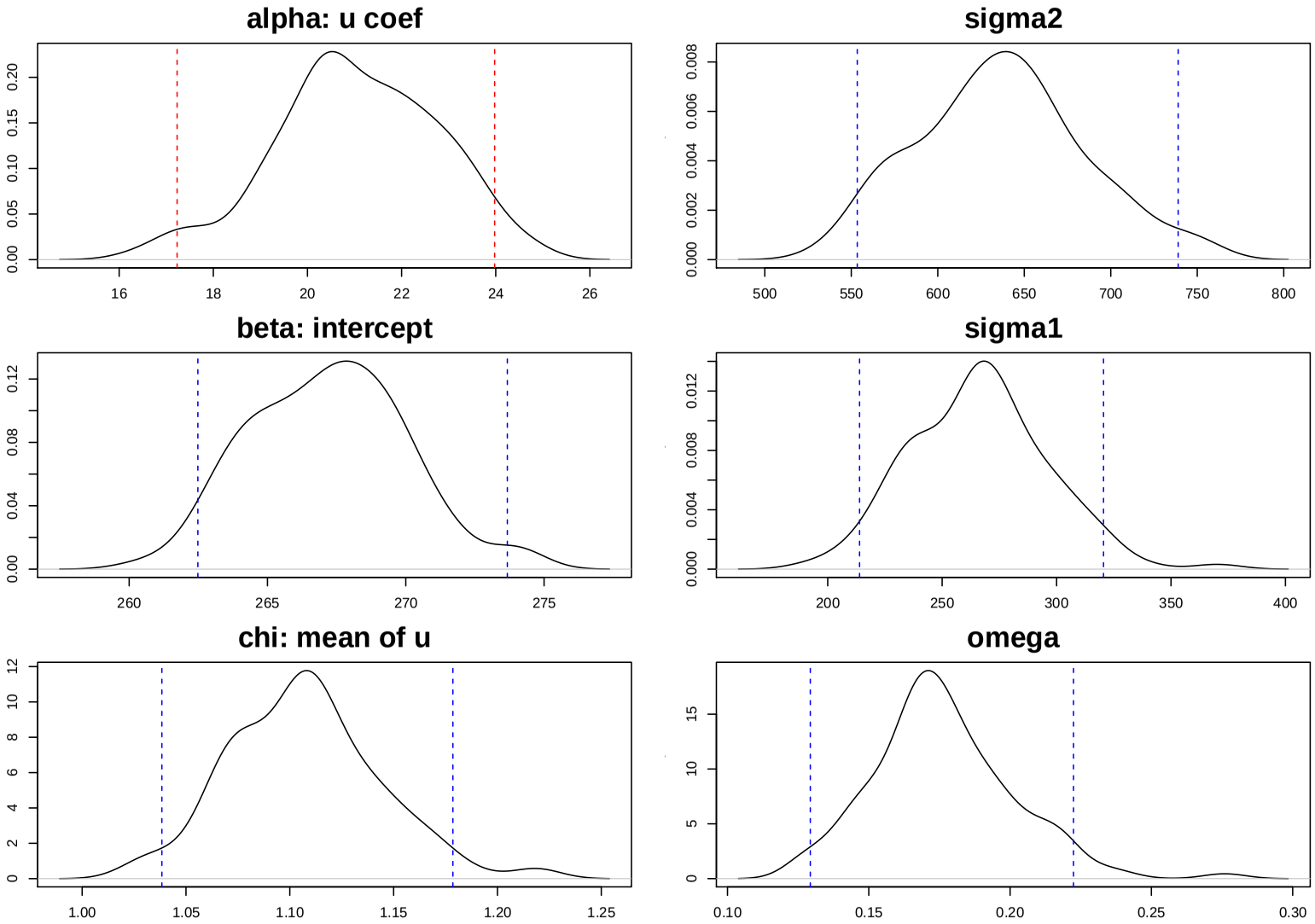

Figure 2.14: Empirical posterior distribution plots for the parameter of interest in the pattern mixture model for the modified NAEP data

For example, these two missing strategy have different slopes for the outcome values. Second, the estimates of σ_2^2 and σ_1^2 are smaller than these of the original data since we delete some extreme case. Third, the location parameters, such as α, β and χ have smaller standard deviation compared with these of the original MCMC chain. Forth, the estimates of χ and ω of the modified data are totally different from these of the original data, since these parameters describe the missing structure and the modified data have different missing structure than the original data. Last but the most importantly, the estimate for the overall mean of this MCMC chain for the modified data is 290.87. The difference of the pattern mixture model estimates between using the modified data and the original data is much more smaller than that for the observed means or the HT estimators.

In Figure 2.15, Figure 2.16 and Figure 2.17, we show the MCMC chain diagnosis plots for the selection model for the modified NAEP data. And in Table 2.4 shows the summary information of the empirical posterior distributions. We can draw similar conclusion as for the pattern mixture model of the modified data. The estimate for the overall mean of this MCMC chain for the modified data is 291.97.

Para	Mean	S.D.	2.50%	Median	97.50%
β	291.928	1.497	289.136	291.854	294.447
σ_2^2	1008.512	26.170	955.356	1007.894	1059.421
λ	0.0147	0.0016	0.0119	0.0148	0.0174
σ_1^2	463.343	50.088	388.219	459.159	565.107
χ	-3.274	0.380	-3.906	-3.244	-2.621
ω^2	0.037	0.011	0.020	0.035	0.065
ρ	0.251	0.045	0.170	0.250	0.334

Table 2.4: Selection Model Parameter Estimates for the modified NAEP Data

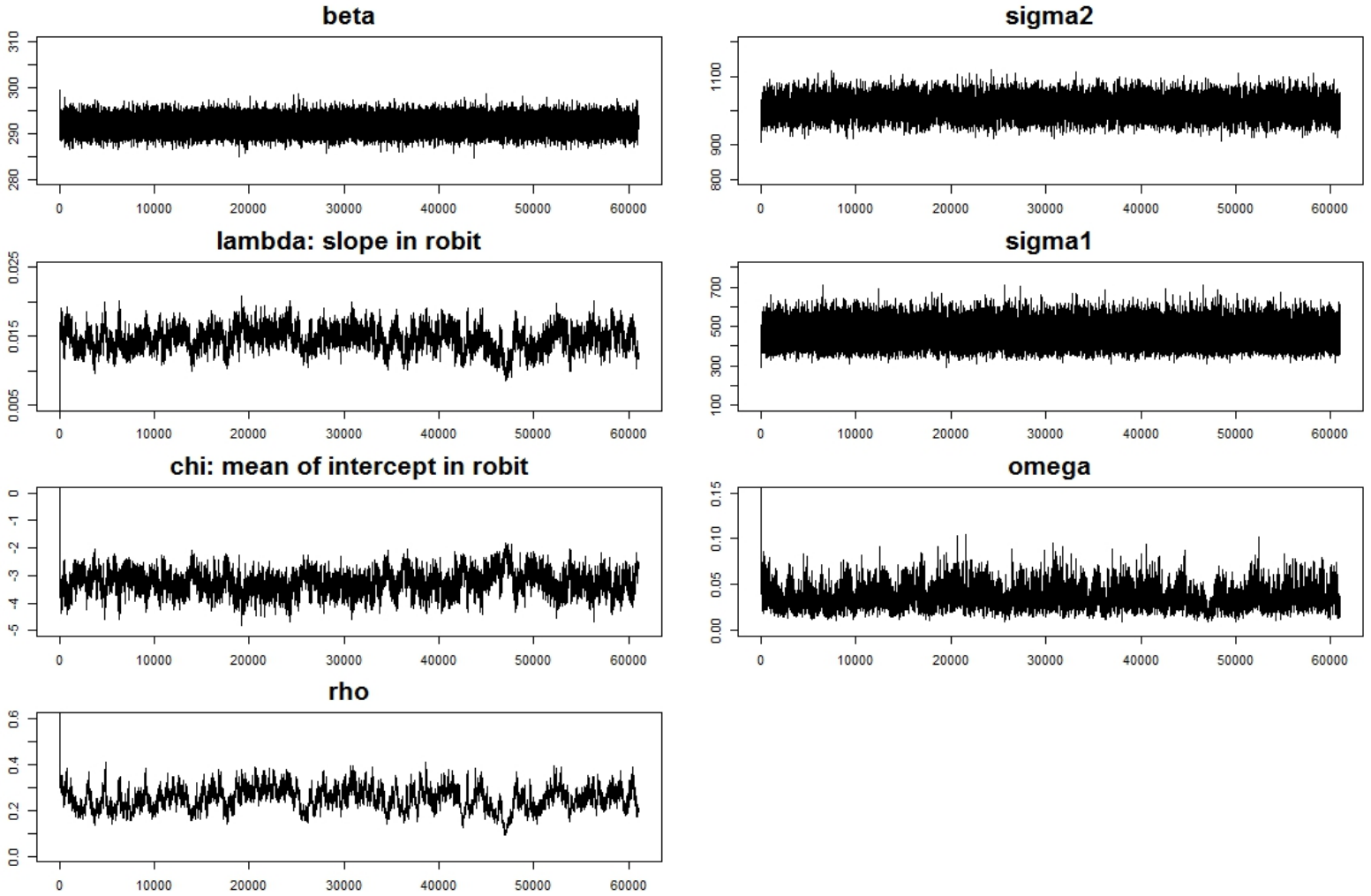

Figure 2.15: Trajectory plots for the parameter of interest in the selection model for the modified NAEP data

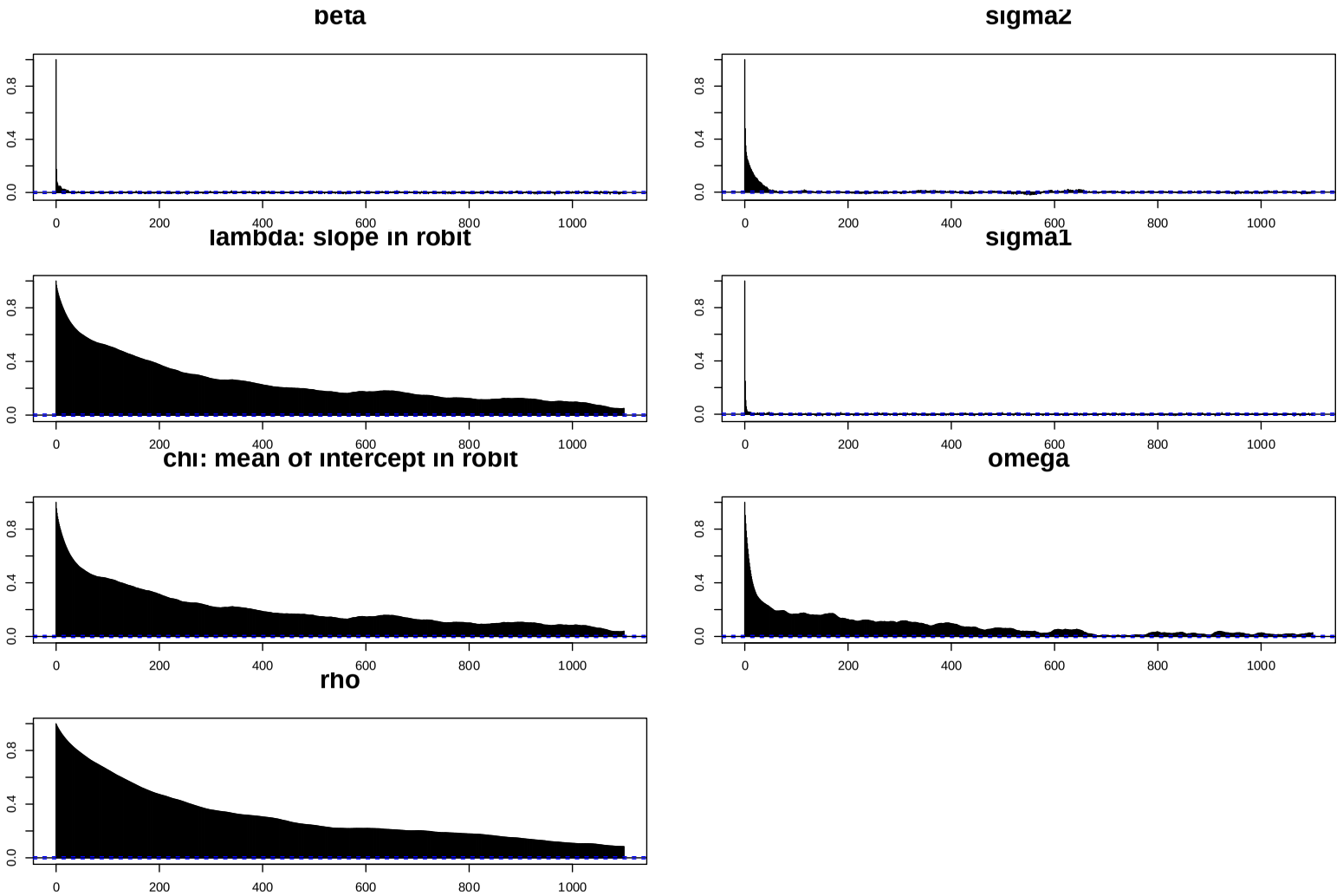

Figure 2.16: Autocorrelation plots for the parameter of interest in the selection model for the modified NAEP data

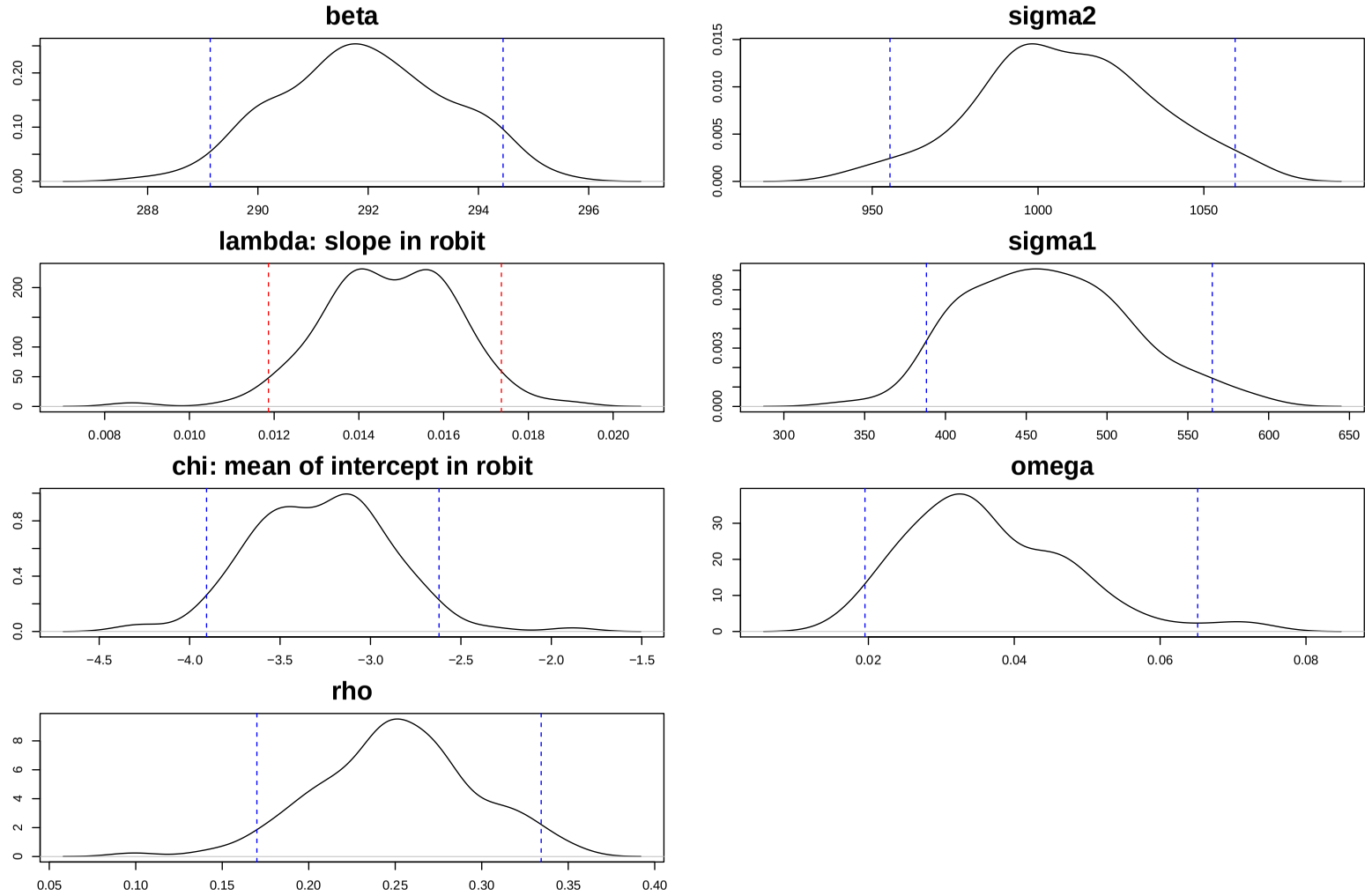

Figure 2.17: Empirical posterior distribution plots for the parameter of interest in the selection model for the modified NAEP data

In this section, we use three plots to evaluate the MCMC chain convergency. These plot are used to make sure that the MCMC chain converges and the subset draws can be treated as independent samples from the posterior distribution. As a summary, we list the results of the overall mean estimators for the original NAEP data and the modified NAEP data from four methods: the observed mean, the HT estimators, and the estimators using pattern mixture model and the selection model in Table 2.5. In order to account for the variability of the MCMC chain, we multiply impute the data $t = 10$ times. Schafer (1997) reviewed some methods for combining the results from the multiply imputed data. Here we utilize Rubin (1987)'s method to pool the results into a single number. The estimates are the average of the l estimates while the numbers in the parentheses are the corresponding pooled standard deviation. The pooled standard deviation aims to not only take the within-chain variability but also the uncertainty brought by the MCMC chain into consideration. It is calculated in the following way.

Suppose $\hat{O}_i$ is the estimate of student average score for MCMC chain i, and V_i is the corresponding standard error within the chain. Then the proposed estimate $\bar{O}$ is the mean of the individual estimates

$$\bar{O} = \frac{1}{t}\sum_{i=1}^{t}\hat{O}_i$$

The overall variance contains two parts, the within-chain variance and the between-chain

Method	NAEP data	Modified NAEP data
Observed Mean	294.20	298.03
HT Estimator	293.62	296.32
Pattern Mixture Model Estimator	292.14(0.25)	290.88(0.51)
Selection Model Estimator	292.34(0.22)	291.99(0.38)

Table 2.5: Overall Mean Estimate

variance. The within-chain variance $\bar{V}$ is of the form

$$\bar{V} = \frac{1}{t}\sum_{i=1}^{t} V_i$$

while the between-chain variance B is

$$B = \frac{1}{t-1}\sum_{i=1}^{t}\left(\hat{O}_i - \bar{O}\right)^2$$

So the total standard deviation v_{pool} is

$$v_{pool} = \sqrt{V_{pool}} = \sqrt{\bar{V} + \left(1 + \frac{1}{t}\right)B}$$

From the table, we can see that compared with the design-based estimators, the model-based Bayesian methods provide us a more robust results. Although the true value of the overall mean is unknown, we believe that the model-based Bayesian methods are less biased. We will use a simulation study to better evaluate the performance in Section 2.8.

2.7 Model Adequacy Checking

For a proposed Bayesian model, after verifying that the corresponding MCMC chain converges, we also want to see how well the model fits the data, which brings the issue of checking model adequacy. The assessment of the adequacy of the proposed hierarchical linear model is always a critical issue in statistics. There is a great amount of literatures in both the classical and Bayesian viewpoints, which try to meet this fundamental need after fitting the model. In Bayesian modeling, a good tool to assess

the model adequacy is to use the replicates of the observed data. If the proposed model is the true model for the observed data, then the replication data generated from the fitted model should have the same, or almost the same distribution as the observed data. Then the observed data can be treated as a random realization of the replication data. Or if we consider some function of the data $T(\mathbf{Y})$, we can also compare the observed value $T(\mathbf{Y}_{obs})$ with the replication values $T(\mathbf{Y}^{rep})$. In Robert (2007), the author reviewed two major approaches with different choices of the reference distributions used to generate the replicates.

1. Prior predictive approach (Box, 1980)

$$f_{y^{rep}}(y) = \int f(y|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$$

2. Posterior predictive approach (Rubin, 1984)

$$f_{y^{rep}|\mathbf{Y}_{obs}}(y|\mathbf{Y}) = \int f(y|\boldsymbol{\theta})f(\boldsymbol{\theta}|\mathbf{Y}_{obs})d\boldsymbol{\theta}$$

The prior predictive approach has the limitation that it is undefined under the scenario of improper prior distribution. This approach requires that all the parameters in the model have informative prior in order to make a reasonable conclusion. This requirement usually is very hard to justify in practice. So in the following section, we examined our proposed models by using the replications generated from the posterior distribution.

For our motivating data example, the model adequacy is assessed for both the pattern mixture model and the selection model. We first compare the performance of replicates with the true observed data. If the model fits the data well, we expect the posterior predictive distribution is a good approximation of the true distribution,

so that the replicates generated from the posterior predictive distribution have similar properties with the observed data. Then, we also evaluate the model by using cross validation analysis, which evaluates the residual values of the observed data in the Bayesian Statistics setting.

2.7.1 Model Adequacy Assessment by Using the Replicates of the Posterior Predictive Distribution

In this subsection, we compare the replicates with the observed data. The replicates are randomly and independently generated from the predictive posterior distribution $f(y_{ij}^{pre}|\mathbf{Y}_{obs}, \mathbf{R})$, $i = 1, \cdots, b$ and $j = 1, \cdots, r_i$:

$$\begin{aligned} f(y_{ij}^{pre}|\mathbf{Y}_{obs}, \mathbf{R}) &= \int_{\mathbf{\Omega}} f_{y^{pre}|\mathbf{\Omega}}(y_{ij}^{pre}|\mathbf{\Omega}, \mathbf{Y}_{obs}, \mathbf{R}) f_{\mathbf{\Omega}}(\mathbf{\Omega}|\mathbf{Y}_{obs}, \mathbf{R}) d\mathbf{\Omega} \\ &= \int_{\mathbf{\Omega}} f_{y^{pre}|\mathbf{\Omega}}(y_{ij}^{pre}|\mathbf{\Omega}) f_{\mathbf{\Omega}}(\mathbf{\Omega}|\mathbf{Y}_{obs}, \mathbf{R}) d\mathbf{\Omega} \end{aligned}$$

where y_{ij}^{pre} represents the replicates of the score for jth student in school i. Here the second equality holds due to the fact that y_{ij}^{pre} and $(\mathbf{Y}_{obs}, \mathbf{R})$ are independent when $\mathbf{\Omega}$ (which represents all the parameters except the latent variables $\mathbf{u}$ in the model structure) is given. In fact, $f_{y^{pre}|\mathbf{\Omega}}$ is the likelihood function. For the ignorable model and the non-ignorable selection model, $f_{y^{pre}|\mathbf{\Omega}}(y_{ij}^{pre}|\mathbf{\Omega}) \sim N\left(\beta_{0i}, \sigma_2^2\right)$. For the non-ignorable pattern mixture model,

$$f_{y^{pre}|\mathbf{\Omega}}(y_{ij}^{pre}|\mathbf{\Omega}) = \int_0^{\infty} f(y_{ij}^{pre}|\mathbf{\Omega}, u_{ij}) \pi(u_{ij}|\mathbf{\Omega}, \mathbf{Y}_{obs}, \mathbf{R}) du_{ij}$$

In the Bayesian MCMC chain setting, we carry out the MCMC chain on the full joint model $f(\mathbf{Y}_{obs}, \mathbf{R}|\mathbf{Y}_{mis}, \mathbf{\Omega})$. Then in order to get the independent samples, we save the updated values of $\mathbf{\Omega}$ for the subset of iterations with 1,000 burn-in period and at a lag of 500. The saved values are denoted as $\mathbf{\Omega}^{(q)}$, $q = 1, \cdots, Q$. Then the posterior predictive distribution is approximated by the empirical distribution of a set of random realizations: $y_{ij}^{(q)}$, $q = 1, \cdots, Q$, where $y_{ij}^{(q)}$ is generated from the likelihood function $f_{y^{pre}|\mathbf{\Omega}}(y_{ij}^{pre}|\mathbf{\Omega}^{(q)}, \mathbf{Y}_{obs}, \mathbf{R})$.

In Figure 2.18, we display five replicate samples generated from the fitted pattern mixture model for the observed data in five schools. In the figure, different rows contain the results for different schools. Take the first row as an example, the first column, which is isolated from the rest columns, is the histogram of the observed student scores in the first selected school. The other five columns display five replicated samples generated from the posterior predictive distribution for the observed data. If the right side histograms of the replicated samples have a common pattern, which is obviously different from the left side observed data, we consider the model as misfit.

In Figure 2.19, we summarize the posterior distribution by using four statistics: the sample minimum, maximum, mean and standard deviation values of the replicated samples. These statistics are selected to give us a quantified description of the shape of the posterior predictive distribution. Take the figure for minimum values (the top left sub-figure) as an example. The histogram is the frequency graph for the minimum values of 200 replicate samples. And the red vertical line marks the observed minimum value in the real data while the p-value is essentially the percentile of the observed minimum value among 200 minimum values of the replicate samples. The other three sub-figures function in the same way as the figure for minimum statistics.

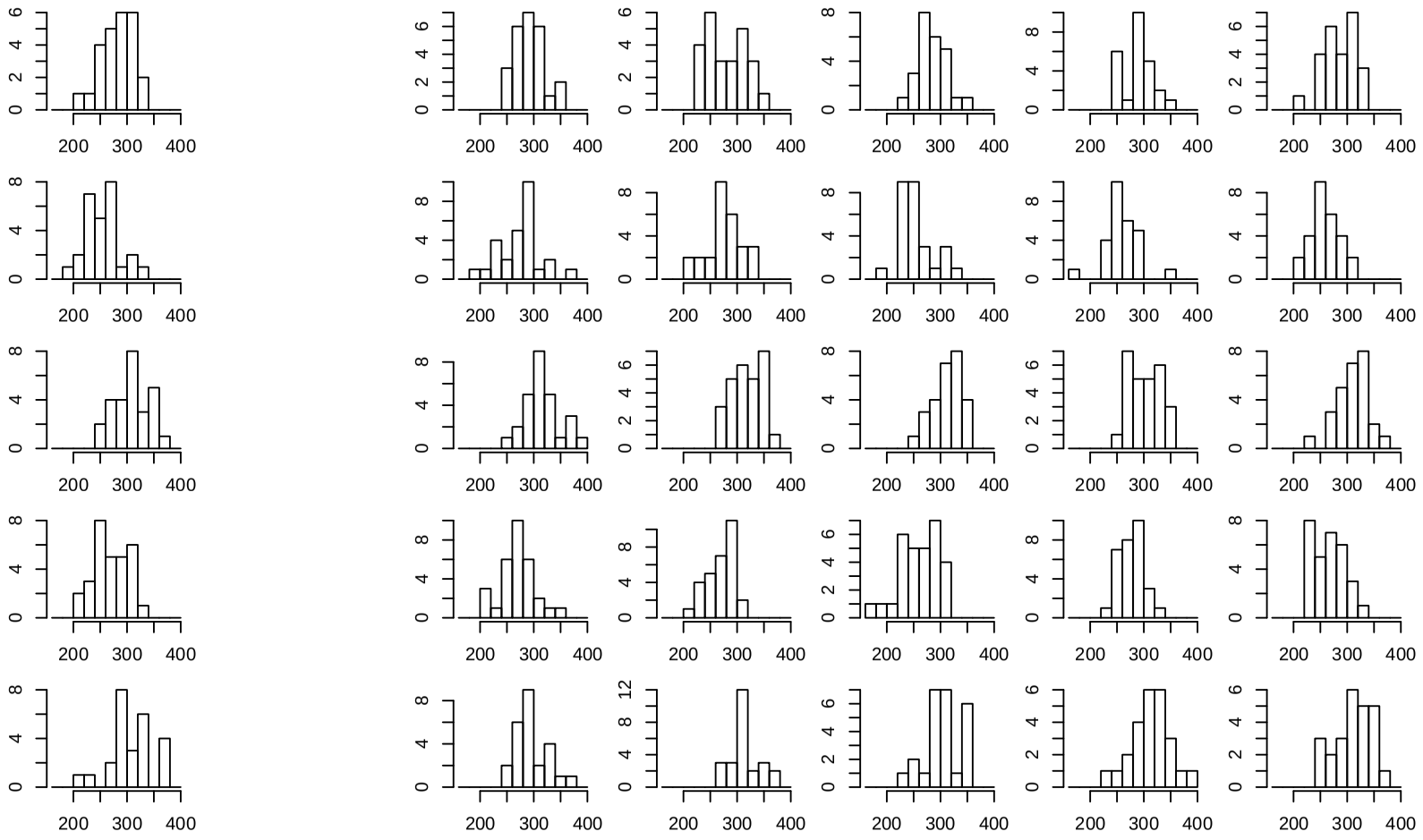

Figure 2.18: Observed and replicates data comparison for selection model: left column displays observed data for the first 5 schools, right columns display 5 replicated data sets y^{rep} from the fitted mixed effect model of $\mathbf{Y}_{obs}|\mathbf{R}$ in the pattern mixture model. No sign of misfit

Since there is no clear pattern of the replicate samples in Figure 2.18 and in Figure 2.19, all the p-values are greater than 0.05 (the significant level), we conclude that the pattern mixture model can be used to fit the real data.

Similarly, in Figure 2.20 and Figure 2.21, we display the five replicates samples for the observed data in five schools and the summary statistics of interest for the fitted selection model. Both of the figures show no indication of misfit for the fitted selection model. So fitting the observed data with the selection model is also acceptable.

2.7.2 Model Adequacy Assessment by Using Residual Plots Based on Cross Validation Analysis

In classical Statistics setting, residual analysis is a very powerful tool for the model adequacy assessment. For a data set $\mathbf{Y} = (y_1, \cdots, y_n)'$, suppose the predicted value calculated from the regression model is denoted as $\hat{y}_1, \cdots, \hat{y}_n$, then the residual of the ith observation is defined as the difference of the observed value and the predicted value:

$$e_i = y_i - \hat{y}_i$$

We usually assume that the response variables y_i, $i = 1, \cdots, n$, are independent and all have the same variance (homoscedasticity). For the general linear regression model, we have one more assumption that the y_i's, $i = 1, \cdots, n$, are normally distributed. In order to evaluate the performance of the proposed regression model, we have a series of residual plots. For example, the plot that the residuals are plotted against the predicted values, which is a scatter plot used to check the homoscedasticity and zero mean assumption, the residuals against independent variables, which is used to exam

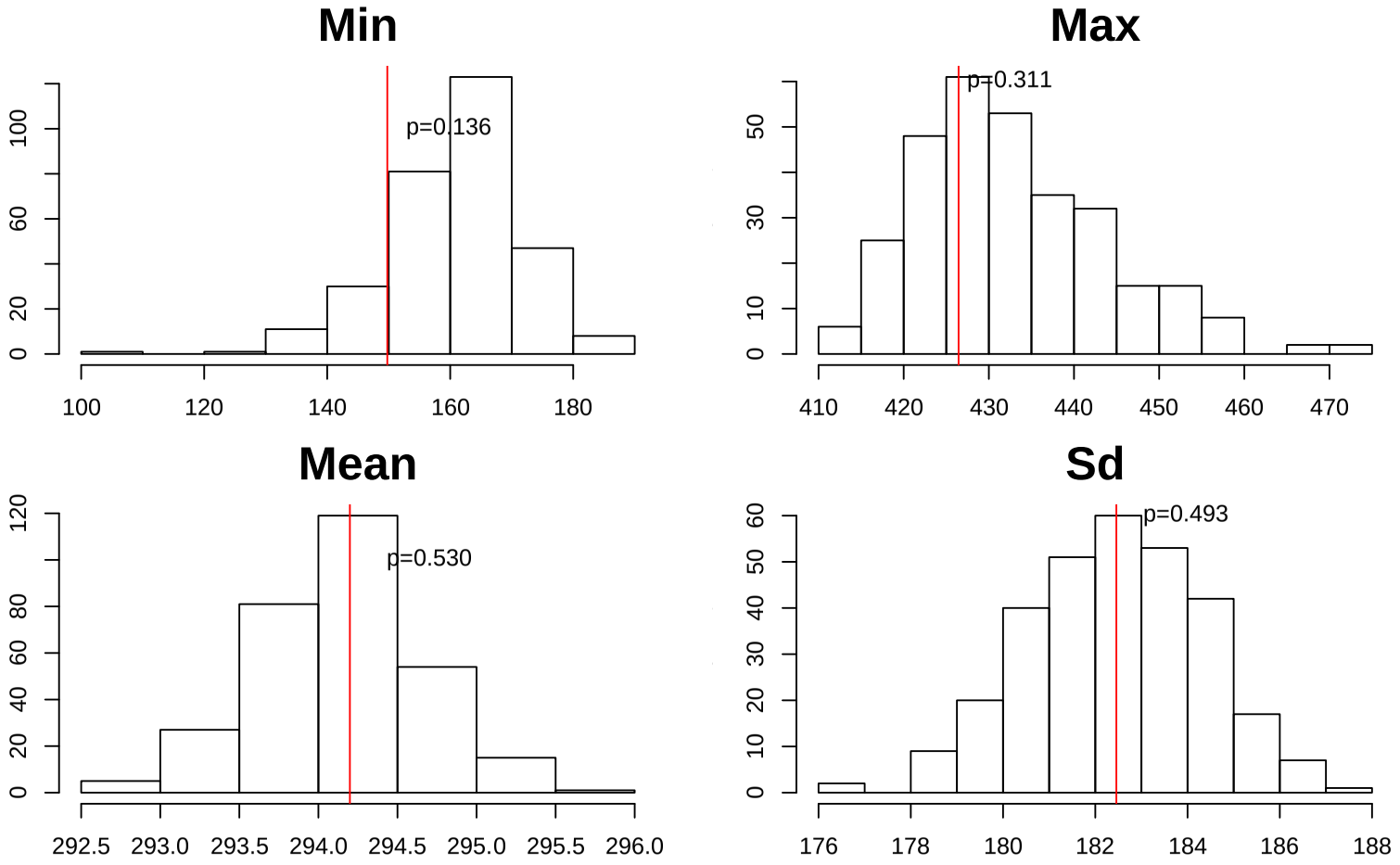

Figure 2.19: Summary statistics for the pattern mixture model: posterior predictive distribution, observed result, and p-value for each of four statistics

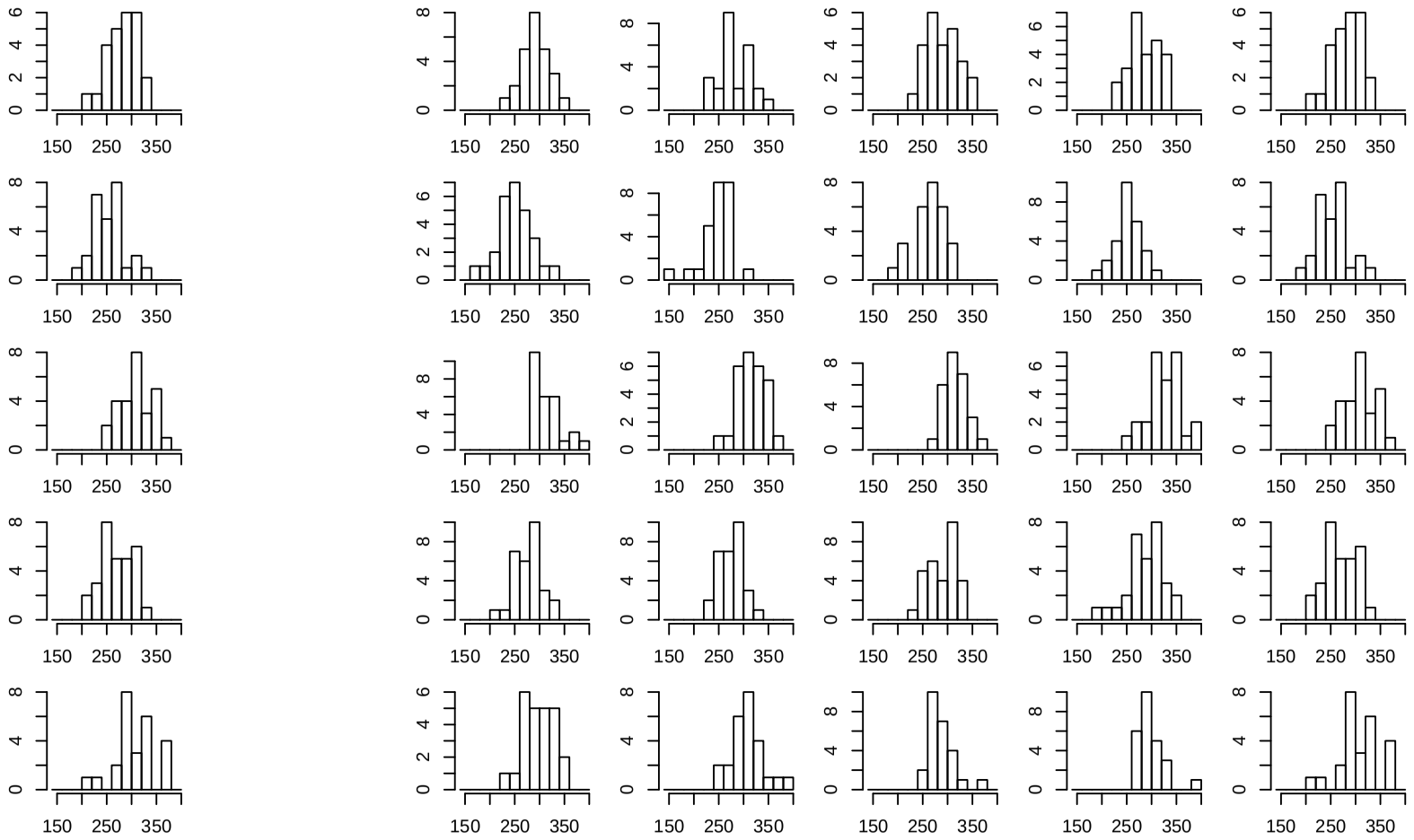

Figure 2.20: Observed and replicates data comparison for selection model: left column displays observed data for the first 5 schools, right columns display 5 replicated data sets y^{rep} from the fitted mixed effect model of $\mathbf{Y}_{obs}|\mathbf{R}$ in the selection model. No sign of misfit

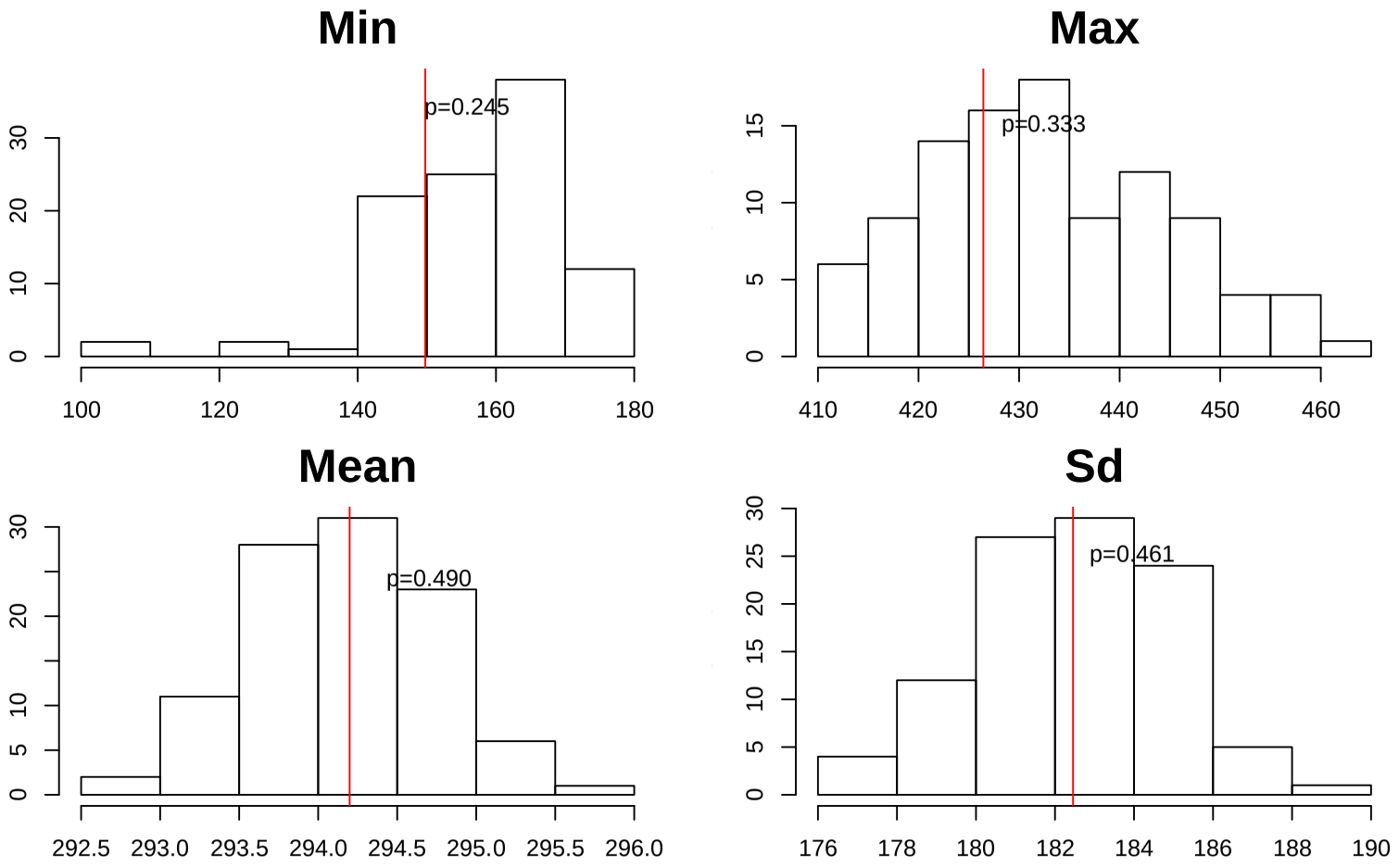

Figure 2.21: Summary statistics for the selection model: posterior predictive distribution, observed result, and p-value for each of four statistics

the randomness, and qq-plot (or normal probability plot), which is used to check the normality assumption.

In addition, people define the Studentized deleted residual as follows. For ith observation y_i, the deleted residual d_i was obtained after the model was re-fit without y_i. Then it was studentized as the division of d_i and its corresponding standard deviation. The studentization allows us to compare residuals across different data points even when the standard deviations of residuals vary greatly from points to points.

In Bayesian framework, the residuals are defined with the help of cross validation in order to analogize the classical Studentized deleted residual setting. Let $(\mathbf{y}_{(ij)}, \mathbf{r}_{(ij)})$ denote the observed outcome variables and the corresponding missing indicators with the absence of the $(ij)^{th}$ pair. We re-evaluate the model by using the data set $(\mathbf{y}_{(ij)}, \mathbf{r}_{(ij)})$, and compare the posterior predictive density function with the true value of this observation, y_{ij}. The standardized residual of y_{ij} is defined as

$$\text{RESID}_{ij} = \frac{y_{ij} - E[y_{ij}|\mathbf{y}_{(ij)}, \mathbf{r}_{(ij)}]}{SD[y_{ij}|\mathbf{y}_{(ij)}, \mathbf{r}_{(ij)}]}$$

where $E[y_{ij}|y_{(ij)}, r_{(ij)}]$ is the posterior mean and $SD[y_{ij}|y_{(ij)}, r_{(ij)}]$ is the posterior standard deviation.

If the proposed model fits the observed data well, we expect the standardized residual have mean 0 and standard deviation 1. Originally, in order to perform the cross validation, we should generate a MCMC chain for each of the data set $(\mathbf{y}_{(ij)}, \mathbf{r}_{(ij)})$, $i = 1, \cdots, b$ and $j = 1, \cdots, r_i$, which will lead to a total of $\sum_{i=1}^{b} r_i$ chains. Here, we utilize the importance sampling technique, which essentially makes the inference without generating a new MCMC chain. The available independent random draws from the original MCMC chain are re-weighted by introducing the weight function to correct

the potential bias by sampling from the "wrong" data. In this way, only one MCMC chain (the MCMC chain for the original observed data set) is needed.

To be more specific, let Ω denote all the parameters in the model and suppose for the $(\mathbf{y}_{(ij)}, \mathbf{r}_{(ij)})$ data set, we want to first generate K realizations of all the parameters, then for the kth realization, which is denoted as $\Omega_{(ij)}^{(k)}$ (from the target density function $p(\Omega_{(ij)}|\mathbf{y}_{(ij)}, \mathbf{r}_{(ij)})$), we can generate $y_{ij}^{(k)}$ from $f(y_{ij}|\Omega_{(ij)}^{(k)}, \mathbf{y}_{(ij)}, \mathbf{r}_{(ij)})$. Now all we need is to generate $\tilde{\Omega}^{(k)}$ (from the proposed density $p(\Omega|\mathbf{y}, \mathbf{r})$), and then by multiplying the weight function $w_{ij}^{(k)}$, we can use the generation of $\tilde{y}_{ij}^{(k)}$ from $f(y_{ij}|\tilde{\Omega}^{(k)}, \mathbf{y}, \mathbf{r})$. The weight function $w_{ij}^{(k)}$ is of the form

$$\begin{aligned} w_{ij}^{(k)} &= \frac{p\left(\tilde{\Omega}^{(k)}|\mathbf{y}_{(ij)}, \mathbf{r}_{(ij)}\right)}{p\left(\tilde{\Omega}^{(k)}|\mathbf{y}, \mathbf{r}\right)} \propto \frac{\pi\left(\tilde{\Omega}^{(k)}\right) \times p\left(\mathbf{y}_{(ij)}, \mathbf{r}_{(ij)}|\tilde{\Omega}^{(k)}\right)}{\pi\left(\tilde{\Omega}^{(k)}\right) \times p\left(\mathbf{y}, \mathbf{r}|\tilde{\Omega}^{(k)}\right)} \\ &= \left(p\left(y_{ij}, r_{ij}|\tilde{\Omega}^{(k)}\right)\right)^{-1} \end{aligned} \tag{2.4}$$

For pattern mixture model, the weight function is

$$\begin{aligned} w_{ij}^{(k)} &= \left(\int_{-\infty}^{+\infty} p\left(y_{ij}|\theta_{ij}^{(k)}, \tilde{\Omega}^{(k)}\right) \times p\left(r_{ij}|\theta_{ij}^{(k)}, \tilde{\Omega}^{(k)}\right) \times p(\theta_{ij}^{(k)}|\tilde{\Omega}^{(k)}) d\theta_{ij}^{(k)}\right)^{-1} \\ &= \left(\int_0^{\infty} \left(2\pi(\sigma_2^2)^{(k)}\right)^{-\frac{1}{2}} exp\left\{-\frac{1}{2(\sigma_2^2)^{(k)}}\left(y_{ij} - \beta_{0i}^{(k)} - \alpha\theta\right)^2\right\}\right)^{-1} \\ &\quad \cdot \left((2\pi)^{-\frac{1}{2}} exp\left\{-\frac{1}{2}\left(\theta - \chi_i^{(k)}\right)^2\right\} d\theta\right)^{-1} \\ &= \left(\phi\left(\frac{y_{ij} - \beta_{0i}^{(k)} - \alpha^{(k)}\chi_i^{(k)}}{\left((\sigma_2^2)^{(k)} + (\alpha^{(k)})^2\right)^{\frac{1}{2}}}\right) \cdot \Phi\left(\frac{\frac{\alpha^{(k)}}{(\sigma_2^2)^{(k)}}\left(y_{ij} - \beta_{0i}^{(k)}\right) + \chi_i^{(k)}}{\left(\frac{\left(\alpha^{(k)}\right)^2}{(\sigma_2^2)^{(k)}} + 1\right)^{\frac{1}{2}}}\right)\right)^{-1} \end{aligned}$$

where $\phi(\cdot)$ denotes the standard normal density function and $\Phi(\cdot)$ denotes the standard normal cumulative distribution function.

For selection model, the weight function is

$$
\begin{aligned}
w_{ij}^{(k)} &= \left(p\left(r_{ij} | y_{ij}, \tilde{\Omega}^{(k)} \right) \times p\left(y_{ij} | \tilde{\Omega}^{(k)} \right) \right)^{-1} \\
&= \left(\phi \left(\frac{y_{ij} - \beta_{0i}^{(k)}}{\left((\sigma_2^2)^{(k)} \right)^{\frac{1}{2}}} \right) \cdot \Phi \left(\frac{\gamma_i^{(k)} + \lambda^{(k)} y_{ij}}{\left(\rho^{(k)} (\tau^2)^{(k)} \right)^{\frac{1}{2}}} \right) \right)^{-1}
\end{aligned}
$$

With the appropriate weight function the posterior mean and standard deviation are

$$
\begin{aligned}
\mathrm{E}[y_{ij} | y_{(ij)}, r_{(ij)}] &= \frac{\sum_{k=1}^{K} w_{ij}^{(k)} \tilde{y}_{ij}^{(k)}}{\sum_{k=1}^{K} w_{ij}^{(k)}} \\
\mathrm{SD}[y_{ij} | y_{(ij)}, r_{(ij)}] &= \sqrt{\frac{\sum_{k=1}^{K} w_{ij}^{(k)} \left(\tilde{y}_{ij}^{(k)} - E[y_{ij} | y_{(ij)}, r_{(ij)}] \right)^2}{\sum_{k=1}^{K} w_{ij}^{(k)}}}
\end{aligned}
$$

When we calculate the predictive posterior mean and standard deviation, by Rao-Blackwell's Theorem, the following ones, which first calculate the conditional expectation of the predictive value given all the parameters, will have smaller variance. For pattern mixture model we have

$$
\begin{aligned}
E[y_{ij} | y_{(ij)}, r_{(ij)}] &= E_\Omega E_{\theta_{ij}|\Omega} E_{y_{ij}|\theta_{ij},\Omega}[y_{ij} | \Omega, y_{(ij)}, r_{(ij)}] = E_\Omega E_{\theta_{ij}|\Omega} \left(\beta_{0i}^{(k)} + \alpha^{(k)} \theta_{ij}^{(k)} \right) \\
&= E_\Omega \left(\beta_{0i}^{(k)} + \alpha^{(k)} \left(\mu_{ij}^{(k)} + \frac{\phi \left(\mu_{ij}^{(k)} \sqrt{\frac{\left(\alpha^{(k)} \right)^2}{\left(\sigma_2^2 \right)^{(k)}} + 1} \right)}{\Phi \left(\mu_{ij}^{(k)} \sqrt{\frac{\left(\alpha^{(k)} \right)^2}{\left(\sigma_2^2 \right)^{(k)}} + 1} \right)} \left(\frac{\left(\alpha^{(k)} \right)^2}{\left(\sigma_2^2 \right)^{(k)}} + 1 \right)^{-\frac{1}{2}} \right) \right) \\
&= \frac{\sum_{k=1}^{K} w_{ij}^{(k)} \left(\beta_{0i}^{(k)} + \alpha^{(k)} \left(\mu_{ij}^{(k)} + \frac{\phi \left(\mu_{ij}^{(k)} \sqrt{\frac{\left(\alpha^{(k)} \right)^2}{\left(\sigma_2^2 \right)^{(k)}} + 1} \right)}{\Phi \left(\mu_{ij}^{(k)} \sqrt{\frac{\left(\alpha^{(k)} \right)^2}{\left(\sigma_2^2 \right)^{(k)}} + 1} \right)} \left(\frac{\left(\alpha^{(k)} \right)^2}{\left(\sigma_2^2 \right)^{(k)}} + 1 \right)^{-\frac{1}{2}} \right) \right)}{\sum_{k=1}^{K} w_{ij}^{(k)}}
\end{aligned}
$$

$$
\begin{aligned}
&Var[y_{ij}|y_{(ij)}, r_{(ij)}] = E_\Omega E_{\theta_{ij}|\Omega} E_{y_{ij}|\theta_{ij},\Omega}[y_{ij}^2|\Omega, \theta_{ij}, y_{(ij)}, r_{(ij)}] - \left(E[y_{ij}|y_{(ij)}, r_{(ij)}]\right)^2 \\
=&E_\Omega E_{\theta_{ij}|\Omega}\left[\left(\beta_{0i}^{(k)} + \alpha^{(k)}\theta_{ij}^{(k)}\right)^2 + \left(\sigma_2^2\right)^{(k)}\right] - \left(E[y_{ij}|y_{(ij)}, r_{(ij)}]\right)^2 \\
=&E_\Omega\left[\left(\beta_{0i}^{(k)}\right)^2 + 2\beta_{0i}^{(k)}\alpha^{(k)}\left(\mu_{ij}^{(k)} + \frac{\phi\left(\mu_{ij}^{(k)}\sqrt{\frac{\left(\alpha^{(k)}\right)^2}{\left(\sigma_2^2\right)^{(k)}}+1}\right)}{\Phi\left(\mu_{ij}^{(k)}\sqrt{\frac{\left(\alpha^{(k)}\right)^2}{\left(\sigma_2^2\right)^{(k)}}+1}\right)}\left(\frac{\left(\alpha^{(k)}\right)^2}{\left(\sigma_2^2\right)^{(k)}}+1\right)^{-\frac{1}{2}}\right)\right. \\
&+\left(\alpha^{(k)}\right)^2\left(\left(\mu_{ij}^{(k)}\right)^2 + \mu_{ij}^{(k)}\frac{\phi\left(\mu_{ij}^{(k)}\sqrt{\frac{\left(\alpha^{(k)}\right)^2}{\left(\sigma_2^2\right)^{(k)}}+1}\right)}{\Phi\left(\mu_{ij}^{(k)}\sqrt{\frac{\left(\alpha^{(k)}\right)^2}{\left(\sigma_2^2\right)^{(k)}}+1}\right)}\left(\frac{\left(\alpha^{(k)}\right)^2}{\left(\sigma_2^2\right)^{(k)}}+1\right)^{-\frac{1}{2}}\right. \\
&\left.\left.+\left(\frac{\left(\alpha^{(k)}\right)^2}{\left(\sigma_2^2\right)^{(k)}}+1\right)^{-1} + \left(\sigma_2^2\right)^{(k)}\right)\right] \\
&-\left(E[y_{ij}|y_{(ij)}, r_{(ij)}]\right)^2
\end{aligned}
$$

where $\mu_{ij}^{(k)} = \chi_i^{(k)} + \frac{\alpha^{(k)}\left(y_{ij} - \beta_{0i}^{(k)} - \alpha^{(k)}\chi_i^{(k)}\right)}{\left(\alpha^{(k)}\right)^2 + \left(\sigma_2^2\right)^{(k)}}$, is the mean parameter in the posterior truncated normal distribution for θ_{ij}.

Similarly for selection model, we have

$$
\begin{aligned}
\mathrm{E}[y_{ij}|y_{(ij)}, r_{(ij)}] &= E_\Omega E[y_{ij}|\Omega, y_{(ij)}, r_{(ij)}] = \frac{\sum_{k=1}^K w_{ij}^{(k)}\beta_{0i}^{(k)}}{\sum_{k=1}^K w_{ij}^{(k)}} \\
\mathrm{SD}[y_{ij}|y_{(ij)}, r_{(ij)}] &= \sqrt{E_\Omega E[y_{ij}^2|\Omega, y_{(ij)}, r_{(ij)}] - \left(E[y_{ij}|y_{(ij)}, r_{(ij)}]\right)^2} \\
&= \sqrt{\frac{\sum_{k=1}^K w_{ij}^{(k)}\left[\left(\beta_{0i}^{(k)}\right)^2 + \left(\sigma_2^2\right)^{(k)}\right]}{\sum_{k=1}^K w_{ij}^{(k)}} - \left(E[y_{ij}|y_{(ij)}, r_{(ij)}]\right)^2}
\end{aligned}
$$

Proof of the last equality in the derivation of the weight function for pattern mixture model:

$$
\begin{aligned}
&\int_0^\infty \left(2\pi(\sigma_2^2)^{(k)}\right)^{-\frac{1}{2}} exp\left\{-\frac{1}{2(\sigma_2^2)^{(k)}}\left(y_{ij}-\beta_{0i}^{(k)}-\alpha z\right)^2\right\} \\
&\cdot (2\pi)^{-\frac{1}{2}} exp\left\{-\frac{1}{2}\left(z-\chi_i^{(k)}\right)^2\right\} dz \\
=&\left(2\pi(\sigma_2^2)^{(k)}\right)^{-\frac{1}{2}}(2\pi)^{-\frac{1}{2}}\int_0^\infty exp\left\{-\frac{1}{2}\left[\left(\frac{\left(\alpha^{(k)}\right)^2}{\left(\sigma_2^2\right)^{(k)}}+1\right)z^2\right.\right. \\
&\left.\left.-2\left(\frac{\alpha^{(k)}}{(\sigma_2^2)^{(k)}}\left(y_{ij}-\beta_{0i}^{(k)}\right)+\chi_i^{(k)}\right)z+\frac{1}{(\sigma_2^2)^{(k)}}\left(y_{ij}-\beta_{0i}^{(k)}\right)^2+\left(\chi_i^{(k)}\right)^2\right]\right\} \\
=&\left(2\pi(\sigma_2^2)^{(k)}\right)^{-\frac{1}{2}}\left(\frac{\left(\alpha^{(k)}\right)^2}{\left(\sigma_2^2\right)^{(k)}}+1\right)^{\frac{-1}{2}} exp\left\{-\frac{1}{2}\left[\frac{1}{(\sigma_2^2)^{(k)}}\left(y_{ij}-\beta_{0i}^{(k)}\right)^2+\left(\chi_i^{(k)}\right)^2\right.\right. \\
&\left.\left.-\frac{\left(\frac{\alpha^{(k)}}{(\sigma_2^2)^{(k)}}\left(y_{ij}-\beta_{0i}^{(k)}\right)+\chi_i^{(k)}\right)^2}{\frac{\left(\alpha^{(k)}\right)^2}{\left(\sigma_2^2\right)^{(k)}}+1}\right]\right\}\times\left(2\pi\left(\frac{\left(\alpha^{(k)}\right)^2}{\left(\sigma_2^2\right)^{(k)}}+1\right)^{-1}\right)^{-\frac{1}{2}} \\
&\int_0^\infty exp\left\{-\frac{1}{2}\left(\frac{\left(\alpha^{(k)}\right)^2}{\left(\sigma_2^2\right)^{(k)}}+1\right)\left(z-\frac{\frac{\alpha^{(k)}}{(\sigma_2^2)^{(k)}}\left(y_{ij}-\beta_{0i}^{(k)}\right)+\chi_i^{(k)}}{\frac{\left(\alpha^{(k)}\right)^2}{\left(\sigma_2^2\right)^{(k)}}+1}\right)^2\right\}dz \\
=&\left(2\pi(\sigma_2^2)^{(k)}\left(\frac{\left(\alpha^{(k)}\right)^2}{\left(\sigma_2^2\right)^{(k)}}+1\right)\right)^{-\frac{1}{2}} exp\left\{-\frac{1}{2}\frac{\left(y_{ij}-\beta_{0i}^{(k)}-\alpha^{(k)}\chi_i^{(k)}\right)^2}{(\sigma_2^2)^{(k)}\left(\frac{\left(\alpha^{(k)}\right)^2}{\left(\sigma_2^2\right)^{(k)}}+1\right)}\right\} \\
&\times\Phi\left(\frac{\frac{\alpha^{(k)}}{(\sigma_2^2)^{(k)}}\left(y_{ij}-\beta_{0i}^{(k)}\right)+\chi_i^{(k)}}{\left(\frac{\left(\alpha^{(k)}\right)^2}{\left(\sigma_2^2\right)^{(k)}}+1\right)^{\frac{1}{2}}}\right) \\
=&\phi\left(\frac{y_{ij}-\beta_{0i}^{(k)}-\alpha^{(k)}\chi_i^{(k)}}{\left((\sigma_2^2)^{(k)}\left(\frac{\left(\alpha^{(k)}\right)^2}{\left(\sigma_2^2\right)^{(k)}}+1\right)\right)^{\frac{1}{2}}}\right)\times\Phi\left(\frac{\frac{\alpha^{(k)}}{(\sigma_2^2)^{(k)}}\left(y_{ij}-\beta_{0i}^{(k)}\right)+\chi_i^{(k)}}{\left(\frac{\left(\alpha^{(k)}\right)^2}{\left(\sigma_2^2\right)^{(k)}}+1\right)^{\frac{1}{2}}}\right)
\end{aligned}
$$

In Figure 2.22, we show the boxplots and histograms of the residuals for the pattern mixture model and the selection model, which indicate that both proposed models have residuals with mean zero and the standard deviation a little bit smaller than 1 (tighter than expected).

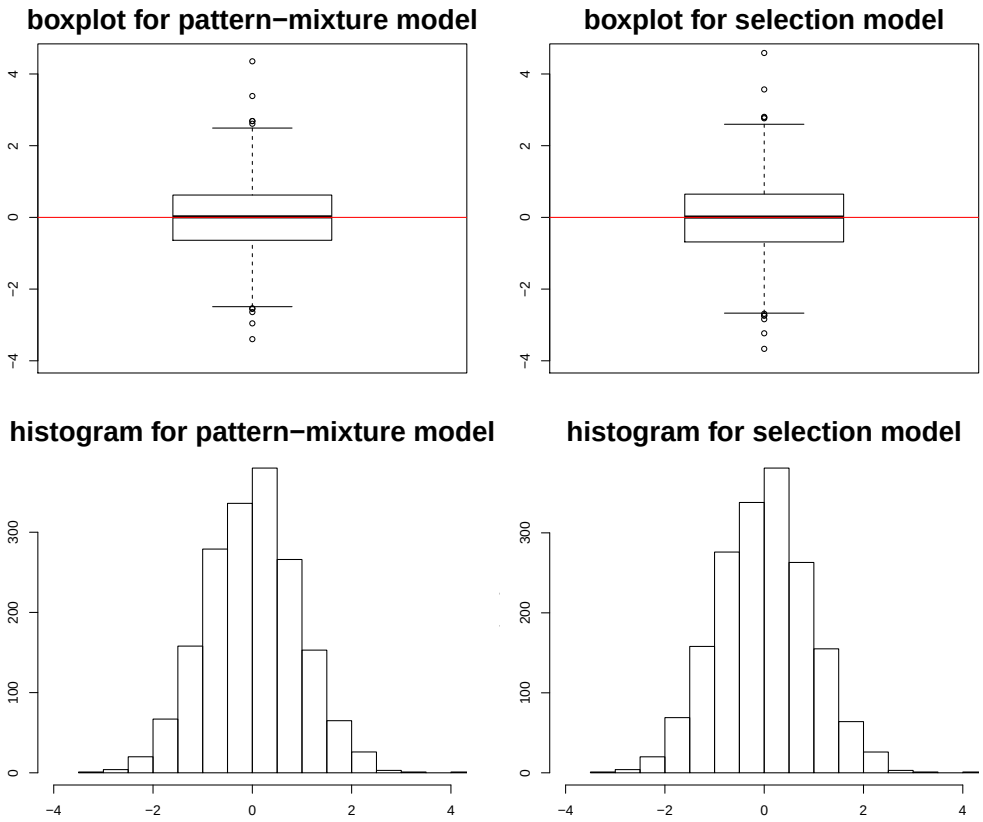

Figure 2.22: Deleted residuals for the cross validation analysis

The performance of both models are similar, so the residual analysis has no preference over the two proposed models based on the observed data only. This result does not surprise us since the residual plot only involve how the model fit the observed part. If the missingness is non-ignorable, we seek for the model that can describe the correct structure for not only the observed part, but more importantly, the missing part. These two approaches can provide limited help on the model fit assessment for the missing part, especially when the missing mechanism is non-ignorable. But they still have positive meanings, since they perform well to assess the model fitting for the observed data.

2.8 Simulation Study

In the previous sections, we applied our proposed Bayesian methods to the NAEP data as well as the modified data, which we artificially delete part of the observed data so that the missing percentage is around 15%. Table 2.5 shows the performance of four methods to estimate the overall average score. The Bayesian methods seem to have smaller differences in the estimates of the original data and the modified data compared with the observed means and the HT estimators. But the true value of the overall average score is unknown in the real data analysis. Also we evaluate the model adequacy of the proposed Bayesian methods by two approaches. But since the true underline structure remains unknown for the real data analysis, we do not have a clear evaluation for these two approaches. So in this section, we conduct a simulation study. We evaluate the performances of the proposed models and the traditional design-based methods under four scenarios which are from a two-factor design. One factor is the true model is either pattern mixture model or selection model while the other factor is the missingness is ignorable or non-ignorable. The details of the simulation setup are given as follows.

First, we determine the sizes of the sampling populations. Each sampling population contains $B = 800$ schools and the number of students in each school N_i is independently and randomly generated from a uniform distribution with a minimum size of 60 and a maximum size of 1000. Then the sampling populations are constructed by using corresponding hierarchical linear model, either the pattern mixture model or the selection model. In the pattern mixture model setup, the parameter α is used to determine the degree of non-ignorability of the missing data mechanism in the student level. We use 0, 20 and 40 as the true values for α. When $\alpha = 0$, the student score value

y_{ij} does not depend on the latent missing variable z_{ij}. Therefore y_{ij} does not depend on the missing indicator r_{ij}, since r_{ij} can be totally determined by the corresponding latent missing variable z_{ij}, for $i = 1, \cdots, B$ and $j = 1, \cdots, N_i$. The generated data have ignorable missing. Similarly, $\alpha = 20$ and $\alpha = 40$, the generated data have non-ignorable missing and for the generated data with $\alpha = 40$, the missingness have more influence on the data, or in another word, the observed data and the missing data are more "separated" than the data generated with $\alpha = 20$. In the selection model setup, λ is the coefficient for y in the $R|Y$ model and serves the same function as α in the pattern mixture model. We use 0, 100 and 200 as the true values. The intercepts for different schools in the $R|Y$ model are randomly and independently generated from a normal distribution with mean 2 and standard deviation 0.5. The hyper parameters in both the pattern mixture model and the selection model are chosen to make sure that the overall student response rates are 85%.

In order to mimic the real data, the student level variance σ_2^2 is chosen to be 1000 while the school level variance σ_1^2 is chosen to be 500. And the student level variability of the missing latent variable z_{ij} is set as 1. χ is set as 1.4, while ω^2 is set as 0.5 in order to obtain various response rate among school and the overall student response rate around 85%. All other parameters in the model will be chosen so that the super-population mean is 318.

Each sampling scheme is repeated for 200 times for each population with the predetermined selection probability which can guarantee that all the students are equally likely to be selected. The finite population mean estimator from each method (UW), the HT estimator (HT) and the model estimator such as the estimator based on the fitted pattern mixture model (PMM) and the estimator based on the fitted selection model (SEM), are computed. We also calculate the real sample mean as a reference point and

denote it as the BD method. We compare those estimators in terms of the following criteria: the empirical bias (which is the mean of the difference between the estimate and the true population mean), the root of the mean squared error (RMSE), the relative rate of the root of the mean squared error (RRMSE), the mean of the estimated standard error (ESTSE) and the coverage rate of the true mean value. Here, RRMSE is defined as the ratio of the difference of the estimate RMSE with the RMSE for the real sample (before deleting the unobserved samples) and the RMSE for the real sample, which is denoted as RMSE(BD). The differences of RMSE in the original scale may be hard to compare. RRMSE, which is a monotone transformation of the RMSE, can amplify the differences and make the comparison result more obvious.

$$\text{RRMSE} = \frac{\text{RMSE} - \text{RMSE(BD)}}{\text{RMSE(BD)}}$$

The following six tables (Table 2.6 - Table 2.11) show the results of the simulation study under six different scenarios that we use to generate the real data: ignorable pattern mixture model structure with $\alpha = 0$, non-ignorable pattern mixture model structure with $\alpha = 20$, non-ignorable pattern mixture model structure with $\alpha = 40$, ignorable selection mixture model structure with $\lambda = 0$, non-ignorable selection model structure with $\lambda = 1$ and non-ignorable selection model structure with $\lambda = 2$.

In Table 2.6 and Table 2.9, which summarize the results of the scenarios that the missingnes is ignorable, the results of the different estimating methods are very similar compared with the reference values of the true sample in terms of all the comparison

	Bias(×100)	RMSE	RRMSE(%)	ESTSE	Non-Coverage Rate(%)
BD	-0.3246	4.4545	0	4.3544	4.2
UW	-19.0780	4.5462	2.0588	4.3793	4.4
HT	3.1234	4.4419	-0.2843	4.4213	4.2
PMM	39.5218	4.7770	7.2396	4.4210	6.2
SEM	29.1811	4.7265	6.1060	4.4160	6.0

Table 2.6: Summary table for data generated from the ignorable pattern mixture model ($\alpha = 0$). After fitting the pattern mixture model, the probability of the credible interval of α covers the true value 0 is 93.4%, the probability of the credible interval of α covers 0 is 0.934. After fitting the selection model, the probability of the credible interval of λ covers 0 is 0.96.

	Bias(×100)	RMSE	RRMSE(%)	ESTSE	Non-Coverage Rate(%)
BD	-5.6842	4.8386	0	5.1487	3.2
UW	613.0934	7.6828	58.7818	4.9008	22.6
HT	461.6481	6.5249	34.8509	5.0472	14.4
PMM	38.7332	5.3324	10.2052	5.1500	6.0
SEM	70.2019	5.3176	9.8992	6.1409	6.2

Table 2.7: Summary table for data generated from the non-ignorable pattern mixture model ($\alpha = 20$). After fitting the pattern mixture model, the probability of the credible interval of α covers the true value 20 is 93%, the probability of the credible interval of α covers 0 is 0.464. After fitting the selection model, the probability of the credible interval of λ covers 0 is 0.516.

criteria. The averages of biases from the true population mean are about the same as that of the true sample mean and the non-coverage rates of mean estimators are all around the nominal level 5%. Table 2.7, Table 2.8, Table 2.10 and Table 2.11 show the results for the data with non-ignorable missing. The sample mean (UW) estimators have large biases, large RMSEs and the non-coverage rates of the mean estimators are higher than the nominal level. The HT (WT) estimators, which adjust the UW estimators with the weights proportional to the response rate, slightly improve the results, but still suffer greatly from the same drawbacks as the UW estimators. The model-based estimators, estimators based on the pattern mixture model or the selection model, outperform the other two design-based estimators in terms of the average bias, RMSE, the average of the standard deviations and the non-coverage rate of the true population

	Bias(×100)	RMSE	RRMSE(%)	ESTSE	Non-Coverage Rate(%)
BD	8.7768	6.5634	0	6.8572	4.6
UW	1097.7829	13.4467	104.8731	6.0777	53.4
HT	840.8640	10.8825	65.8049	6.1902	32.4
PMM	-5.0384	6.7636	3.0492	6.8579	5.2
SEM	26.4852	6.6591	1.4573	6.8388	5.0

Table 2.8: Summary table for data generated from non-ignorable pattern mixture model ($\alpha = 40$). After fitting the pattern mixture model, the probability of the credible interval of α covers the true value 40 is 91.2%, the probability of the credible interval of α covers 0 is 0.002. After fitting the selection model, the probability of the credible interval of λ covers 0 is 0.006.

	Bias(×100)	RMSE	RRMSE(%)	ESTSE	Non-Coverage Rate(%)
BD	22.1321	4.4847	0	4.5538	5.4
UW	8.6748	4.5520	1.5013	4.5760	5.6
HT	24.9697	4.5403	1.2405	4.5903	5.8
PMM	86.2967	5.3263	18.7679	4.6023	9.2
SEM	59.6201	4.7811	6.6093	4.5786	5.6

Table 2.9: Summary table for data generated from the ignorable selection model ($\lambda = 0$). After fitting the selection model, the probability of the credible interval of λ covers the true value 0 is 94.6%, the probability of the credible interval of λ covers 0 is 0.946. After fitting the pattern mixture model, the probability of the credible interval of α covers 0 is 0.978.

mean if the missing mechanism is non-ignorable missing. Furthermore, if the degree of non-ignorability is higher, which associates with the situation that the value of α in pattern mixture model or λ in selection model is larger, then the model-based estimators have bigger improvement especially when we correctly specified model.

We are especially interested in the model fitting when we use the wrong model to fit the data. In Figure 2.23, Figure 2.24 and Figure 2.25, we show the model adequacy check when we fit pattern mixture model to the data that are generated from selection model with $\lambda = 0.02$. In Figure 2.24, the p-value for the observed minimum is 0.975; while in Figure 2.25, the residuals have a non-zero mean. Both figures suggest that the pattern mixture model does not fit the data well.

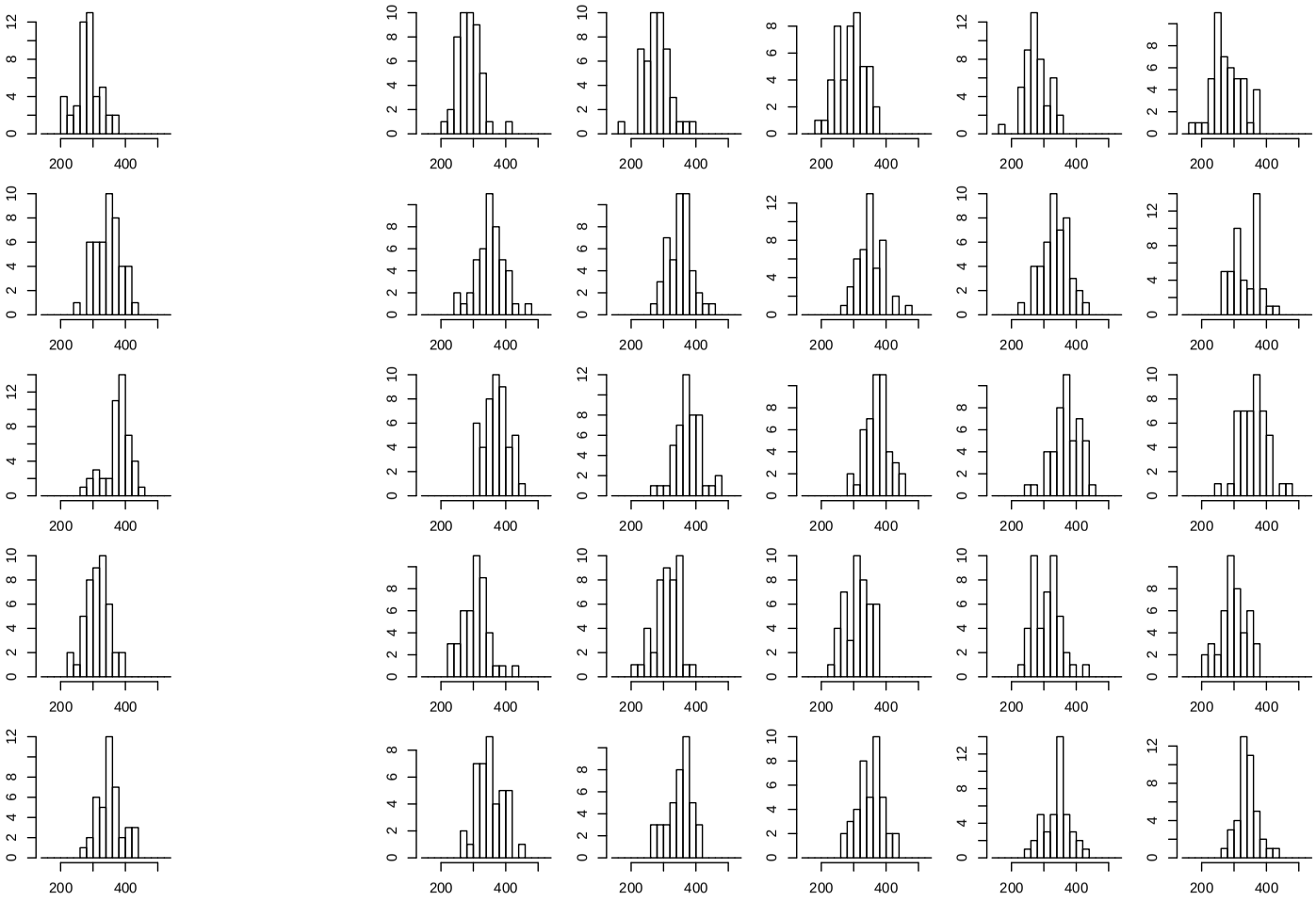

Figure 2.23: Left column displays observed data for the first 5 schools, right columns display 5 replicated data sets y^{rep} from the fitted pattern mixture model of Y when data are generated from selection model

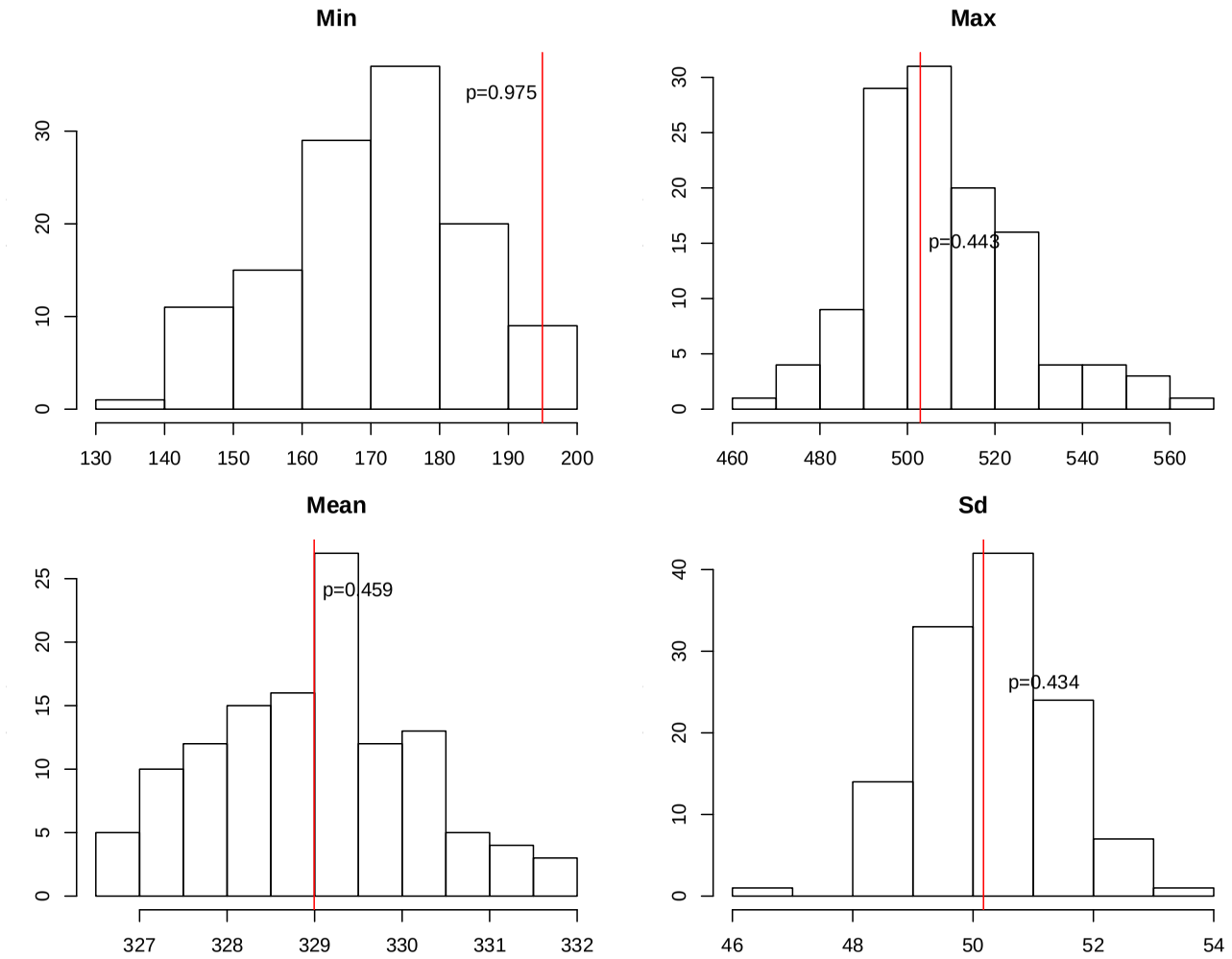

Figure 2.24: Posterior predictive distribution, observed result, and p-value for each of four statistics for PMM model when data are generated from selection model

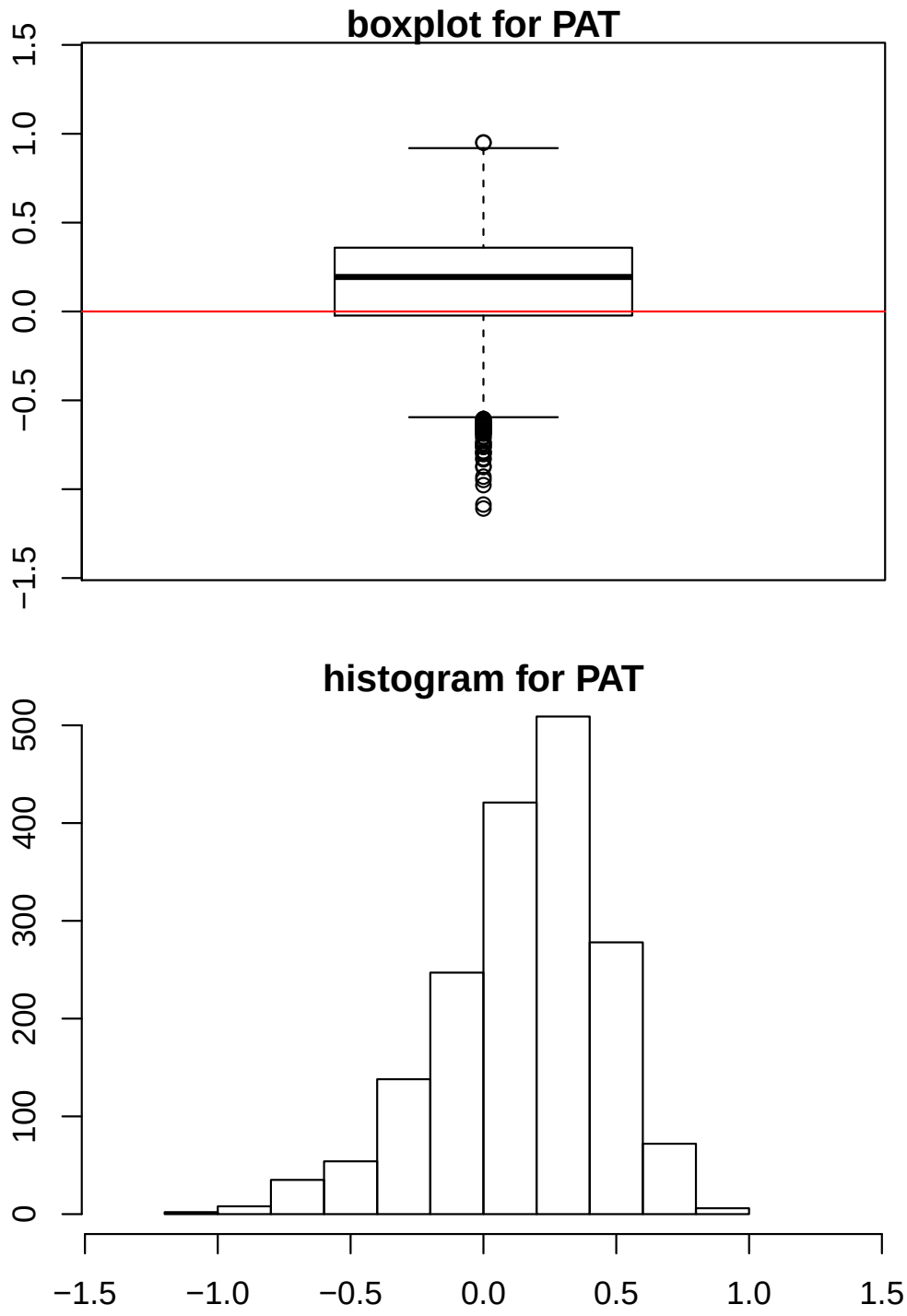

Figure 2.25: Residual plot for the fitted pattern mixture model when data are generated from selection model

	Bias($\times$100)	RMSE	RRMSE(%)	ESTSE	Non-Coverage Rate(%)
BD	3.7445	4.5018	0	4.5239	5.0
UW	304.3346	5.4119	20.2166	4.4769	11.0
HT	214.3838	4.9410	9.7554	4.4881	8.4
PMM	-134.4051	4.9660	10.3103	4.5993	6.6
SEM	-26.4022	4.6253	2.7437	4.5526	5.0

Table 2.10: Summary table for data generated from the non-ignorable selection model ($\lambda = 0.01$). After fitting the selection model, the probability of the credible interval of λ covers the true value 0.01 is 94.2%, the probability of the credible interval of λ covers 0 is 0.466. After fitting the pattern mixture model, the probability of the credible interval of α covers 0 is 0.722.

	Bias($\times$100)	RMSE	RRMSE(%)	ESTSE	Non-Coverage Rate(%)
BD	-1.2639	4.4884	0	4.6227	5.4
UW	535.7515	6.8643	52.9328	4.4169	22.0
HT	388.4103	5.7783	28.7368	4.4243	13.8
PMM	-139.7375	4.9404	10.0711	4.6974	6.8
SEM	-7.9069	4.5950	2.3749	4.6215	6.6

Table 2.11: Summary table for data generated from the non-ignorable selection model ($\lambda = 0.02$). After fitting the selection model, the probability of the credible interval of λ covers the true value 0.02 is 98.6%, the probability of the credible interval of λ covers 0 is 0.238. After fitting the pattern mixture model, the probability of the credible interval of α covers 0 is 0.064.

In Figure 2.26, Figure 2.27 and Figure 2.28, we show the model adequacy checking when we fit selection model to the data which are generated from pattern mixture model with $\alpha = 40$. In Figure 2.27, the observed minimum has p-value 0.984 while the observed mean has p-value 0.992, which are greater than 0.975 (we use a two-sided 95% confidence interval here). In addition, the mean of the residuals in Figure 2.28 is slightly greater than 0. All of the above information suggest that the selection model does not fit the data well.

From the above discussion, we can see that, when we use the wrong model to fit the data, these model checking plots suggest that the wrong model does not fit the data well. In the real data analysis, since we do not know the true distribution of the

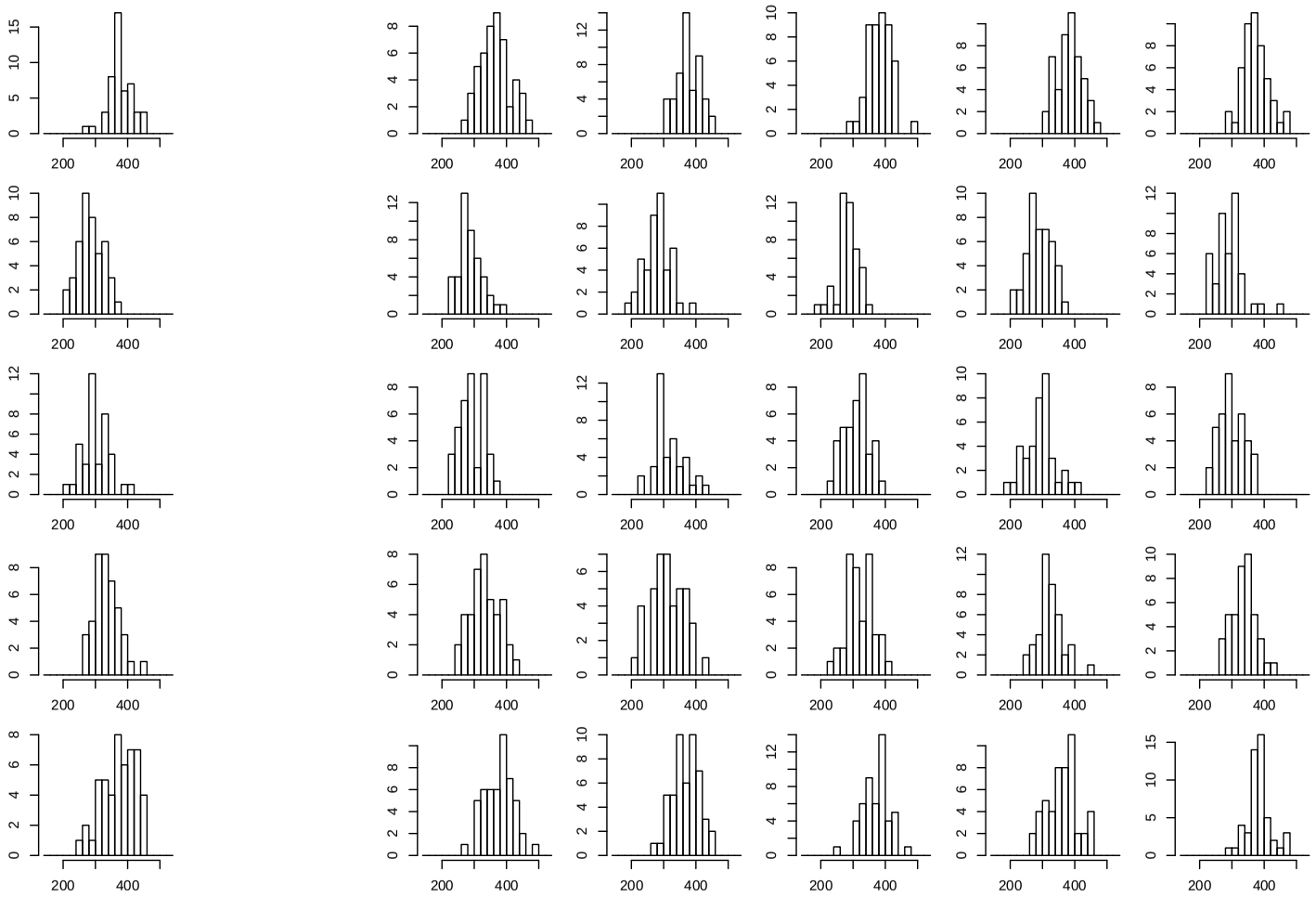

Figure 2.26: Left column displays observed data for the first 5 schools, right columns display 5 replicated data sets y^{rep} from the fitted selection model of Y when data are generated from pattern mixture model

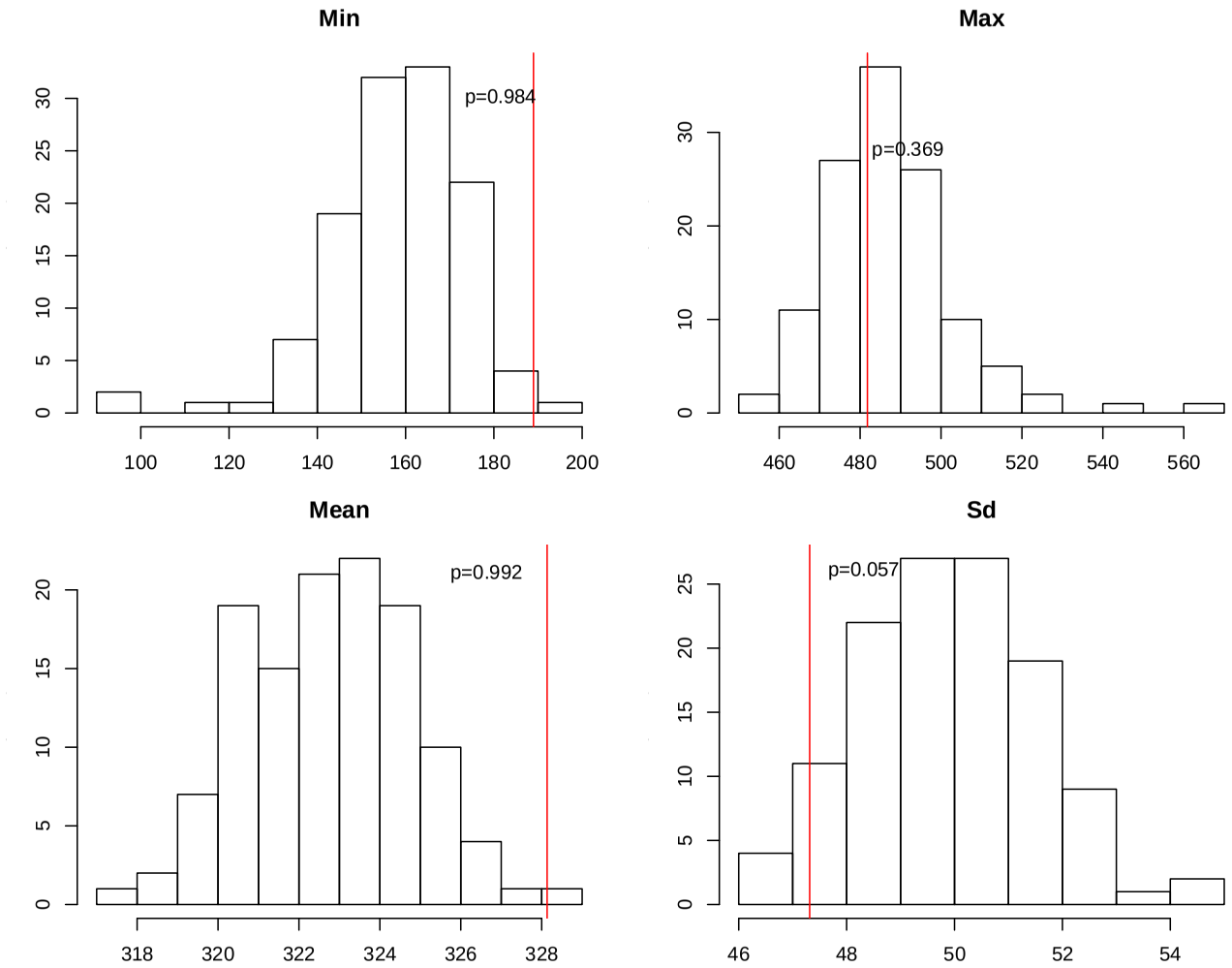

Figure 2.27: Posterior predictive distribution, observed result, and p-value for each of four statistics for SEM model when data are generated from pattern mixture model

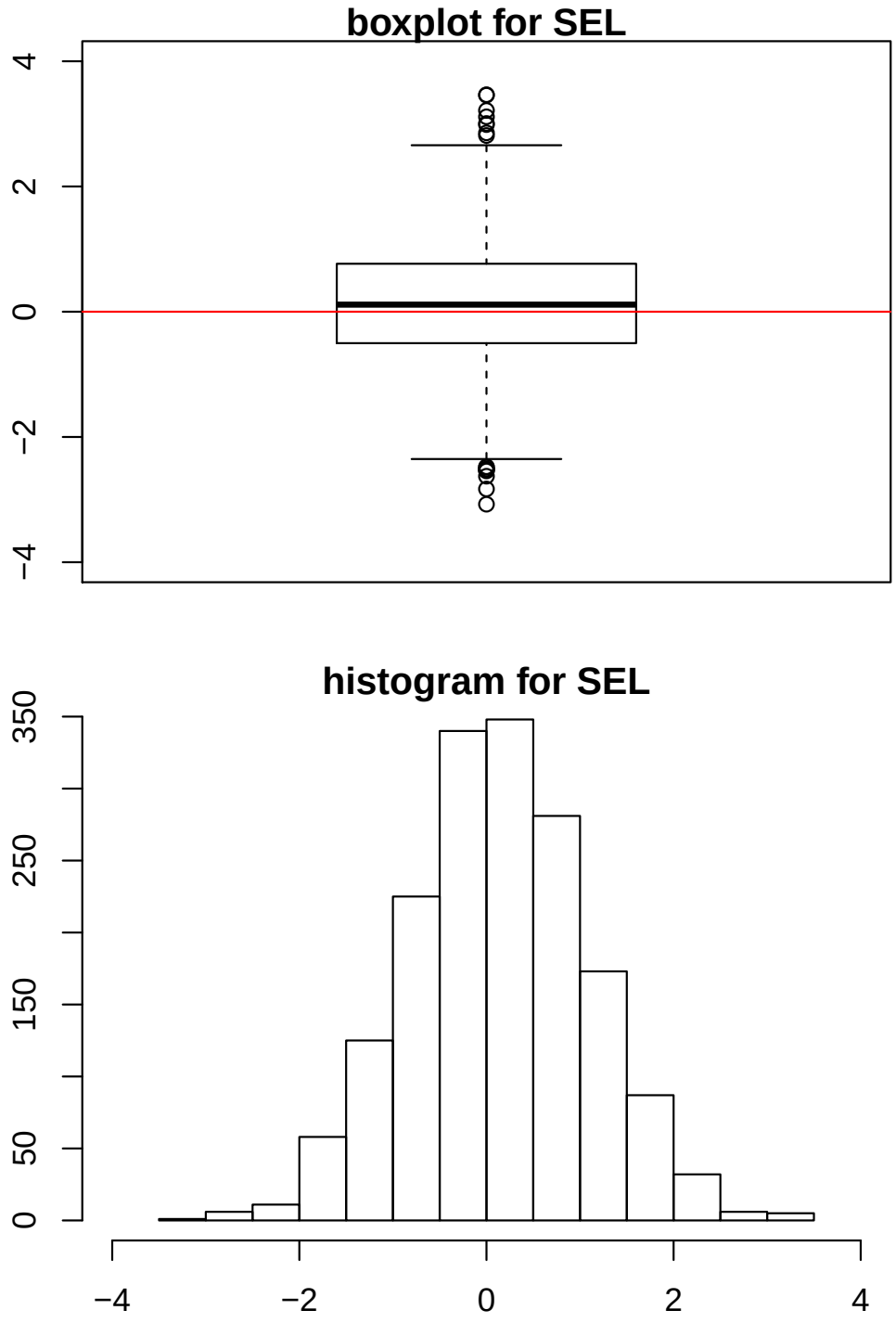

Figure 2.28: Residual plots for the fitted selection model when data are generated from pattern mixture model

data, we need a model selection criterion which can identify the right model even when we have non-ignorable missing.

2.9 Model Selection Problems in Bayesian Statistics

In this section, we discuss several existing approaches in the literature for model comparison. In model selection problems, two great concerns are how the model fits the data and the model complexity. There are generally two kinds of situations: the nested models, and models that have totally different model structures. Two models are said to be nested models if in the proposed models, both of the models contain the same terms and one of the model has at least one more term. The model with the extra term(s) is called full model and the other model is called reduced model. For example the ignorable pattern mixture model and non-ignorable pattern mixture model in Section 2.4. For the nested models, the credible interval technique is used to draw the conclusion of whether to use the full model or the reduced model. To be more specific, at the significant level α, if the parameter in question has a $(1-\alpha)100\%$ credible interval does not contain zero, then we say that this parameter is significant at the α level and we will select the full model.

In Section 2.4 and Sectioin 2.5, we have proposed two different model structures: pattern mixture model and selection model. It is natural to raise the question that which model is true model of the data. In the literature, there are three types of model selection criteria: the Bayes factor, the deviance information criterion (DIC) and the minimum posterior predictive loss approach.

2.9.1 Bayes Factor Approach

The Bayes factor approach is the analogue of likelihood ratio test in the classical hypothesis testing as they share the same formula. Given two models, Bayes factor is the ratio of marginal likelihood of the two proposed models. In our data example, we denote the pattern mixture model as $\mathcal{M}_{pmm}$, and the corresponding parameters as $\boldsymbol{\theta}_{pmm}$, while we denote the selection model as $\mathcal{M}_{sem}$ and the corresponding parameters as $\boldsymbol{\theta}_{sem}$. Then given the observed data $\mathbf{Y}$, the Bayes factor B is

$$B = \frac{P\left(\mathbf{Y}|\mathcal{M}_{pmm}\right)}{P\left(\mathbf{Y}|\mathcal{M}_{sem}\right)} = \frac{\int P\left(\boldsymbol{\theta}_{pmm}|\mathcal{M}_{pmm}\right) P\left(\mathbf{Y}|\boldsymbol{\theta}_{pmm}, \mathcal{M}_{pmm}\right) d\boldsymbol{\theta}_{pmm}}{\int P\left(\boldsymbol{\theta}_{sem}|\mathcal{M}_{sem}\right) P\left(\mathbf{Y}|\boldsymbol{\theta}_{sem}, \mathcal{M}_{sem}\right) d\boldsymbol{\theta}_{sem}}$$

So that the posterior odds is

$$\frac{P(\mathcal{M}_{pmm}|\mathbf{Y})}{P(\mathcal{M}_{sem}|\mathbf{Y})} = \frac{\pi(\mathcal{M}_{pmm})}{\pi(\mathcal{M}_{sem})} \times \frac{P(\mathbf{Y}|\mathcal{M}_{pmm})}{P(\mathbf{Y}|\mathcal{M}_{sem})}$$

$$\text{posterior odds} = \text{prior odds} \times \text{Bayes factor}$$

Usually, the prior odds ratio is set to be 1, so we can directly use the Bayes factor to make the decision.

There are some differences between the classical likelihood ratio test and the Bayes factor approach. The Bayes factor does not depend on any single set of parameters as it solely depend on the model structure assumptions by integrating over all parameters. In addition, for Bayes factor approach, the candidate parameter sets do not need to share the same parameter space. One more thing worth to point out is that the Bayes factor also depends on the prior information of parameters.

While the Bayes factor approach is quite intuitive, unfortunately, it has some drawbacks. First, it is often very difficult to calculate in practice due to complexity of the integrating step. Second, the Bayes factor approach is based on the assumption that one of the candidate models is the true model, while in reality, the truth is unknown. Third, for certain choice of the prior distributions, such as diffuse prior, the Bayes factor approach may make different conclusion from what the classical likelihood ratio test makes. This phenomenon is called Lindley-Barlett's paradox (Lindley, 1957 and Shafer, 1982).

Our motivating data is a multilevel data, and the model structure involves the missing latent variable. The calculation of the Bayes factor is a non-trivial task. So we do not conduct the Bayes factor due to its computation complexity.

2.9.2 Deviance Information Criterion Approach

DIC is a very popular criterion in Bayesian model selection. It was proposed by Spiegelhalter et al. (2002) and is a generalization of AIC and BIC. The deviance is defined as

$$D(\boldsymbol{\theta}) = -2\log f(\mathbf{Y}|\boldsymbol{\theta}) + 2\log g(\mathbf{Y})$$

where $g(\mathbf{Y})$ is a function of $\mathbf{Y}$ alone and is fully specified. It can serve both as the measurement of model fit and the measurement of model complexity. The model complexity measurement, which is described as the effective number of parameters, p_D, is defined as

$$p_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}})$$

where $\overline{D(\boldsymbol{\theta})}$ is the posterior mean deviance,

$$\overline{D(\boldsymbol{\theta})} = E_{\boldsymbol{\theta}}\left[-2\log f(\mathbf{Y}|\boldsymbol{\theta})|\mathbf{Y}\right] + 2\log g(\mathbf{Y})$$

which is also functioned as the measurement of goodness of fit.

The DIC is constructed as the combination of the model fit measurement and the model complexity measurement.

$$\begin{aligned}\text{DIC} &= \overline{D(\boldsymbol{\theta})} + p_D \\ &= 2\overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}) \\ &= D(\bar{\boldsymbol{\theta}}) + 2p_D\end{aligned}$$

The last equality shows that DIC agrees with the classical AIC

$$\text{AIC} = -2\log L(\hat{\boldsymbol{\theta}}) + 2p$$

Where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate and p is the number of parameters in the classical setting. Furthermore, for a non-hierarchical model with non-informative prior of $\boldsymbol{\theta}$, we have

$$\text{DIC} = \text{AIC}$$

Based on the above definition of DIC, one may choose the candidate model with the smallest DIC value.

Although right now, DIC is a popular choice of Bayesian modeling assessment and comparison, some people also discussed the ambiguity of DIC, especially when the model is a mixture model (DeIorio, 2002). Celeux (2006) further introduced the eight

different versions of DIC for hierarchical models when missing data present. The author stated that

> "The fundamental versatility of the DIC criterion is that, in hierarchical models, basic notions like parameters and deviance may take several equally acceptable meanings, with direct consequences for the properties of the corresponding DICs."

In our data example, both the pattern mixture model and selection model are solved with the help of the missing latent variables and the missing latent variables also have the hierarchy feature. Treating the missing latent variables as parameters will definitely increase the total number of parameters, but the question of which layer we want to include or which level is our focus for the missing latent variables is very hard to answer. In addition, one of the advantages of DIC is easy to implement. With a converged MCMC chain, the measurement of fit is the average of the log-likelihoods of the selected iterations, while the measurement of model complexity, $D(\bar{\boldsymbol{\theta}})$ is the log-likelihood calculated at the end of the Gibbs sampler using the posterior means (or modes) of $\boldsymbol{\theta}$. For our situation, the existence of missing value $\mathbf{Y}_{mis}$ and the missing latent variables $\mathbf{u}$ makes the calculation a non-trivial work. What's more, the DIC approach inherits all the drawbacks from the AIC approach. So when sample size increases, the DIC approach tends to select the more complicated model. Based on the above reasons, especially that we are not clear which level should we focus on, although DIC is well-known in Bayesian model selection, we do not conduct the DIC approach.

2.9.3 Minimum Posterior Predictive Loss Approach

The third solution is the minimum posterior predictive loss approach (Gelman, 1996). With the commonly used squared error loss function, the proposed model

selection criterion is

$$\text{PPL}_k(\mathcal{M}_i) = \sum_{l=1}^{n} Var\left(y_l^{rep}|\mathbf{Y}_{obs}, \mathcal{M}_i\right) + \frac{k}{k+1}\sum_{l=1}^{n}\left(E\left(y_l^{rep}|\mathbf{Y}_{obs}, \mathcal{M}_i\right) - y_{l,obs}\right)^2 \quad (2.5)$$

where $y_{l,obs}$ denotes the lth observed observation in $\mathbf{Y}_{obs}$, $l = 1, \cdots, n$. We select the candidate model with the smallest PPL_k.

In particular, let $k \to \infty$ in (2.5), we have

$$\text{PPL}(\mathcal{M}_i) = \lim_{k\to\infty} \text{PPL}_k(\mathcal{M}_i) = \sum_{l=1}^{n} Var\left(y_l^{rep}|\mathbf{Y}_{obs}, \mathcal{M}_i\right) + \sum_{l=1}^{n}\left(E\left(y_l^{rep}|\mathbf{Y}_{obs}, \mathcal{M}_i\right) - y_{l,obs}\right)^2$$

which corresponds to the mean square error criterion in the classical statistics setting.

We implement the posterior predictive loss approach in our data example, and surprisingly find out that this minimum posterior predictive loss approach fails. In Table 2.12, we list the chosen rates of the model structures under three circumstances which we use to generate the data: the ignorable model, the non-ignorable pattern mixture model with $\alpha = 20$ and the non-ignorable selection model with $\lambda = 1$. The number of candidate models are three instead of four since the ignorable pattern mixture model and the ignorable selection model have the same structure. We can see that no matter what model structure we used to generate the data, this approach always favors the non-ignorable pattern mixture model compared with ignorable and non-ignorable selection models in the simulation study described in the subsection 2.8. So we further studied the following toy example to evaluate the performance of PPL approach.

Suppose we have two candidate models for a complete data set $\mathbf{Y}$ containing $b \times m$ observations without missingness.

True Model	Ignorable	Non-ignorable PMM	Non-ignorable SEM
Ignorable	0	100	0
Non-ignorable PMM	0	100	0
Non-ignorable SEM	0	100	0

Table 2.12: The chosen rate (%)

In the first model, we treat the data as they are from b separate groups. Then the proposed model is

$$y_{ij} \sim N\left(\beta_{0i}, \sigma_2^2\right) \quad \text{for } i = 1, \cdots, b,\ j = 1, \cdots, m.$$

Then, the posterior distributions for the parameters are:

$$\sigma_2^2|\mathbf{Y}_{obs} \sim IG\left(a_1 + \frac{1}{2}b(m-1), b_1 + \frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{m}\left(y_{ij} - \bar{y}_i\right)^2\right)$$

$$\beta_{0i}|\mathbf{Y}_{obs}, \sigma_2^2 \sim N\left(\bar{y}_i, \frac{\sigma_2^2}{m}\right)$$

The posterior predictive distribution of y_{ij} is

$$\begin{aligned}
&f\left(y_{ij}^{rep}|\mathbf{Y}_{obs}, \sigma_2^2\right) = \int \exp\left\{-\frac{1}{2\sigma_2^2}\left(y - \beta_{0i}\right)^2\right\} \times \exp\left\{-\frac{m}{2\sigma_2^2}\left(\beta_{0i} - \bar{y}_i\right)^2\right\} d\beta_{0i} \\
=&\int \exp\left\{-\frac{1}{2\sigma_2^2}\left[(m+1)\beta_{0i}^2 - 2\left(y + m\bar{y}_i\right)\beta_{0i} + \left(y^2 + m\bar{y}_i^2\right)\right]\right\} d\beta_{0i} \\
\propto &\exp\left\{-\frac{1}{2\sigma_2^2}\left(y^2 - \frac{\left(y + m\bar{y}_i\right)^2}{m+1}\right)\right\} \\
=&N\left(\bar{y}_i, \frac{m+1}{m}\sigma_2^2\right)
\end{aligned}$$

Following the law of total variance, we have

$$\begin{aligned}
&Var\left(y_{ij}^{rep}|\mathbf{Y}_{obs}\right) = Var\left(E\left(y_{ij}^{rep}|\mathbf{Y}_{obs},\sigma_2^2\right)\right) + E\left(Var\left(y_{ij}^{rep}|\mathbf{y}_{obs},\sigma_2^2\right)\right) \\
=&Var\left(\bar{y}_i|\mathbf{Y}_{obs},\sigma_2^2\right) + E\left(\frac{m+1}{m}\sigma_2^2\right) \\
=&\frac{m+1}{m}\cdot\frac{b_1+\frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{m}\left(y_{ij}-\bar{y}_i\right)^2}{a_1+\frac{1}{2}b(m-1)-1}
\end{aligned}$$

And also the square of the bias is

$$\left(E\left(y_{ij}^{rep}|\mathbf{y}_{obs}\right)-y_{ij}\right)^2 = \left(\bar{y}_i-y_{ij}\right)^2$$

Denote the posterior predictive loss for the first model as $\mathrm{PPL}_k(\mathcal{M}_1)$, then

$$\begin{aligned}
\mathrm{PPL}_k(\mathcal{M}_1) &= \sum_{i=1}^{b}\sum_{j=1}^{m} Var\left(y_{ij}^{rep}|\mathbf{y}_{obs}\right) + \frac{k}{k+1}\sum_{i=1}^{b}\sum_{j=1}^{m}\left(E\left(y_{ij}^{rep}|\mathbf{y}_{obs}\right)-y_{ij}\right)^2 \\
&= b(m+1)\cdot\frac{b_1+\frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{m}\left(y_{ij}-\bar{y}_i\right)^2}{a_1+\frac{1}{2}b(m-1)-1} + \frac{k}{k+1}\sum_{i=1}^{b}\sum_{j=1}^{m}\left(\bar{y}_i-y_{ij}\right)^2
\end{aligned}$$

In the second model, we treat all the data are all from one group. Then the second proposed model is

$$y_{ij} \sim N\left(\beta_0,\sigma_1^2\right) \quad \text{for } i=1,\cdots,b,\ j=1,\cdots,m.$$

and denote the posterior predictive loss for the second model as $\mathrm{PPL}_k(\mathcal{M}_2)$, then simi-

larly we have

$$\begin{aligned}\mathrm{PPL}_k(\mathcal{M}_2) &= \sum_{i=1}^{b}\sum_{j=1}^{m} Var\left(y_{ij}^{rep}|\mathbf{y}_{obs}\right) + \frac{k}{k+1}\sum_{i=1}^{b}\sum_{j=1}^{m}\left(E\left(y_{ij}^{rep}|\mathbf{y}_{obs}\right) - y_{ij}\right)^2 \\ &= (bm+1)\cdot\frac{b_1 + \frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{m}(y_{ij}-\bar{y})^2}{a_1 + \frac{1}{2}(bm-1) - 1} + \frac{k}{k+1}\sum_{i=1}^{b}\sum_{j=1}^{m}(\bar{y}-y_{ij})^2\end{aligned}$$

So the difference between $\mathrm{PPL}_k(\mathcal{M}_1)$ and $\mathrm{PPL}_k(\mathcal{M}_2)$ is

$$\begin{aligned}&\mathrm{PPL}_k(\mathcal{M}_1) - \mathrm{PPL}_k(\mathcal{M}_2) \\ =&b(m+1)\cdot\frac{b_1 + \frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{m}(y_{ij}-\bar{y}_i)^2}{a_1 + \frac{1}{2}b(m-1) - 1} - (bm+1)\cdot\frac{b_1 + \frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{m}(y_{ij}-\bar{y})^2}{a_1 + \frac{1}{2}(bm-1) - 1} \\ &- \frac{k}{k+1}\cdot m\sum_{i=1}^{b}(\bar{y}_i-\bar{y})^2 \\ =&-\frac{k}{k+1}\cdot m\sum_{i=1}^{b}(\bar{y}_i-\bar{y})^2 + \frac{1}{\left(a_1 + \frac{1}{2}b(m-1)-1\right)\left(a_1 + \frac{1}{2}(bm-1)-1\right)} \\ &\left\{\left[b(m+1)\left(a_1 + \frac{1}{2}(bm-1) - 1\right) - (bm+1)\left(a_1 + \frac{1}{2}b(m-1) - 1\right)\right]\right. \\ &\left.\times\left(b_1 + \frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{m}(y_{ij}-\bar{y}_i)^2\right) - (bm+1)\left(a_1 + \frac{1}{2}b(m-1)-1\right)\cdot\frac{m}{2}\sum_{i=1}^{b}(\bar{y}_i-\bar{y})^2\right\} \\ =&\frac{1}{\left(a_1 + \frac{1}{2}b(m-1)-1\right)\left(a_1 + \frac{1}{2}(bm-1)-1\right)}\left\{(a_1+bm-1)\left(b_1 + \frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{m}(y_{ij}-\bar{y}_i)^2\right.\right. \\ &\cdot(b-1) - 2\left(a_1 + \frac{1}{2}b(m-1)-1\right)\left(a_1 + bm - 1 - \frac{1}{k+1}\left(a_1 + \frac{1}{2}(bm-1)-1\right)\right) \\ &\left.\left.\cdot\frac{m}{2}\sum_{i=1}^{b}(\bar{y}_i-\bar{y})^2\right\}\end{aligned}$$

Since the hyper parameter of the scale parameter are chosen to be non-influential ($a_1 =$

0.1 and $b_1 = 0.1$), if

$$\begin{aligned}
&\frac{\sum_{i=1}^{b}\left(\bar{y}_i-\bar{y}\right)^2/(b-1)}{2\left(b_1+\frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{m}\left(y_{ij}-\bar{y}_i\right)^2\right)/(b(m-1))} \\
&> \frac{b(m-1)}{2m\left(a_1+\frac{1}{2}b(m-1)-1\right)} \times \frac{a_1+bm-1}{\left(a_1+bm-1-\frac{1}{k+1}\left(a_1+\frac{1}{2}(bm-1)-1\right)\right)} \\
&\approx \frac{2(k+1)}{2k+1}\cdot\frac{1}{m}
\end{aligned}$$

then $\mathrm{PPL}_k(\mathcal{M}_1) < \mathrm{PPL}_k(\mathcal{M}_2)$.

Here $\frac{\sum_{i=1}^{b}(\bar{y}_i-\bar{y})^2/(b-1)}{2\left(b_1+\frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{m}(y_{ij}-\bar{y}_i)^2\right)/(b(m-1))} \approx \frac{\sum_{i=1}^{b}(\bar{y}_i-\bar{y})^2/(b-1)}{\sum_{i=1}^{b}\sum_{j=1}^{m}(y_{ij}-\bar{y}_i)^2/(b(m-1))}$, which is the F statistics that we usually use in the ANOVA analysis of k-sample equal means test. If $F > F_{\alpha,b-1,b(m-1)}$, we reject the null hypothesis of equal means and treat the data as they are from different groups (the first proposed model). We can easily choose some value d, such that $\frac{2(k+1)}{2k+1}\cdot\frac{1}{m} < d < F_{\alpha,b-1,b(m-1)}$. Since the following two inequalities usually hold: $\frac{2(k+1)}{2k+1}\cdot\frac{1}{m} < 1$ and $F_{\alpha,b-1,b(m-1)} > 1$. Then the PPL approach and the ANOVA analysis draw conflicting conclusions. So the PPL approach fails to correctly identify the right model structure.

Concluding remarks for model selection criteria

We draw the following conclusions of the model selection issue for our motivating multi-level data example. First, the Bayes factor analysis involves the undesirable complicated integration calculation. Secondly, there are uncertainty problems for the DIC approach: we are not clear about which level of the information we want to focus on. Last, we conduct the minimum posterior predictive loss approach in a simulation study, which shows that this approach fails to correctly identify the true underline data structure. In addition, we show that it has conflicting conclusions with the traditional ANOVA test even for a completely observed data set.

Furthermore, we should note that all the above three Bayesian model selection criteria are only built on the observed data while the missingness is not taken into consideration. We can easily construct two data sets that have the observed data in the same range but have different missing mechanisms. For example, we can generate the data from standard normal distribution and then we randomly mark 15% of the observation. The first data set includes the observations that are not marked, while in the second data set, we add 100 to the marked data so that the marked data have a normal distribution with mean 100 and standard deviation 1. Based on the modified data, we delete the observations which are greater than 50. In the above setting, we artificially make the two data sets have the same observed data. But obviously the first data set is missing completely at random while the second data set is not missing at random. In this case, the observed likelihood of these two data sets are exactly the same, but obviously this two data sets have different missing mechanisms. Therefore, the model selection result by any of the three model selection criteria will be the same regardless which missing mechanism is the true one. This inherent identifiability issue will make it more difficult to identify the underlying missing mechanism by any model selection criterion.

Chapter 3

Non-parametric Test of Missing Completely at Random for Multivariate Missing Data

3.1 Background

In previous chapter, we show that when missing data present, the choice of the appropriate statistical approaches not only relies on the data structure but also relies on the type of the missing mechanism. For ignorable missing, both the design-based methods and the model-based Bayesian methods can provide appropriate unbiased estimators. While for non-ignorable missing, the model-based Bayesian methods outperform the design-based methods by providing the estimators with less biases, less mean square errors and better coverage rates of the true value. This phenomenon suggests that correctly identifying the missing mechanism is very important before adopting any statistical approaches. In addition, with a correctly specified model, the Bayesian meth-

ods are very powerful tools in the missing data analysis. The Bayesian approaches can not only identify the missing mechanism, whether the missingness is ignorable or non-ignorable, by performing a hypothesis test on the coefficient parameter in the conditional model of $\mathbf{Y}$ and $\mathbf{R}$, one given another, but also provide a model-based estimator, which eliminates the possible bias bringing by the non-ignorable missing. But sometimes, the Bayesian methods are inappropriate since these methods require a pre-specified model structure. The wrongly specified model structure may cause the analysis results inaccurate. Therefore, there is a great need for a more flexible method, such as a model-free test, of missing mechanism in the missing data analysis.

Our motivating data is a multilevel data with complex structures, directly working on the non-parametric solution for this data is a challenging task. Here we start from a simpler case, the test for incomplete multivariate data. The multivariate data have arisen frequently in real data analysis, such as surveys with multiple questions and analyses that each subject is measured for several different variables of interest. The missingness may occur due to different reasons, some of the participants do not show up or refuse to answer some of the questions, while some of the participants accidently skip some of the questions. The missing data issue is extremely critical in multivariate data analysis due to the complexity of the multivariate data structure. Traditionally, researchers usually only use the information from the subjects that are fully observed, which may not only largely reduce the sample size, but more importantly, may cause the sample not a good representative of the true population. So in this chapter, we propose a non-parametric test of missing completely at random (MCAR) for multivariate missing data.

MCAR is the strictest type of missing mechanism among the three of them. A lot of statistical analysis approaches rely on the assumption of MCAR, such as the

generalized estimating equations method. In the literature, test of MCAR has gained a lot of attention. It has been developed in several different areas ((e.g., contigency tables, Fuchs, 1982; generalized estimating equations, Chen and Little, 1999, Qu and Song, 2002). For multivariate missing data, people found out that the test of MCAR is equivalent to test the homogeneity of distributions among different missing-pattern groups. The missing pattern is defined by the missing indicator: if two subjects have identical missing indicators, we say they have the same missing pattern and belong to the same missing-pattern group. By this way, the whole data matrix is re-arranged and divided into groups according to their missing patterns. The missing pattern concept is illustrated in Figure 3.1 quoted from Little and Rubin (2002).

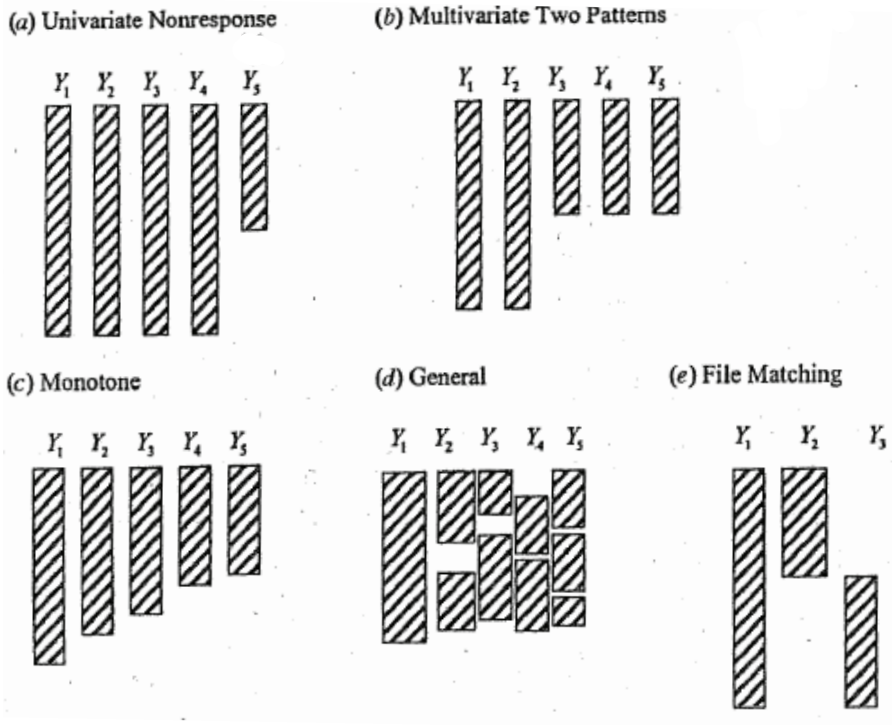

Figure 3.1: Examples of Missing Patterns: rows correspond to observations, columns to variables

In Figure 3.1, we show five examples of missing patterns. In the figure, the rows correspond to the subjects and columns correspond to the variables of interest.

The subplot (a) shows the case that missingness is confined to a single variable. The subplot (b) shows the case that a subset of subjects does not complete the questionnaire. The subplot (c) shows the case that are very common in longitudinal study: dropout situation. The subplot (d) is the general case that the missingness can show up anywhere. The subplot (e) shows the case of matching files: there are two versions of the questionnaires and different people answer different versions; we need to match two sets of incomplete data together to get the complete data set. Our aim is to propose a test that can be used even for the general case.

Little (1988) first proposed a test of MCAR for incomplete multivariate data by testing the homogeneity of means across different missing-pattern groups. The test is based on the likelihood ratio test assuming the normality for the data. Little (1988) also mentioned a likelihood ratio test for testing homogeneity of both means and covariances across different missing-pattern groups as another possible test of MCAR. However, as noticed both in Little (1988) and Kim and Bentler (2002), this test may not work well for small or medium sized samples. To overcome this restriction, Kim and Bentler (2002) proposed a test of homogeneity of both means and covariances across different missing-pattern groups based on generalized least squares. Under the normality assumption, this test can be also used to test MCAR assumption.

The tests in both Little (1988) and Kim and Bentler (2002) were developed under normality assumption, which is not always appropriate in real data analysis. For example, we may encounter the skewed data, such as survival data, or heavy-tail data. The above mentioned tests will fail on the non-normality scenarios. In comparison, nonparametric tests are usually more flexible in accommodating different distributions, hence are more desirable. Recently, Jamshidian and Jalal (2010) proposed a nonparametric test of MCAR, which focuses only on testing homogeneity of covariances across

different missing-pattern groups. However, in practice, if the missingness is not completely at random, it may cause other distributional differences, for example, mean differences, skewness differences, among the different missing-pattern groups. In this chapter, we propose a nonparametric test of MCAR which is capable of detecting any distributional differences in the observed data across the different missing-pattern groups if the missingness is not completely at random. The test is completely nonparametric and therefore does not require any distributional assumption for the data. Furthermore, unlike most of the existing tests for MCAR which require fairly large number of observations in each missing-pattern group, our test can be applied to any multivariate data with missing values, no matter how small the sample size is within each missing-pattern group, as long as we have reasonable size of completely observed cases. Our simulation study shows that the proposed test has well controlled Type I errors under a variety of simulation settings and also has good power against a variety of MAR and MNAR alternatives.

3.2 Notation and the Hypothesis Testing Problem

For incomplete multivariate data, we use the following notation. Let $\mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_n)'$ be the data matrix with n cases, and for each k ($k = 1, ..., n$), $\mathbf{y}_k$ is a vector of p variables. Some of the n cases do not have complete observations for all the p variables. Based on their missing-patterns, the data matrix is divided into different missing-pattern groups so that the data have the same set of missing variables within each group. Let s be the total number of missing-pattern groups in the data, n_i be the number of cases in the ith missing-pattern group and $\sum_{i=1}^{s} n_i = n$. We denote the data matrix for the ith missing-pattern group by $\mathbf{Y}_i = (\mathbf{y}_{i1}, \ldots, \mathbf{y}_{in_i})'$, where $\mathbf{y}_{ij}$ is the vector

for the jth case in the ith missing-pattern group. We assume that there always exist some cases where all the p variables are completely observed. Without loss of generality, we choose the group that contains all the complete cases as our first missing-pattern group $\mathbf{Y}_1$.

Let F_i be the complete joint distribution of all the p variables for the data in the ith missing pattern group $\mathbf{Y}_i$, $i = 1, ..., s$. As mentioned in the previous section, data are MCAR if missingness does not depend on the data, missing or observed. This definition of MCAR implies the following.

Proposition 1. *(Equivalency) The missingness is MCAR if and only if* $F_1 = \cdots = F_s$.

The above result then suggests that testing MCAR for incomplete multivariate data is equivalent to the following hypothesis testing problem,

$$H_0 : F_1 = \cdots = F_s \text{ versus } H_1 : \text{there exists } i \neq j \in \{1, ..., s\}, \text{ such that } F_i \neq F_j. \quad (3.1)$$

Since the data have the same set of missing variables within each missing-pattern group, we use $\mathbf{o}_i$ and $\mathbf{m}_i$ as the subsets of $\{1, 2, ..., p\}$ indicating which variables are observed and which variables are missing for group i, respectively. We further define $F_{i,\mathbf{o}_i}$ and $F_{i,\mathbf{m}_i}$ as the joint distributions of the observed variables and missing variables, respectively, in group i. For example, with $p = 5$, for missing-pattern group i, if the first and last variables are missing and the others are observed, then $\mathbf{o}_i = \{2, 3, 4\}$, $\mathbf{m}_i = \{1, 5\}$, $F_{i,\mathbf{o}_i}$ is the joint distribution of the second, third and fourth variables, and $F_{i,\mathbf{m}_i}$ is the joint distribution of the first and fifth variables. Clearly, the complete joint distribution F_i is the joint distribution of $F_{i,\mathbf{o}_i}$ and $F_{i,\mathbf{m}_i}$. Since there is no observation available for the variables in $\mathbf{m}_i$ in missing-pattern group i, it is impossible to make inferences about the underlying distribution $F_{i,\mathbf{m}_i}$ without any further assumptions

about the missing mechanism. Therefore, there is no way to know whether the complete joint distribution F_i are all equal, since the $F_{i,\mathbf{m}_i}$ as part of the F_i are not inferable. This implies that the null hypothesis in (3.1) or MCAR can not be tested without any further assumptions about the missing mechanism. Kim and Bentler (2002) also recognized this difficulty by stating "this (testing MCAR) is basically impossible without making strong assumptions about the missing data process". In Little (1988), Kim and Bentler (2002) and Jamshidian and Jalal (2010), MAR is assumed for the missing mechanism in order to carry out the proposed tests. Different from their approaches, we do not impose any assumption about the underlying missing mechanism, since this is what we want to test in the first place. Instead, we consider a null hypothesis implied by that in (3.1). Based on this modified null hypothesis, a nonparametric testing procedure can be developed.

Before we introduce our new null hypothesis, we briefly explain the idea behind it. We first define $\mathbf{o}_{ij}$ as the intersection of sets $\mathbf{o}_i$ and $\mathbf{o}_j$. Therefore, $\mathbf{o}_{ij}$ indicates the variables that are observed for both groups i and j. If $\mathbf{o}_{ij} \neq \emptyset$, we further denote the joint distributions of the variables in $\mathbf{o}_{ij}$ from groups i and j by $F_{i,\mathbf{o}_{ij}}$ and $F_{j,\mathbf{o}_{ij}}$, respectively. As mentioned above, for groups i and j, no information can be drawn for $F_{i,\mathbf{m}_i}$ and $F_{j,\mathbf{m}_j}$ without any further assumption. If $\mathbf{o}_{ij} = \emptyset$, it is not possible to compare F_i and F_j. If $\mathbf{o}_{ij} \neq \emptyset$, in order to compare F_i and F_j, the best we can do based on the observed data is to compare $F_{i,\mathbf{o}_{ij}}$ and $F_{j,\mathbf{o}_{ij}}$. This motivates us to consider the following hypothesis testing problem:

$$H_0 : F_{i,\mathbf{o}_{ij}} = F_{j,\mathbf{o}_{ij}} \text{ for all } i \neq j \in \{1, ..., s\} \text{ and } \mathbf{o}_{ij} \neq \emptyset$$

versus

$$H_1 : \text{there exists } i \neq j \in \{1, ..., s\} \text{ and } \mathbf{o}_{ij} \neq \emptyset \text{ such that } F_{i,\mathbf{o}_{ij}} \neq F_{j,\mathbf{o}_{ij}}. \tag{3.2}$$

It is clear that the null hypothesis in (3.1), i.e., MCAR, implies the null hypothesis in (3.2). Therefore, the testing procedure proposed for testing H_0 in (3.2) in the following section can be also used for testing MCAR. When used for testing MCAR, the proposed testing procedure can still control the type I error at the nominal level. In other words, when the missingness is MCAR, the probability for our proposed testing procedure to falsely reject the null hypothesis that the missingness is MCAR is controlled at the α-level. Therefore, our proposed testing procedure remains a valid test for MCAR.

When we reject H_0 in (3.2), it implies that the null hypothesis in (3.1) can not be true either, therefore we can easily conclude that the missingness is not MCAR. When we fail to reject H_0 in (3.2), we may not be able to conclude that the null hypothesis in (3.1) is true as well, since the F_i may be different and the difference between the F_i lies in the $F_{i,\mathbf{m}_i}$. In those cases, the missingness is MNAR. In other cases, the F_i may be the same, and hence the missingness is MCAR. In the following, we give a simple example showing these two possibilities.

First, we generate a random sample, $\mathbf{y}_1, ..., \mathbf{y}_{20}$, from a bivariate normal distribution with mean $(0, 0)'$ and covariance matrix I_2, where I_p stands for the p-dimensional identity matrix. For the first 10 observations, $\mathbf{y}_1, ..., \mathbf{y}_{10}$, we keep both variables of each observation, therefore both variables of those observations are completely observed. In the last 10 observations, $\mathbf{y}_{11}, ..., \mathbf{y}_{20}$, we delete the second variable of each observation, therefore only the first variables of those observations are observed. For this incomplete bivariate data, the first 10 observations form one missing-pattern group and the last 10 observations form another group. Therefore, we have $s = 2$, $\mathbf{o}_1 = \{1, 2\}$, $\mathbf{m}_1 = \emptyset$, $\mathbf{o}_2 = \{1\}$, $\mathbf{m}_2 = \{2\}$, and $\mathbf{o}_{1,2} = \{1\}$. The hypothesis testing problem in (3.2) is then equivalent to testing whether the marginal distributions of the first variable in these two

missing-pattern groups are the same. It is clear that they are the same. Therefore, we would not reject H_0 in (3.2). Also based on how we generate this incomplete data, we know the complete joint distributions of both variables before the deletion in these two groups are the same, and therefore the missingness is MCAR in this case.

Now we generate another set of incomplete data as follows. We first take the original $\mathbf{y}_1, ..., \mathbf{y}_{20}$ before the deletion in the above study, add 100 to the second variable in each of the last 10 observations, and other observations remain the same. This way we can view the second variables of the last 10 observations as being drawn from the normal distribution with mean 100 and variance 1. Based on this modified data, we delete the observations from the second variable which are larger than 50. With a very high probability, this will lead to the deletion of the second variables in the last 10 observations. Therefore, we get the same incomplete data as in the previous study. Again, we would not reject H_0 in (3.2), since the marginal distributions of the first variable in the two groups are the same. However, based on the way we make up this incomplete data, the complete joint distributions of both variables before the deletion in these two groups are no longer the same, and the missing mechanism is clearly MNAR.

From the above example, we can see that, even with the same set of incomplete data, if we fail to reject H_0 in (3.2), the complete joint distributions can the same or different across the different missing-pattern groups, and hence the missingness can be MCAR or MNAR, depending on what are the $F_{i,\mathbf{m}_i}$. With no information about the $F_{i,\mathbf{m}_i}$, it is impossible to know which one is the real missing mechanism. The same phenomenon exists for the procedures in Little (1988), Kim and Bentler (2002) and Jamshidian and Jalal (2010). That is, if those procedures fail to reject the null hypothesis, they do not automatically guarantee that the missingness is MCAR. The true missingness can be MCAR or MNAR, depending on the distributions of the missing

data, which unfortunately will not be known. Therefore, the difficulty associated with interpreting our testing result when we fail to reject H_0 in (3.2) is due to the nature of the data. If we fail to reject H_0 in (3.2), at least we can conclude that the missingness can be MCAR in this case.

3.3 The Proposed Non-parametric Test

In this section, we describe the proposed procedure for the hypothesis testing problem in (3.2) without any distributional assumptions on the F_i. Notice that, to test H_0 in (3.2), we need to compare each pair of $(F_{i,\mathbf{o}_{ij}}, F_{j,\mathbf{o}_{ij}})$ for all $\{(i,j) : i = 1, \cdots, s, j = i+1, \cdots, s, \mathbf{o}_{ij} \neq \emptyset\}$. For this purpose, we first introduce a dissimilarity measurement used in Rizzo and Székely (2010) to quantify the difference between any two multivariate random samples. This sample-based dissimilarity measurement can help to identify the difference between their underlying distributions.

Suppose there are two random samples $\{\mathbf{x}_1, ..., \mathbf{x}_{n_1}\}$ and $\{\mathbf{z}_1, ..., \mathbf{z}_{n_2}\}$ in $\mathbb{R}^p$. Define the data matrices based on these two samples as $\mathbb{X} = (\mathbf{x}_1', ..., \mathbf{x}_{n_1}')'$ and $\mathbb{Z} = (\mathbf{z}_1', ..., \mathbf{z}_{n_2}')'$. Then the dissimilarity measurement between the two samples is defined as

$$d(\mathbb{X}, \mathbb{Z}) = 2g(\mathbb{X}, \mathbb{Z}) - g(\mathbb{X}, \mathbb{X}) - g(\mathbb{Z}, \mathbb{Z}),$$

where

$$g(\mathbb{X},\mathbb{Z}) = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\|\mathbf{x}_i - \mathbf{z}_j\|,$$
$$g(\mathbb{X},\mathbb{X}) = \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\|\mathbf{x}_i - \mathbf{x}_j\|,$$
$$g(\mathbb{Z},\mathbb{Z}) = \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}\|\mathbf{z}_i - \mathbf{z}_j\|,$$

and $\|\cdot\|$ denotes the Euclidean norm. The above dissimilarity measurement can be considered as the sample version of the following measure,

$$d(F_X, F_Z) = 2E\|\boldsymbol{X}_1 - \boldsymbol{Z}_1\| - E\|\boldsymbol{X}_1 - \boldsymbol{X}_2\| - E\|\boldsymbol{Z}_1 - \boldsymbol{Z}_2\|,$$

where F_X and F_Z are the underlying distributions of samples $\mathbb{X}$ and $\mathbb{Z}$, respectively, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent random observations drawn from F_X, and $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are independent random observations drawn from F_Z. It is well known that $d(F_X, F_Z) \geq 0$ with equality if and only if $F_X = F_Z$. As a result, if the two samples $\mathbb{X}$ and $\mathbb{Z}$ have the same underlying distributions, we expect that $d(\mathbb{X},\mathbb{Z})$ would be close to 0. On the other hand, if the two underlying distributions are different, we expect that $d(\mathbb{X},\mathbb{Z})$ would be large. And if there is another sample $\tilde{\mathbf{Z}}$, which is obviously far different from $\mathbf{X}$, we would expect $d(\mathbb{X},\tilde{\mathbb{Z}})$ has a even larger value than $d(\mathbb{X},\mathbb{Z})$. Therefore, the dissimilarity measure $d(\mathbb{X},\mathbb{Z})$ can help detect whether the two underlying distributions are the same.

Before we apply this dissimilarity measurement to our missing-pattern groups i and j (i.e., $\mathbb{Y}_i$ and $\mathbb{Y}_j$), we introduce a few more notations. Recall that our data have the same set of missing variables within each missing-pattern group. We define $\mathbb{Y}_{i,\mathbf{o}_i}$ as the sub-matrix of $\mathbb{Y}_i$ consisting only the columns associated with the variables in $\mathbf{o}_i$,

$i = 1, ..., s$, and $\mathbb{Y}_{i,\mathbf{o}_{ij}}$ as the sub-matrix of $\mathbb{Y}_i$ consisting only the columns associated with the variables in $\mathbf{o}_{ij}$. Since $\mathbb{Y}_{i,\mathbf{o}_{ij}}$ and $\mathbb{Y}_{j,\mathbf{o}_{ij}}$ can be considered as the sample drawn from $F_{i,\mathbf{o}_{ij}}$ and $F_{j,\mathbf{o}_{ij}}$, respectively, the above dissimilarity measure when applying to $\mathbb{Y}_{i,\mathbf{o}_{ij}}$ and $\mathbb{Y}_{j,\mathbf{o}_{ij}}$, i.e., $d(\mathbb{Y}_{i,\mathbf{o}_{ij}}, \mathbb{Y}_{j,\mathbf{o}_{ij}})$, can provide information about the difference between $F_{i,\mathbf{o}_{ij}}$ and $F_{j,\mathbf{o}_{ij}}$.

To compare all the possible pairs of $(F_{i,\mathbf{o}_{ij}}, F_{j,\mathbf{o}_{ij}})$, we define the following overall dissimilarity measure between all the s missing-pattern groups,

$$B = \sum_{\substack{1 \le i < j \le s \\ \mathbf{o}_{ij} \neq \emptyset}} (\frac{n_i n_j}{2n}) d(\mathbb{Y}_{i,\mathbf{o}_{ij}}, \mathbb{Y}_{j,\mathbf{o}_{ij}}).$$

If H_0 in (3.2) is true, we would expect that B is close to 0. If H_1 in (3.2) is true, we would expect a large value of B.

The above dissimilarity measurement B is very similar to the between-sample variability measurement used in ANOVA. We can also define the following measure which resembles the within-sample variability measure in ANOVA,

$$W = \sum_{i=1}^{s} n_i g(\mathbb{Y}_{i,\mathbf{o}_i}, \mathbb{Y}_{i,\mathbf{o}_i})/2.$$

Similar to the F statistic for ANOVA, we define the following statistic for testing H_0 in (3.2),

$$F = \frac{B/(s-1)}{W/(n-s)}. \tag{3.3}$$

Intuitively, a larger value of F implies relatively larger between-sample variability compared with the within-sample variability, which indicates that it is more likely for the underlying distributions of the s missing-pattern groups to be different. Therefore, we reject H_0 in (3.2) if $F > c_\alpha$, where c_α is the upper α quantile of the distribution of

F under H_0 in (3.2). Before we describe how to determine c_α, we first study some properties of our proposed test.

Theorem 2. *(Consistency) The above F test for the hypothesis testing problem in (3.2) is statistically consistent against all alternatives with finite second moments.*

The above result implies that our proposed F test is capable of detecting any distributional difference in the observed data among the s missing-pattern groups.

Next we describe in details how to determine c_α, the critical value of our F test. From above, c_α is the upper α-quantile of the null distribution of F. In general, the null distribution of F is not easy to obtain, and so is c_α. To circumvent this difficulty, we resort to the bootstrap method to approximate the null distribution of F. Recall that $\mathbb{Y}_1$ consists of n_1 cases with p variables completely observed. To generate a bootstrap resample under the null hypothesis, we first randomly draw n cases with replacement from the n_1 cases in $\mathbb{Y}_1$ and put them in a n-by-p matrix. We denote this matrix by $\mathbb{Y}^*_{complete}$. We choose the first n_1 rows of $\mathbb{Y}^*_{complete}$ as the bootstrap resample of $\mathbb{Y}_1$. We then choose the next n_2 rows of $\mathbb{Y}^*_{complete}$ and delete the observations for the variables in $\mathbf{m}_2$. This way we obtain a bootstrap resample of $\mathbb{Y}_2$. We continue this procedure, and for any missing pattern group $\mathbb{Y}_i$, we choose their corresponding rows in $\mathbb{Y}^*_{complete}$, delete the observations for the variables in $\mathbf{m}_i$, and obtain the bootstrap resample of $\mathbb{Y}_i$. After we finish all the missing-pattern groups, we are able to obtain a bootstrap resample of the incomplete data matrix $\mathbb{Y}$. We denote it by $\mathbb{Y}^*$. In the above procedure, the way we generate those missing data in $\mathbb{Y}^*$ guarantees that the missingness is MCAR and $F_{i,\mathbf{o}_{ij}} = F_{j,\mathbf{o}_{ij}}$ for all $i \neq j \in \{1, ..., s\}$ and $\mathbf{o}_{ij} \neq \emptyset$. Therefore, our bootstrap resample is generated under H_0 in (3.2). We repeat the above bootstrap procedure B times. Here B is sufficiently large. We denote the bootstrap data matrix $\mathbb{Y}^*$ from the

k-th bootstrap resample by $\mathbb{Y}^{*k}$. For each $\mathbb{Y}^{*k}$, we calculate the F test statistic defined in (3.3) and denoted it by F^{*k}, $k = 1, ..., B$. Let $\hat{c}_\alpha$ be the upper α empirical quantile of $F^{*1}, F^{*2}, ..., F^{*B}$. Then this $\hat{c}_\alpha$ is our estimate of c_α based on our bootstrap method, and we reject H_0 in (3.2) if $F > \hat{c}_\alpha$. Alternatively, we can also calculate the p-value of the F test based on this bootstrap approximation, i.e.,

$$\hat{p} = \sum_{k=1}^{B} I\left\{F^{*k} > F_{obs}\right\}/B, \tag{3.4}$$

where F_{obs} is the value of F based on the original data matrix $\mathbb{Y}$.

3.4 Simulation Study

In this section, we present some simulation studies to demonstrate the performance of our proposed testing procedure. The simulation settings we use are similar to those reported in Jamshidian and Jalal (2010).

3.4.1 Type I Error Study

The first simulation study we conduct is to assess the type I error rates for our proposed F test. The simulation settings consist of the following 64 scenarios from a four-factor design:

(1) The dimension of the data p: we consider two settings $p = 4$ and 10.

(2) The total sample size n: three settings are considered, $n = 200$ and 1000.

(3) The missing percentage q: two settings are considered, $q = 0.35$ and 0.65.

(4) The underlying distributions F: eight distributions will be considered and they are: (i) a standard multivariate normal distribution with mean $\mathbf{0}$ and covariance

I_p (denoted by N); (ii) a correlated multivariate normal distribution with mean $\mathbf{0}$ and covariance Σ (denoted by Corr-N); (iii) a multivariate t distribution with mean $\mathbf{0}$, covariance I_p and degree of freedom 4 (denoted by t); (iv) a correlated multivariate t distribution with mean $\mathbf{0}$, covariance Σ and degree of freedom 4 (denoted by Corr-t); (v) a multivariate uniform distribution which has independent uniform$(0, 1)$ marginal distributions (denoted by U); (vi) a correlated multivariate uniform distribution obtained by multiplying $\Sigma^{\frac{1}{2}}$ to the multivariate uniform distribution in (v) (denoted by Corr-U); (vii) a multivariate distribution obtained by generating $W = Z + 0.1Z^3$, where Z is from the standard multivariate normal distribution (denoted by W); (viii) a multivariate Weibull distribution which has independent Weibull marginal distribution and each Weibull marginal distribution has scale parameter 1 and shape parameter 2 (denoted by Weibull).

Among these eight distributions, t, Corr-t and W are examples of heavy-tailed distributions, while U and Corr-U are examples of light-tailed distributions. Weibull distribution can be treated as an example of skewed distributions. For the above correlated distributions in (ii), (iv) and (vi), we choose $\Sigma = 0.7\mathbf{1}_p\mathbf{1}_p^{'} + 0.3I_p$, where $\mathbf{1}_p$ is a vector of p ones.

Although our proposed F test is for the hypothesis testing problem in (3.2), as mentioned in Section 2, this test remains valid for testing MCAR, i.e., it can still control the type I error at the nominal level for testing MCAR. Therefore, in this section we investigate the type I error rate of our proposed F test when the missingness is MCAR. To this end, for each combination of n, p and F, we first generate the complete n-by-p data matrix $\mathbf{Y}_{complete}$ with each row corresponding to a random observation from F. To generate missing data which are MCAR, we generate another n-by-p matrix $\mathbf{U}$ with

elements $[\mathbf{U}]_{ij}$ being independently drawn from uniform$(0, 1)$. For the given missing percentage q, we compare each element of $\mathbf{U}$, $[\mathbf{U}]_{ij}$, with a threshold, which is chosen so that the percentage of cases with missing values is q. If $[\mathbf{U}]_{ij}$ is less than the threshold, we delete the corresponding element in $\mathbf{Y}_{complete}$. After we finish all the deletion, we obtain the incomplete multivariate data matrix $\mathbf{Y}$. From the above procedure how we generate those missing data, it is clear that the missingness is independent of the data $\mathbf{Y}_{complete}$, and therefore the missingness is MCAR. After we obtain the incomplete multivariate data matrix $\mathbf{Y}$, we apply the proposed F test to $\mathbf{Y}$ and calculate $\hat{p}$ as in (3.4) with the number of bootstrap resamples $B = 499$. If $\hat{p}$ is smaller than the nominal level $\alpha = 0.05$, we reject the null hypothesis. We repeat the above procedure 1000 times and the percentage of times when the null hypothesis is rejected is the simulated type I error rate. Table 3.1 presents the simulated type I error rates for our proposed F test for testing MCAR for each of the 64 settings. As we can see from the table, all the simulated Type I errors are all close to the nominal level 5%, which indicates great performance of our F test under the null hypothesis for testing MCAR.

3.4.2 Power Study

In this subsection, we report two simulation studies to evaluate the power of our proposed non-parametric F test against the other two missing mechanisms, MAR and MNAR. We first study the power of our F test when the missingness is MAR. To generate an incomplete multivariate data with missing data being MAR, we first start with the complete data matrix $\mathbb{Y}_{complete}$ for each combination of n, p and F as in the previous type I error study. Different from the type I error study where the missingness is independent of observed or missing data, we need to make the missingness depend on

Dist.	q	p=4		p=10	
		n=200	n=1000	n=200	n=1000
N	0.35	4.7	3.7	4.1	6.0
	0.65	5.3	5.1	4.9	3.7
Corr-N	0.35	4.7	5.4	5.1	5.3
	0.65	4.0	4.6	5.5	4.3
t	0.35	4.4	4.6	6.2	5.2
	0.65	5.2	5.3	7.2	4.9
Corr-t	0.35	5.1	4.5	5.9	5.5
	0.65	5.0	4.5	4.1	5.8
U	0.35	5.0	4.2	4.7	4.2
	0.65	4.6	5.0	4.1	5.0
Corr-U	0.35	4.5	4.6	5.0	4.9
	0.65	4.3	5.3	4.0	5.9
W	0.35	4.7	5.9	4.9	5.1
	0.65	4.6	5.7	4.9	5.1
Weibull	0.35	4.6	5.0	5.0	4.4
	0.65	4.7	4.9	4.3	4.7

Table 3.1: The type I error rates (%)

the observed data in the MAR case. For this purpose, we first denote the jth variable of the ith case by y_{ij}, and then we keep the first variables of all n cases, $y_{11}, ..., y_{n1}$, as observed. If y_{i1} is larger than a threshold c, then each of the other variables from the same case, $y_{i2}, ..., y_{ip}$, will be independently subject to missing with probability q_1. If y_{i1} is smaller than c, then each of $y_{i2}, ..., y_{ip}$ will be independently subject to missing with probability q_2. Based on this procedure, we obtain the incomplete data matrix $\mathbb{Y}$. From the above, we can see that the missingness only depends on the values of the y_{i1}, which are observed. Therefore, the missingness is MAR. In our simulation, we choose c as the 60 percentile of $y_{11}, ..., y_{n1}$, and q_1 and q_2 are determined so that the percentage of cases with missing values is q.

We apply the proposed F test to the incomplete data matrix $\mathbb{Y}$ and calculate $\hat{p}$ as in (3.4). We repeat this procedure 1000 times and the percentage of times when $\hat{p}$ is smaller than 0.05 is the simulated power of our F test under this particular MAR

Dist.	q	p=4		p=10	
		n=200	n=1000	n=200	n=1000
N	0.35	22.4	96.2	26.2	96.8
	0.65	24.9	94.0	18.6	90.4
Corr-N	0.35	33.4	98.3	64.9	100.0
	0.65	26.3	95.8	27.4	98.0
t	0.35	20.5	93.5	25.9	58.3
	0.65	23.2	94.8	37.3	54.6
Corr-t	0.35	27.6	97.7	43.1	99.4
	0.65	23.2	94.8	37.3	100.0
U	0.35	27.2	96.6	36.4	99.4
	0.65	26.0	95.4	26.1	99.2
Corr-U	0.35	39.6	99.7	87.2	100.0
	0.65	31.4	98.4	75.3	100.0
W	0.35	20.1	94.2	15.8	77.1
	0.65	20.7	92.0	12.8	65.1
Weibull	0.35	27.6	96.8	28.3	98.6
	0.65	27.2	95.4	30.1	97.6

Table 3.2: Power (%) of the F test with MAR alternatives

alternative. Table 3.2 shows the simulated power of our F test for each of the 64 settings.

To study the power of our F test when the missingness is MNAR, we generate the incomplete data matrix $\mathbb{Y}$ with missing data being MNAR as follows. We first generate $\mathbb{Y}$ as that in the type I error study for each combination of n, p and F, and then we replace the first missing-pattern group $\mathbb{Y}_1$ (the one where all the p variables are observed) by a different n_1-by-p matrix $\mathbb{Y}_1^{new}$. In $\mathbb{Y}_1^{new}$, each row is a random observation from G, a different distribution from F. Therefore, the incomplete data matrix $\mathbb{Y}$ we obtain this way has $F_1 = G$ and $F_2 = \cdots = F_s = F$. Since our theorem in Section 3.3 suggests that our F test is consistent against all the alternatives, we choose G to have different location, or different covariance structure, or different distribution form from F. The different F/G settings we considered are summarized in Table 3.3. In the first panel of Table 3.3, the G counterpart of each F distribution (for example, "N1" is the G counterpart of N in the first row) represents the distribution of $\mathbf{y}^{new}$, where $\mathbf{y}^{new}$ is obtained by $\mathbf{y}^{new} = \mathbf{y} + (0.6, 0, ..., 0)'$, and $\mathbf{y}$ follows the distribution F. Therefore, the

first panel consists of the settings where F and G differ in the location. The notations in the second and third panels of Table 3.3 are the same as those in the type I error study. Therefore, it is clear that the second panel of Table 3.3 consists of the settings where F and G have different covariance structures, and that the third panel of Table 3.3 consists of the settings where F and G come from different distribution families. For each $\mathbb{Y}$, we apply our F test and calculate $\hat{p}$ based on (3.4). We repeat this procedure 1000 times and Table 3.3 reports the simulated power of our F test under different alternatives. As we can see from Table 3.2 and Table 3.3, our F test performs very well against all the MAR and MNAR alternatives we consider here. It is not surprising since the observed data in all the settings have different distributions among different missing-pattern groups and our F test is capable of detecting any distributional difference among them. With sample size n increasing, the power of our F test also increases accordingly. When $n = 1000$, the power is approaching 1 in many settings, which further confirms the consistency property of our F test. In the first panel of Table 3.3 where F and G differ in location, some of the power is not very high. This is mainly due to the fact that the difference between F and G we consider here is only a location shift of 0.6 in one out of p variables, which is a very small difference especially for the heavy tailed distributions (t and W) with $p = 10$. If we increase the magnitude of the location shift, the power of our test will increase significantly in those settings.

Dist.	q	p=4		p=10	
F/G		n=200	n=1000	n=200	n=1000
$N/N1$	0.35	37.5	100.0	18.1	74.5
	0.65	35.6	99.6	7.4	47.4
$t/t1$	0.35	18.9	94.9	10.1	17.7
	0.65	18.6	95.9	7.8	9.1
$U/U1$	0.35	100.0	100.0	100.0	100.0
	0.65	100.0	100.0	99.8	100.0
$W/W1$	0.35	20.2	95.6	14.9	25.2
	0.65	22.5	97.6	6.8	14.8
Weibull/Weibull1	0.35	100.0	100.0	85.7	100.0
	0.65	100.0	100.0	54.0	100.0
N/Corr-N	0.35	15.7	96.9	16.5	99.4
	0.65	8.2	94.4	7.2	76.7
Corr-N/N	0.35	17.6	97.3	22.3	82.0
	0.65	16.0	98.0	14.0	59.2
t/Corr-t	0.35	10.2	80.2	11.3	48.1
	0.65	6.5	78.7	11.4	31.6
Corr-t/t	0.35	12.4	77.2	8.9	28.3
	0.65	9.4	88.0	10.6	21.6
U/Corr-U	0.35	100.0	100.0	100.0	100.0
	0.65	100.0	100.0	100.0	100.0
Corr-U/U	0.35	100.0	100.0	100.0	100.0
	0.65	100.0	100.0	100.0	100.0
t/N	0.35	33.9	68.7	79.5	97.8
	0.65	20.1	47.7	72.3	91.7
W/N	0.35	31.4	67.9	91.3	99.6
	0.65	15.5	47.7	77.8	98.3
Weibull/N	0.35	100.0	100.0	100.0	100.0
	0.65	100.0	100.0	100.0	100.0

Table 3.3: Power (%) with MNAR alternatives

Chapter 4

Concluding Remarks

In this dissertation, we propose two approaches: the Bayesian approach and the non-parametric approach, to handle missing data analysis.

The Bayesian project is motivated by a real data, the California NAEP data, which contains the student performance scores. We build the hierarchical linear models to describe the hierarchy feature of the data and the corresponding missing structures. In particular, we construct the pattern mixture model and the selection model for this purpose. Due to the complexity of the model, we propose to use Bayesian methods to fit the proposed models. The Bayesian methods turn out can not only identify the missing mechanism but also provide model-based estimators, which outperform the traditional design-based estimators in terms of bias, mean square error and the converge rate of the true value if the missing is not at random as shown in the simulation study.

During the research, we find out that the ordinary Gibbs sampler methods do not work well for our motivating data example. So for the pattern mixture model, we propose to use the empirical Bayesian method, which essentially utilize the data information to set the threshold for the parameter of the model, to avoid the possible false

convergency issue. On the other hand, to describe the relationship between the outcome values and the corresponding missing indicators in the ordinary selection model, people usually use the logistic regression, where the involving parameters usually have posterior distribution not in a close form. We propose to use the robit regression as an approximation of the logistic regression in the hierarchical linear model so that all the parameters having conjugate priors. By using the robit regression, the consuming time of each parameter updated draw is much shorter compared with the original logistic regression, which needs to fit a logistic regression in each iteration. But the robit regression involves the missing latent variables, which bring a big auto-correlation between draws. Overall, the robit regression approximation is still more efficient than the original logistic regression. We evaluate the MCMC chain of both the pattern mixture model and the selection model by a series of diagnosis plots for the parameters of interest, such as the trajectory plots, the auto-correlation plots and the empirical posterior plots to verify that the MCMC chains converge correctly and we can make valid conclusion based on the MCMC chain we run. Furthermore, we check the fitted models by two methods, the posterior predictive distribution with its summary information method and the residual plots by using the cross validation method. Both the pattern mixture model and the selection model seem to fit our motivating data well under these two criteria. We further investigate three commonly used model selection criteria in the sense of Bayesian data analysis, the Bayes factor method, the deviance information criterion and the minimum posterior predictive loss method namely. We point out that the first two methods have some limitations in our data example and the last one fails to identify the correct model structure under the non-ignorable missing.

In addition, although in this dissertation, we focus on the two-level data with only the student level containing missing values, we can extend our methods to more

general cases. First, our methods can be extended to involve other covariates in either the outcome model or the missing indicator model, or both. The explaining covariates can be gender, race, age and the highest parents' education level. Second, they can be extended to solve the nationwide three-level data with both the school level and the student level containing missingness. For this situation, we use a three-level hierarchical model to describe the data structure, and introduce two different kinds of missing latent variables for the student level and the school level missing separately.

The nonparametric project, as an alternative, is about testing the missing mechanism for multivariate missing data. The test of missing mechanism is very important since most of the statistical procedures we use are built on the assumption of MCAR, such as the estimating equation method. For this kind of test, people divide the data into groups by their missing patterns and the test will be equivalent a homogeneity test among several groups. In literature, there are several tests based on the normality assumption, but normality is not always appropriate. Our test is the first distribution free one. We first develop a nonparametric dispersion measurement for multivariate missing data and used the between and within distance ratio based on this dispersion measurement as the non-parametric test statistics, which is similar to ANOVA ratio. And actually ANOVA ratio can be treated as a special case for the nonparametric ratio. We also prove the consistency of the test. The simulation result shows this nonparametric test has type I error well controlled at the nominal level and good power under various alternatives.

Bibliography

[1] Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**, 669–679.

[2] Amemiya, T. (1984). Tobit models: A survey. *Journal of Econometrics*, **24**, 3–61.

[3] Baker, S. G. and Laird, N. M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, **83**, 62–69.

[4] Box, G. E. P. (1980). Sampling and bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A*, **143**, 383–430.

[5] Celeux, G., Forbes, F., Robert, C. P., and Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, **1**, 651–674.

[6] Chen, H. Y. and Little, R. (1999). A test of missing completely at random from generalised estimating equation with missing data. *Biometrika*, **86**, 1–13.

[7] Chib, S. and Greenberg, E. (1995). Understanding the metropolischastings algorithm. *The American Statistician*, **49**, 327–335.

[8] Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge Univ. Press.

[9] DeIorio, M. and Robert, C. P. (2002). Discussion on the paper by Spiegelhalter et al. *Journal of the Royal Statistical Society, Series B*, **64**, 629–630.

[10] Diggle, P. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics*, **43**, 49–93.

[11] Efron, B. and Tibshirani, R. (1993). *An Introduction to Bootstrap*. Chapman & Hall.

[12] Follmann, D. and Wu, M. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics*, **51**, 151–168.

[13] Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data. *Journal of the American Statistical Association*, **77**, 270–278.

[14] Gelfand, A. E. and Ghosh, S. K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika*, **85**, 1–11.

[15] Gelman, A., Meng, X., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, **6**, 733–807.

[16] Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. CRC Press.

[17] Givens, G. H. and Hoeting, J. A. (2005). *Computational Statistics*. Wiley-Interscience.

[18] Glynn, R. J., Laird, N. M., and Rubin, D. B. (1986). Selection modelling versus mixture modelling with nonignorable nonresponse. Wainer, H. (ed.), *Drawing Inferences from Self-Selected Samples*, pp. 115–142, Springer.

[19] Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

[20] Ibrahim, J. G., Chen, M. H., and Lipsitz, S. R. (2001). Missing responses in generalized linear mixed models when the missing data mechanism is nonignorable. *Biometrika*, **88**, 551–564.

[21] Jamshidian, M. and Jalal, S. (2010). Tests of homoscedasticity, normality and missing completely at random for incomplete multivariate data. *Psychometrika*, **75**, 649–674.

[22] Kim, K. H. and Bentler, P. M. (2002). Tests of homogeneity of means and covariance matrices for multivariate incomplete data. *Psychometrika*, **67**, 609–623.

[23] Lindley, D. V. (1957). A statistical paradox. *Biometrika*, **44**, 187–192.

[24] Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, **83**, 1198–1202.

[25] Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, **88**, 125–134.

[26] Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, **90**, 1112–1121.

[27] Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley-Interscience, 2nd edn.

[28] Little, R. J. A. (2004). To model or not to model? competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, **99**, 546–556.

[29] Liu, J. S. (1994). The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, **89**, 958–966.

[30] Qu, A. and Song, P. X. K. (2002). Testing ignorable missingness in estimating equation approaches for longitudinal data. *Biometrika*, **89**, 841–850.

[31] Raudenbush, S. W. and Bryk, A. S. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage.

[32] Rizzo, M. L. and Székely, G. J. (2010). Disco analysis: A nonparametric extension of analysis of variance. *The Annals of Applied Statistics*, **4**, 1034–1055.

[33] Robert, C. P. (2007). *The Bayesian Choice: from Decision-theoretic Foundations to Computational Implementation*. Springer.

[34] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

[35] Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics*, **12**, 1151–1172.

[36] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. J. Wiley & Sons.

[37] Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.

[38] Scott, A. and Smith, T. M. F. (1969). Estimation in multi-stage surveys. *Journal of the American Statistical Association*, **64**, 830–840.

[39] Shafer, G. (1982). Lindley's paradox. *Journal of the American Statistical Association*, **77**, 325–334.

[40] Smith, A. F. M. and Robers, G. O. (1993). Bayesian computation via the gibs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society, Series B*, **55**, 3–23.

[41] Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 528–540.

[42] Spiegelhalter, D. J., Best, N., Carlin, B., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, **64**, 583–639.

[43] Srivastava, M. S. (2002). *Methods of Multivariate Statistics*. Wiley.

[44] Székely, G. J. and Rizzo, M. L. (2005). A new test for multivariate normality. *Journal of Multivariate Analysis*, **93**, 58–80.

[45] Székely, G. J. and Rizzo, M. L. (2005). Hierarchical clustering via joint between-within distances: Extending wards minimum variance method. *Journal of Classification*, **22**, 151–183.

[46] Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 528–540.

Appendix A

Derivation of the Conditional Posterior Distribution in the Pattern Mixture Model Approach

a. The proof of

$$\begin{aligned}&\left[\theta_{ij} \mid \mathbf{Y}_{obs}, r_{ij}=1, \beta_{0ij}, \alpha, \sigma_2^2, \chi_{ij}\right] \\ \propto & TN_{[\theta_{ij}>0]}\left(\chi_{ij}+\frac{\alpha(y_{ij}-\beta_{0ij}-\alpha\chi_{ij})}{\alpha_2^2+\sigma_2^2}, \frac{\sigma_2^2}{\alpha_2^2+\sigma_2^2}\right)\end{aligned}$$

Since $r_{ij} = 1 \Rightarrow \theta_{ij} > 0$

$$
\begin{aligned}
&P(\theta_{ij} \mid \mathbf{Y}_{obs}, r_{ij} = 1, \beta_{0ij}, \alpha, \sigma_2^2, \chi_{ij}) \\
\propto &P\left(\mathbf{Y}_{obs} \mid \beta_{0ij}, \alpha, \theta_{ij}, \sigma_2^2\right) \times P\left(\theta_{ij} \mid \chi_{ij}\right) / P\left(\theta_{ij} > 0\right) \\
\propto &exp\left\{-\frac{1}{2\sigma_2^2}\left(y_{ij} - \left(\beta_{0ij} + \alpha\theta_{ij}\right)\right)^2\right\} \times exp\left\{-\frac{1}{2}\left(\theta_{ij} - \chi_{ij}\right)^2\right\} / P\left(\theta_{ij} > 0\right) \\
\propto &exp\left\{-\frac{1}{2}\left[\left(\frac{\alpha_2^2}{\sigma_2^2} + 1\right)\theta_{ijk}^2 - 2\left(\frac{\alpha}{\sigma_2^2}\left(y_{ij} - \beta_{0ij}\right) + \chi_{ij}\right)\theta_{ij}\right]\right\} / P\left(\theta_{ij} > 0\right) \\
\propto &TN_{[\theta_{ij}>0]}\left(\frac{\frac{\alpha}{\sigma_2^2}(y_{ij} - \beta_{0ij}) + \chi_{ij}}{\frac{\alpha_2^2}{\sigma_2^2} + 1}, \frac{1}{\frac{\alpha_2^2}{\sigma_2^2} + 1}\right) \\
\propto &TN_{[\theta_{ij}>0]}\left(\chi_{ij} + \frac{\alpha(y_{ij} - \beta_{0ij} - \alpha\chi_{ij})}{\alpha_2^2 + \sigma_2^2}, \frac{\sigma_2^2}{\alpha_2^2 + \sigma_2^2}\right)
\end{aligned}
$$

b. The proof of

$$
\begin{aligned}
&\left[\beta_{0ij} \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, r_{ij} = 1, \alpha, \boldsymbol{\theta}_{ij(k)}, \sigma_2^2, \beta_{0i}, \alpha_1, \theta_{ij}, \sigma_1^2\right] \\
\propto &N\left(\frac{n_{ij}\sigma_1^2\left(\bar{y}_{ij} - \alpha\bar{\theta}_{ij}\right) + \sigma_2^2\left(\beta_{0i} + \alpha_1\theta_{ij}\right)}{n_{ij}\sigma_1^2 + \sigma_2^2}, \frac{\sigma_2^2\sigma_1^2}{n_{ij}\sigma_1^2 + \sigma_2^2}\right)
\end{aligned}
$$

$$
\begin{aligned}
&P\left(\beta_{0ij} \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, r_{ij} = 1, \alpha, \boldsymbol{\theta}_{ij(k)}, \sigma_2^2, \beta_{0i}, \alpha_1, \theta_{ij}, \sigma_1^2\right) \\
\propto &P\left(\mathbf{Y}_{obs}, \mathbf{Y}_{mis} \mid \beta_{0ij}, \alpha, \boldsymbol{\theta}_{ij(k)}, \sigma_2^2\right) \times P\left(\beta_{0ij} \mid \beta_{0i}, \alpha_1, \theta_{ij}, \sigma_1^2\right) \\
\propto &P\left(\bar{y}_{ij} \mid \beta_{0ij}, \alpha, \boldsymbol{\theta}_{ij(k)}, \sigma_2^2\right) \times P\left(\beta_{0ij} \mid \beta_{0i}, \alpha_1, \theta_{ij}, \sigma_1^2\right) \\
\propto &exp\left\{-\frac{n_{ij}}{2\sigma_2^2}\left(\bar{y}_{ij} - \left(\beta_{0ij} + \alpha\bar{\theta}_{ij}\right)\right)^2\right\} \times exp\left\{-\frac{1}{2\sigma_1^2}\left(\beta_{0ij} - \left(\beta_{0i} + \alpha_1\theta_{ij}\right)\right)^2\right\} \\
\propto &exp\left\{-\frac{1}{2}\left[\left(\frac{n_{ij}}{\sigma_2^2} + \frac{1}{\sigma_1^2}\right)\beta_{0ij}^2 - 2\left(\frac{n_{ij}}{\sigma_2^2}\left(\bar{y}_{ij} - \alpha\bar{\theta}_{ij}\right) + \frac{1}{\sigma_1^2}\left(\beta_{0i} + \alpha_1\theta_{ij}\right)\right)\beta_{0ij}\right]\right\} \\
\propto &N\left(\frac{\frac{n_{ij}}{\sigma_2^2}\left(\bar{y}_{ij} - \alpha\bar{\theta}_{ij}\right) + \frac{1}{\sigma_1^2}\left(\beta_{0i} + \alpha_1\theta_{ij}\right)}{\frac{n_{ij}}{\sigma_2^2} + \frac{1}{\sigma_1^2}}, \frac{1}{\frac{n_{ij}}{\sigma_2^2} + \frac{1}{\sigma_1^2}}\right) \\
\propto &N\left(\frac{n_{ij}\sigma_1^2\left(\bar{y}_{ij} - \alpha\bar{\theta}_{ij}\right) + \sigma_2^2\left(\beta_{0i} + \alpha_1\theta_{ij}\right)}{n_{ij}\sigma_1^2 + \sigma_2^2}, \frac{\sigma_2^2\sigma_1^2}{n_{ij}\sigma_1^2 + \sigma_2^2}\right)
\end{aligned}
$$

where $\bar{y}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ij}$; $\bar{\theta}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} \theta_{ij}$;

$\boldsymbol{\theta}_{ij(k)}$ denote the vector of $\theta_{ij}, k = 1, \cdots, n_{ij}$; and

$\left[\bar{y}_{ij} \mid \beta_{0ij}, \alpha, \bar{\theta}_{ij}, \sigma_2^2\right] \sim N\left(\beta_{0ij} + \alpha\bar{\theta}_{ij}, \frac{\sigma_2^2}{n_{ij}}\right)$.

c. The proof of

$$\begin{aligned}
&\left[\theta_{ij} \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, r_{ij} = 1, \beta_{0ij}, \beta_{0i}, \alpha_1, \sigma_1^2, \chi_i\right] \\
\propto &TN_{[\theta_{ij}>0]}\left(\chi_i + \frac{\alpha_1\left(\beta_{0ij} - \beta_{0i} - \alpha_1\chi_i\right)}{\alpha_1^2 + \sigma_1^2}, \frac{\sigma_1^2}{\alpha_1^2 + \sigma_1^2}\right)
\end{aligned}$$

Since $r_{ij} = 1 \Rightarrow \theta_{ij} > 0$

$$\begin{aligned}
&P\left(\theta_{ij} \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, r_{ij} = 1, \beta_{0ij}, \beta_{0i}, \alpha_1, \sigma_1^2, \chi_i\right) \\
\propto &P\left(\beta_{0ij} \mid \beta_{0i}, \alpha_1, \theta_{ij}, \sigma_1^2\right) \times P\left(\theta_{ij} \mid \chi_i\right) / P\left(\theta_{ij} > 0\right) \\
\propto &exp\left\{-\frac{1}{2\sigma_1^2}\left(\beta_{0ij} - \left(\beta_{0i} + \alpha_1\theta_{ij}\right)\right)^2\right\} \times P\left\{-\frac{1}{2}\left(\theta_{ij} - \chi_i\right)^2\right\} / P\left(\theta_{ij} > 0\right) \\
\propto &exp\left\{-\frac{1}{2}\left[\left(\frac{\alpha_1^2}{\sigma_1^2} + 1\right)\theta_{ij}^2 - 2\left(\frac{\alpha_1}{\sigma_1^2}\left(\beta_{0ij} - \beta_{0i}\right) + \chi_i\right)\theta_{ij}\right]\right\} \\
\propto &TN_{[\theta_{ij}>0]}\left(\frac{\frac{\alpha_1}{\sigma_1^2}\left(\beta_{0ij} - \beta_{0i}\right) + \chi_i}{\frac{\alpha_1^2}{\sigma_1^2} + 1}, \frac{1}{\frac{\alpha_1^2}{\sigma_1^2} + 1}\right) \\
\propto &TN_{[\theta_{ij}>0]}\left(\chi_i + \frac{\alpha_1\left(\beta_{0ij} - \beta_{0i} - \alpha_1\chi_i\right)}{\alpha_1^2 + \sigma_1^2}, \frac{\sigma_1^2}{\alpha_1^2 + \sigma_1^2}\right)
\end{aligned}$$

d. The proof of

$$\left[\theta_{ij} \mid \mathbf{Y}_{obs}, r_{ij} = 0, \beta_{0i}, \alpha_1, \sigma_1^2, \chi_i\right] \propto TN_{[\theta_{ij}<0]}\left(\chi_i, 1\right)$$

$$
\begin{aligned}
& P\left(\theta_{ij} \mid \mathbf{Y}_{obs}, r_{ij}=0, \beta_{0i}, \alpha_1, \sigma_1^2, \chi_i\right) \\
\propto & P(\theta_{ij} \mid \chi_i) / P(\theta_{ij}<0) \\
\propto & TN_{[\theta_{ij}<0]}\left(\chi_i, 1\right)
\end{aligned}
$$

e. The proof of

$$
\begin{aligned}
& \left[\mathbf{y}_{ij(k)} \mid r_{ij}=0, \alpha, \boldsymbol{\theta}_{ij(k)}, \sigma_2^2, \beta_{0i}, \alpha_1, \theta_{ij}, \sigma_1^2\right] \\
\propto & N\left(\alpha \boldsymbol{\theta}_{ij(k)}+\left(\beta_{0i}+\alpha_1 \theta_{ij}\right) \mathbf{1}_{ij}, \sigma_2^2 \mathbf{I}_{ij}+\sigma_1^2 \mathbf{J}_{ij}\right)
\end{aligned}
$$

and

$$
\begin{aligned}
& \left[\beta_{0ij} \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, r_{ij}=0, \alpha, \boldsymbol{\theta}_{ij(k)}, \sigma_2^2, \beta_{0i}, \alpha_1, \theta_{ij}, \sigma_1^2\right] \\
\propto & N\left(\frac{n_{ij} \sigma_1^2\left(\bar{y}_{ij}-\alpha \bar{\theta}_{ij}\right)+\sigma_2^2\left(\beta_{0i}+\alpha_1 \theta_{ij}\right)}{n_{ij} \sigma_1^2+\sigma_2^2}, \frac{\sigma_2^2 \sigma_1^2}{n_{ij} \sigma_1^2+\sigma_2^2}\right)
\end{aligned}
$$

where $\mathbf{1}_{ij}=(\overbrace{1, \cdots, 1}^{n_{ij}})$;

$\mathbf{I}_{ij}=diag(\overbrace{1, \cdots, 1}^{n_{ij}})$;

$\mathbf{J}_{ij}=\mathbf{1}_{ij} \times \mathbf{1}_{ij}^{\prime}$;

$\bar{\theta}_{ij}=\frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} \theta_{ij}$; and

$\boldsymbol{\theta}_{ij(k)}$ denote the vector of θ_{ij}, $k=1, \cdots, n_{ij}$;

$\mathbf{y}_{ij(k)}$ denote the vector of y_{ij}, $k=1, \cdots, n_{ij}$.

For those two-level missing data, not only y_{ij}, for $i=1, \cdots, n_{ij}$, but also β_{0ij} are treated as missing. So we need to generate the conditional distribution of both of them at the same time. To solve this problem, I will generate the joint distribution of y_{ij}, $i=1, \cdots, n_{ij}$ and β_{0ij} first, and then integrate β_{0ij} out to get the marginal distribution

of y_{ij}, $i = 1, \cdots, n_{ij}$. This step is based on the fact

$$P(A, B \mid C) = P(A \mid B, C) \times P(B \mid C)$$

Since the joint distribution of y_{ij}, $i = 1, \cdots, n_{ij}$ and β_{0ij} is

$$\begin{aligned}
&P\left(\mathbf{y}_{ij(k)}, \beta_{0ij} \mid r_{ij} = 0, \alpha, \boldsymbol{\theta}_{ij(k)}, \sigma_2^2, \beta_{0i}, \alpha_1, \theta_{ij}, \sigma_1^2\right) \\
\propto &P\left(\mathbf{y}_{ij(k)} \mid \beta_{0ij}, \alpha, \boldsymbol{\theta}_{ij(k)}, \sigma_2^2\right) \times P\left(\beta_{0ij} \mid \beta_{0i}, \alpha_1, \theta_{ij}, \sigma_1^2\right) \\
\propto &exp\left\{-\frac{1}{2\sigma_2^2}\sum_{k=1}^{n_{ij}}(y_{ij} - (\beta_{0ij} + \alpha\theta_{ij}))^2\right\} \times exp\left\{-\frac{1}{2\sigma_1^2}\left(\beta_{0ij} - (\beta_{0i} + \alpha_1\theta_{ij})\right)^2\right\} \\
\propto &exp\left\{-\frac{1}{2}\left[\left(\frac{n_{ij}}{\sigma_2^2} + \frac{1}{\sigma_1^2}\right)\beta_{0ij}^2 - 2\left(\frac{1}{\sigma_2^2}\sum_{k=1}^{n_{ij}}(y_{ij} - \alpha\theta_{ij}) + \frac{1}{\sigma_1^2}\left(\beta_{0i} + \alpha_1\theta_{ij}\right)\right)\beta_{0ij}\right]\right\} \\
\times &exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma_2^2}\sum_{k=1}^{n_{ij}}(y_{ij} - \alpha\theta_{ij})^2\right]\right\} \\
\propto &exp\left\{-\frac{1}{2}\left[\left(\frac{n_{ij}}{\sigma_2^2} + \frac{1}{\sigma_1^2}\right)\left(\beta_{0ij} - \frac{\frac{1}{\sigma_2^2}\sum_{k=1}^{n_{ij}}(y_{ij} - \alpha\theta_{ij}) + \frac{1}{\sigma_1^2}\left(\beta_{0i} + \alpha_1\theta_{ij}\right)}{\frac{n_{ij}}{\sigma_2^2} + \frac{1}{\sigma_1^2}}\right)^2\right]\right\} \\
\times &exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma_2^2}\sum_{k=1}^{n_{ij}}(y_{ij} - \alpha\theta_{ij})^2 - \frac{\left(\frac{1}{\sigma_2^2}\sum_{k=1}^{n_{ij}}(y_{ij} - \alpha\theta_{ij}) + \frac{1}{\sigma_1^2}\left(\beta_{0i} + \alpha_1\theta_{ij}\right)\right)^2}{\frac{n_{ij}}{\sigma_2^2} + \frac{1}{\sigma_1^2}}\right]\right\}
\end{aligned}$$

Let $\tilde{y}_{ijk} = y_{ij} - \alpha\theta_{ij}$, $A_{ij} = \beta_{0i} + \alpha_1\theta_{ij}$. Integrate β_{0ij} out

$$
\begin{aligned}
&P\left(y_{ij}, k = 1, \cdots, n_{ij} \mid r_{ij} = 0, \alpha, \boldsymbol{\theta}_{ij(k)}, \sigma_2^2, \beta_{0i}, \alpha_1, \theta_{ij}, \sigma_1^2\right) \\
\propto & exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma_2^2}\sum_{k=1}^{n_{ij}}\left(y_{ij} - \alpha\theta_{ij}\right)^2 - \frac{\left(\frac{1}{\sigma_2^2}\sum_{k=1}^{n_{ij}}\left(y_{ij} - \alpha\theta_{ij}\right) + \frac{1}{\sigma_1^2}\left(\beta_{0i} + \alpha_1\theta_{ij}\right)\right)^2}{\frac{n_{ij}}{\sigma_2^2} + \frac{1}{\sigma_1^2}}\right]\right\} \\
\propto & exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma_2^2}\sum_{k=1}^{n_{ij}}\tilde{y}_{ijk}^2 - \frac{\left(\frac{1}{\sigma_2^2}\sum_{k=1}^{n_{ij}}\tilde{y}_{ijk} + \frac{1}{\sigma_1^2}A_{ij}\right)^2}{\frac{n_{ij}}{\sigma_2^2} + \frac{1}{\sigma_1^2}}\right]\right\} \\
\propto & exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma_2^2}\sum_{k=1}^{n_{ij}}\tilde{y}_{ijk}^2 - \frac{1}{\frac{n_{ij}}{\sigma_2^2} + \frac{1}{\sigma_1^2}}\left(\frac{1}{\sigma_2^2}\sum_{k=1}^{n_{ij}}\left(\tilde{y}_{ijk} - A_{ij}\right) + \left(\frac{n_{ij}}{\sigma_2^2} + \frac{1}{\sigma_1^2}\right)A_{ij}\right)^2\right]\right\} \\
\propto & exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma_2^2}\sum_{k=1}^{n_{ij}}\tilde{y}_{ijk}^2 - \frac{1}{\frac{n_{ij}}{\sigma_2^2} + \frac{1}{\sigma_1^2}}\left(\frac{1}{\sigma_2^4}\left(\sum_{k=1}^{n_{ij}}\left(\tilde{y}_{ijk} - A_{ij}\right)\right)^2 + \frac{1}{\sigma_2^2}\left(\frac{n_{ij}}{\sigma_2^2} + \frac{1}{\sigma_1^2}\right)2A_{ij}\right.\right.\right. \\
& \left.\left.\left.\cdot\sum_{k=1}^{n_{ij}}\tilde{y}_{ijk}\right)\right]\right\} \\
\propto & exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma_2^2}\left(\sum_{k=1}^{n_{ij}}\tilde{y}_{ijk}^2 - 2A_{ij}\sum_{k=1}^{n_{ij}}\tilde{y}_{ijk}\right) - \frac{\frac{1}{\sigma_2^4}}{\frac{n_{ij}}{\sigma_2^2} + \frac{1}{\sigma_1^2}}\left(\sum_{k=1}^{n_{ij}}\left(\tilde{y}_{ijk} - A_{ij}\right)\right)^2\right]\right\} \\
\propto & exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma_2^2}\sum_{k=1}^{n_{ij}}\left(\tilde{y}_{ijk} - A_{ij}\right)^2 - \frac{\frac{1}{\sigma_2^4}}{\frac{n_{ij}}{\sigma_2^2} + \frac{1}{\sigma_1^2}}\left(\sum_{k=1}^{n_{ij}}\left(\tilde{y}_{ijk} - A_{ij}\right)\right)^2\right]\right\}
\end{aligned}
$$

So

$$
\begin{aligned}
&\left[y_{ij}, k = 1, \cdots, n_{ij} \mid r_{ij} = 0, \alpha, \boldsymbol{\theta}_{ij(k)}, \sigma_2^2, \beta_{0i}, \alpha_1, \theta_{ij}, \sigma_1^2\right] \\
\propto & N\left(\alpha\boldsymbol{\theta}_{ij(k)} + \beta_{0i} + \alpha_1\theta_{ij}, \mathbf{C}_{ij}^{-1}\right)
\end{aligned}
$$

where $\mathbf{C}_{ij} = (c_{ij,st})_{n_{ij}\times n_{ij}}$; and

$$
c_{ij,st} = \begin{cases} d\left(1 - s_{ij}\right) & \text{if } s = t, \\ d\left(-s_{ij}\right) & \text{otherwise.} \end{cases}
$$

where $d = \frac{1}{\sigma_2^2}$, $s_{ij} = \frac{\frac{1}{\sigma_2^2}}{\frac{n_{ij}}{\sigma_2^2}+\frac{1}{\sigma_1^2}}$, then $\mathbf{C}_{ij} = d\left(\mathbf{I}_{ij} - s_{ij}\mathbf{J}_{ij}\right)$. Based on intraclass correlation matrix inverse formula,

$$[\mathbf{I} - s\mathbf{J}]^{-1} = \mathbf{I} + \frac{s}{1-ns}\mathbf{J}$$

we have

$$\begin{aligned}\mathbf{C}_{ij}^{-1} &= d^{-1}\left(\mathbf{I}_{ij} + \frac{s_{ij}}{1-n_{ij}s_{ij}}\mathbf{J}_{ij}\right)\\ &= \sigma_2^2\mathbf{I}_{ij} + \sigma_1^2\mathbf{J}_{ij}\end{aligned}$$

Then, use conditional probability formula, then β_{0ij} can be drawn from

$$\begin{aligned}&\left[\beta_{0ij} \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, r_{ij}=0, \alpha, \boldsymbol{\theta}_{ij(k)}, \sigma_2^2, \beta_{0i}, \alpha_1, \theta_{ij}, \sigma_1^2\right]\\ \propto &\frac{P(\left(\mathbf{y}_{ij(k)}, \beta_{0ij} \mid r_{ij}=0, \alpha, \boldsymbol{\theta}_{ij(k)}, \sigma_2^2, \beta_{0i}, \alpha_1, \theta_{ij}, \sigma_1^2\right))}{P\left(\mathbf{y}_{ij(k)} \mid r_{ij}=0, \alpha, \boldsymbol{\theta}_{ij(k)}, \sigma_2^2, \beta_{0i}, \alpha_1, \theta_{ij}, \sigma_1^2\right)}\\ \propto &N\left(\frac{n_{ij}\sigma_1^2\left(\bar{y}_{ij} - \alpha\bar{\theta}_{ij}\right) + \sigma_2^2\left(\beta_{0i} + \alpha_1\theta_{ij}\right)}{n_{ij}\sigma_1^2 + \sigma_2^2}, \frac{\sigma_2^2\sigma_1^2}{n_{ij}\sigma_1^2 + \sigma_2^2}\right)\end{aligned}$$

This is the same distribution as at most one-level missing case, i.e. $r_{ij} = 1$.

f. The proof of

$$\begin{aligned}&\left[\sigma_2^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \beta_{0(ij)}, \boldsymbol{\theta}_{(ijk)}\right]\\ \propto &IG\left(a_2 + \frac{1}{2}\left(\sum_{i=1}^{b}\sum_{j=1}^{b_i} n_{ij} - 1\right), b_2 + \frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{b_i}\sum_{k=1}^{n_{ij}}(y_{ij} - (\beta_{0ij} + \hat{\alpha}_2\theta_{ij}))^2\right)\end{aligned}$$

and

$$\left[\alpha \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\beta}_{0(ij)}, \boldsymbol{\theta}_{(ijk)}, \sigma_2^2\right] \propto N\left(\hat{\alpha}_2, \frac{\sigma_2^2}{\sum_{i=1}^{b}\sum_{j=1}^{b_i}\sum_{k=1}^{n_{ij}}\theta_{ijk}^2}\right)$$

where

$$\hat{\alpha}_2 = \frac{\sum_{i=1}^{b}\sum_{j=1}^{b_i}\sum_{k=1}^{n_{ij}}\theta_{ij}\left(y_{ij} - \beta_{0ij}\right)}{\sum_{i=1}^{b}\sum_{j=1}^{b_i}\sum_{k=1}^{n_{ij}}\theta_{ijk}^2}$$

and

$\beta_{0(ij)}$ denote the vector of β_{0ij}, for $i = 1, \cdots, b, j = 1, \cdots, b_i$

$\boldsymbol{\theta}_{(ijk)}$ denote the vector of θ_{ij}, for $i = 1, \cdots, b, j = 1, \cdots, b_i, k = 1, \cdots, n_{ij}$.

Since the joint distribution of α and σ_2^2 is

$$P\left(\alpha, \sigma_2^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\beta}_{0(ij)}, \boldsymbol{\theta}_{(ijk)}\right)$$

$$\propto P\left(\mathbf{Y}_{obs}, \mathbf{Y}_{mis} \mid \boldsymbol{\beta}_{0(ij)}, \alpha, \boldsymbol{\theta}_{(ijk)}, \sigma_2^2\right) \times P\left(\alpha, \sigma_2^2\right)$$

$$\propto \left(\sigma_2^2\right)^{-\frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{b_i} n_{ij}} exp\left\{-\frac{1}{2\sigma_2^2}\sum_{i=1}^{b}\sum_{j=1}^{b_i}\sum_{k=1}^{n_{ij}}(y_{ij} - (\beta_{0ij} + \alpha\theta_{ij}))^2\right\} \times \left(\sigma_2^2\right)^{-(a_2+1)} exp\left\{-\frac{b_2}{\sigma_2^2}\right\}$$

$$\propto \left(\sigma_2^2\right)^{-\left(a_2+\frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{b_i} n_{ij}+1\right)} exp\left\{-\frac{1}{\sigma_2^2}\left[b_2 + \frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{b_i}\sum_{k=1}^{n_{ij}}(y_{ij} - (\beta_{0ij} + \alpha\theta_{ij}))^2\right]\right\}$$

Then integrating α out to get the marginal distribution of σ_2^2

$$
\begin{aligned}
& P\left(\sigma_2^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \beta_{0(ij)}, \boldsymbol{\theta}_{(ijk)}\right) \\
& \propto \left(\sigma_2^2\right)^{-\left(a_2+\frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{b_i} n_{ij}+1\right)} exp\left\{-\frac{b_2}{\sigma_2^2}\right\} \\
& \times \int_{-\infty}^{+\infty} exp\left\{-\frac{1}{2\sigma_2^2}\left[\left(\sum_{i=1}^{b}\sum_{j=1}^{b_i}\sum_{k=1}^{n_{ij}}\theta_{ijk}^2\right)\alpha_2^2 - 2\left(\sum_{i=1}^{b}\sum_{j=1}^{b_i}\sum_{k=1}^{n_{ij}}\theta_{ij}\left(y_{ij}-\beta_{0ij}\right)\right)\alpha\right]\right\} d\alpha \\
& \times exp\left\{-\frac{1}{2\sigma_2^2}\left[\sum_{i=1}^{b}\sum_{j=1}^{b_i}\sum_{k=1}^{n_{ij}}\left(y_{ij}-\beta_{0ij}\right)^2\right]\right\} \\
& \propto \left(\sigma_2^2\right)^{-\left(a_2+\frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{b_i} n_{ij}+1\right)} exp\left\{-\frac{b_2}{\sigma_2^2}\right\} \\
& \times \int_{-\infty}^{+\infty} exp\left\{-\frac{1}{2\sigma_2^2}\left(\sum_{i=1}^{b}\sum_{j=1}^{b_i}\sum_{k=1}^{n_{ij}}\theta_{ijk}^2\right)\left(\alpha - \frac{\sum_{i=1}^{b}\sum_{j=1}^{b_i}\sum_{k=1}^{n_{ij}}\theta_{ij}\left(y_{ij}-\beta_{0ij}\right)}{\sum_{i=1}^{b}\sum_{j=1}^{b_i}\sum_{k=1}^{n_{ij}}\theta_{ijk}^2}\right)^2\right\} d\alpha \\
& \times exp\left\{-\frac{1}{2\sigma_2^2}\left[\sum_{i=1}^{b}\sum_{j=1}^{b_i}\sum_{k=1}^{n_{ij}}\left(y_{ij}-\beta_{0ij}\right)^2 - \frac{\left(\sum_{i=1}^{b}\sum_{j=1}^{b_i}\sum_{k=1}^{n_{ij}}\theta_{ij}\left(y_{ij}-\beta_{0ij}\right)\right)^2}{\sum_{i=1}^{b}\sum_{j=1}^{b_i}\sum_{k=1}^{n_{ij}}\theta_{ijk}^2}\right]\right\} \\
& \propto \left(\sigma_2^2\right)^{-\left(a_2+\left(\frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{b_i} n_{ij}+1\right)\right)} exp\left\{-\frac{b_2}{\sigma_2^2}\right\} \\
& \times exp\left\{-\frac{1}{2\sigma_2^2}\left[\sum_{i=1}^{b}\sum_{j=1}^{b_i}\sum_{k=1}^{n_{ij}}\left(y_{ij}-\beta_{0ij}\right)\left(y_{ij}-\beta_{0ij}-\hat{\alpha}_2\theta_{ij}\right)\right]\right\} \\
& \propto \left(\sigma_2^2\right)^{-\left(a_2+\left(\frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{b_i} n_{ij}+1\right)\right)} exp\left\{-\frac{1}{\sigma_2^2}\left[b_2 + \frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{b_i}\sum_{k=1}^{n_{ij}}\left(y_{ij}-\beta_{0ij}-\hat{\alpha}_2\theta_{ij}\right)^2\right]\right\} \\
& \propto IG\left(a_2 + \frac{1}{2}\left(\sum_{i=1}^{b}\sum_{j=1}^{b_i} n_{ij} - 1\right), b_2 + \frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{b_i}\sum_{k=1}^{n_{ij}}\left(y_{ij}-\left(\beta_{0ij}+\hat{\alpha}_2\theta_{ij}\right)\right)^2\right)
\end{aligned}
$$

Then the conditional probability of α is

$$
\begin{aligned}
& P\left(\alpha \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \beta_{0(ij)}, \boldsymbol{\theta}_{(ijk)}, \sigma_2^2\right) \\
=& \frac{P\left(\alpha, \sigma_2^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \beta_{0(ij)}, , \boldsymbol{\theta}_{(ijk)}\right)}{P\left(\sigma_2^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \beta_{0(ij)}, , \boldsymbol{\theta}_{(ijk)}\right)} \\
\propto & \left(\sigma_2^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{\sigma_2^2} \sum_{i=1}^{b} \sum_{j=1}^{b_i} \sum_{k=1}^{n_{ij}} \left[\left(y_{ij} - \beta_{0ij} - \alpha\theta_{ij}\right)^2 - \left(y_{ij} - \beta_{0ij} - \hat{\alpha}_2\theta_{ij}\right)^2\right]\right\} \\
\propto & \left(\sigma_2^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{\sigma_2^2} \left(\sum_{i=1}^{b} \sum_{j=1}^{b_i} \sum_{k=1}^{n_{ij}} \theta_{ijk}^2\right) \left(\alpha - \frac{\sum_{i=1}^{b} \sum_{j=1}^{b_i} \sum_{k=1}^{n_{ij}} \theta_{ij}\left(y_{ij} - \beta_{0ij}\right)}{\sum_{i=1}^{b} \sum_{j=1}^{b_i} \sum_{k=1}^{n_{ij}} \theta_{ijk}^2}\right)^2\right\} \\
\propto & N\left(\hat{\alpha}_2, \frac{\sigma_2^2}{\sum_{i=1}^{b} \sum_{j=1}^{b_i} \sum_{k=1}^{n_{ij}} \theta_{ijk}^2}\right)
\end{aligned}
$$

g. The proof of

$$
\left[\chi_{ij} \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\theta}_{ij(k)}, \chi_{i0}, \omega^2\right] \propto N\left(\frac{\omega^2 n_{ij} \bar{\theta}_{ij} + \chi_{i0}}{\omega^2 n_{ij} + 1}, \frac{\omega^2}{\omega^2 n_{ij} + 1}\right)
$$

where

$$
\bar{\theta}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} \theta_{ij}
$$

and

$$
\left[\bar{\theta}_{ij} \mid \chi_{ij}\right] \sim N\left(\chi_{ij}, \frac{1}{n_{ij}}\right)
$$

$$
\begin{aligned}
& P\left(\chi_{ij} \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\theta}_{ij(k)}, \chi_{i0}, \omega^2\right) \\
\propto & P\left(\boldsymbol{\theta}_{ij(k)} \mid \chi_{ij}\right) \times P\left(\chi_{ij} \mid \chi_{i0}, \omega^2\right) \\
\propto & P\left(\bar{\theta}_{ij} \mid \chi_{ij}\right) \times P\left(\chi_{ij} \mid \chi_{i0}, \omega^2\right) \\
\propto & exp\left\{-\frac{n_{ij}}{2}\left(\bar{\theta}_{ij}-\chi_{ij}\right)^2\right\} \times exp\left\{-\frac{1}{2\omega^2}\left(\chi_{ij}-\chi_{i0}\right)^2\right\} \\
\propto & exp\left\{-\frac{1}{2}\left[\left(n_{ij}+\frac{1}{\omega^2}\right)\chi_{ij}^2-2\left(n_{ij}\bar{\theta}_{ij}+\frac{\chi_{i0}}{\omega^2}\right)\chi_{ij}\right]\right\} \\
\propto & N\left(\frac{n_{ij}\bar{\theta}_{ij}+\frac{\chi_{i0}}{\omega^2}}{n_{ij}+\frac{1}{\omega^2}}, \frac{1}{n_{ij}+\frac{1}{\omega^2}}\right) \\
\propto & N\left(\frac{\omega^2 n_{ij}\bar{\theta}_{ij}+\chi_{i0}}{\omega^2 n_{ij}+1}, \frac{\omega^2}{\omega^2 n_{ij}+1}\right)
\end{aligned}
$$

h. The proof of

$$
\left[\chi_{i0} \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\chi}_{i(j)}, \omega^2, \chi, \omega_0^2\right] \propto N\left(\frac{\omega_0^2 b_i \bar{\chi}_i+\omega^2\chi}{\omega_0^2 b_i+\omega^2}, \frac{\omega_0^2\omega^2}{\omega_0^2 b_i+\omega^2}\right)
$$

where

$$
\boldsymbol{\chi}_{i(j)} \text{ denote the vector of } \chi_{ij}, \text{ for } j=1,\cdots,b_i,
$$

and

$$
\begin{aligned}
\bar{\chi}_i &= \frac{1}{b_i}\sum_{j=1}^{b_i}\chi_{ij} \\
[\bar{\chi}_i \mid \chi_{i0}, \omega_i] &\sim N(\chi_{i0}, \frac{\omega^2}{b_i})
\end{aligned}
$$

$$
\begin{aligned}
& P\left(\chi_{i0} \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\chi}_{i(j)}, \omega^2, \chi, \omega_0^2\right) \\
\propto & P\left(\boldsymbol{\chi}_{i(j)} \mid \chi_{i0}, \omega^2\right) \times P\left(\chi_{i0} \mid \chi, \omega_0^2\right) \\
\propto & P\left(\bar{\chi}_i \mid \chi_{i0}, \omega^2\right) \times P\left(\chi_{i0} \mid \chi, \omega_0^2\right) \\
\propto & exp\left\{-\frac{b_i}{2\omega^2}\left(\bar{\chi}_i - \chi_{i0}\right)^2\right\} \times exp\left\{-\frac{1}{2\omega_0^2}\left(\chi_{i0} - \chi\right)^2\right\} \\
\propto & exp\left\{-\frac{1}{2}\left[\left(\frac{b_i}{\omega^2} + \frac{1}{\omega_0^2}\right)\chi_{i0}^2 - 2\left(\frac{b_i}{\omega^2}\bar{\chi}_i + \frac{1}{\omega_0^2}\chi\right)\chi_{i0}\right]\right\} \\
\propto & N\left(\frac{\frac{b_i}{\omega^2}\bar{\chi}_i + \frac{1}{\omega_0^2}\chi}{\frac{b_i}{\omega^2} + \frac{1}{\omega_0^2}}, \frac{1}{\frac{b_i}{\omega^2} + \frac{1}{\omega_0^2}}\right) \\
\propto & N\left(\frac{\omega_0^2 b_i \bar{\chi}_i + \omega^2 \chi}{\omega_0^2 b_i + \omega^2}, \frac{\omega_0^2 \omega^2}{\omega_0^2 b_i + \omega^2}\right)
\end{aligned}
$$

i. The proof of

$$
\left[\omega^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\chi}_{(ij)}, \boldsymbol{\chi}_{(i)0}\right] \propto IG\left(a_3 + \frac{\sum_{i=1}^{b} b_i}{2}, b_3 + \frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{b_i}\left(\chi_{ij} - \chi_{i0}\right)^2\right)
$$

Where $\boldsymbol{\chi}_{(ij)}$ denote the vector of χ_{ij}, for $i = 1, \cdots, b$, $j = 1, \cdots, b_i$.

$\boldsymbol{\chi}_{(i)0}$ denote the vector of χ_{i0}, for $i = 1, \cdots, b$.

$$
\begin{aligned}
& P\left(\omega^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\chi}_{(ij)}, \boldsymbol{\chi}_{(i)0}\right) \\
\propto & P\left(\boldsymbol{\chi}_{(ij)} \mid \boldsymbol{\chi}_{(i)0}, \omega^2\right) \times P\left(\omega^2\right) \\
\propto & \left(\omega^2\right)^{-\frac{\sum_{i=1}^{b} b_i}{2}} exp\left\{-\frac{1}{2\omega^2}\sum_{i=1}^{b}\sum_{j=1}^{b_i}(\chi_{ij} - \chi_{i0})^2\right\} \times \left(\omega^2\right)^{-(a_3+1)} exp\left\{-\frac{b_3}{\omega^2}\right\} \\
\propto & \left(\omega^2\right)^{-\left(a_3 + \frac{\sum_{i=1}^{b} b_i}{2} + 1\right)} exp\left\{-\frac{1}{\omega^2}\left[b_3 + \frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{b_i}(\chi_{ij} - \chi_{i0})^2\right]\right\} \\
\propto & IG\left(a_3 + \frac{\sum_{i=1}^{b} b_i}{2}, b_{3i} + \frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{b_i}\left(\chi_{ij} - \chi_{i0}\right)^2\right)
\end{aligned}
$$

j. The proof of

$$\left[\omega_0^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\chi}_{(i)0}\right] \propto IG\left(a_3 + \frac{b-1}{2}, b_3 + \frac{1}{2}\sum_{i=1}^{b}\left(\chi_{i0} - \bar{\chi}_0\right)^2\right)$$

and

$$\left[\chi \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\chi}_{(i)0}, \omega_0^2\right] \propto N\left(\bar{\chi}_0, \frac{\omega_0^2}{b}\right)$$

where

$$\boldsymbol{\chi}_{(i)0} \text{ denote the vector of } \chi_{i0}, \text{ for } i = 1, \cdots, b,$$
$$\bar{\chi}_0 = \frac{1}{b}\sum_{i=1}^{b}\chi_{i0}$$

The joint distribution of χ and ω_0^2 is

$$\begin{aligned}
&P\left(\chi, \omega_0^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\chi}_{(i)0}\right) \\
\propto &P\left(\chi_{i0} \mid \chi, \omega_0^2\right) \times P\left(\chi, \omega_0^2\right) \\
\propto &\left(\omega_0^2\right)^{-\frac{b}{2}} exp\left\{-\frac{1}{2\omega_0^2}\sum_{i=1}^{b}\left(\chi_{i0} - \chi\right)^2\right\} \times \left(\omega_0^2\right)^{-(a_3+1)} exp\left\{-\frac{b_3}{\omega_0^2}\right\} \\
\propto &\left(\omega_0^2\right)^{-(a_3+\frac{b}{2}+1)} exp\left\{-\frac{1}{\omega_0^2}\left[b_3 + \frac{1}{2}\sum_{i=1}^{b}\left(\chi_{i0} - \chi\right)^2\right]\right\}
\end{aligned}$$

Integrating χ out, we will get the marginal distribution of ω_0^2

$$
\begin{aligned}
& P\left(\omega_0^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\chi}_{(i)0}\right) \\
\propto & \left(\omega_0^2\right)^{-(a_3+\frac{b}{2}+1)} exp\left\{-\frac{b_3}{\sigma_0^2}\right\} \times \int_{-\infty}^{+\infty} exp\left\{-\frac{1}{2\omega_0^2}\sum_{i=1}^{b}(\chi_{i0}-\chi)^2\right\} d\chi \\
\propto & \left(\omega_0^2\right)^{-(a_3+\frac{b}{2}+1)} exp\left\{-\frac{b_3}{\sigma_0^2}\right\} \times \int_{-\infty}^{+\infty} exp\left\{-\frac{b}{2\omega_0^2}\left(\chi-\bar{\chi}_0\right)^2\right\} d\chi \\
\times & exp\left\{-\frac{1}{2\omega_0^2}\sum_{i=1}^{b}\left(\chi_{i0}-\bar{\chi}_0\right)^2\right\} d\chi \\
\propto & \left(\omega_0^2\right)^{-\left(a_3+\frac{b+1}{2}\right)} exp\left\{-\frac{1}{\sigma_0^2}\left(b_3+\frac{1}{2}\sum_{i=1}^{b}\left(\chi_{i0}-\bar{\chi}_0\right)^2\right)\right\} \\
\propto & IG\left(a_3+\frac{b-1}{2}, b_3+\frac{1}{2}\sum_{i=1}^{b}\left(\chi_{i0}-\bar{\chi}_0\right)^2\right)
\end{aligned}
$$

Then the conditional distribution of χ is

$$
\begin{aligned}
& P\left(\chi \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\chi}_{(i)0}, \omega_0^2\right) \\
= & \frac{P\left(\chi, \omega_0^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\chi}_{(i)0}\right)}{P\left(\omega_0^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\chi}_{(i)0}\right)} \\
\propto & \left(\omega_0^2\right)^{\frac{1}{2}} exp\left\{-\frac{1}{2\omega_0^2}\sum_{i=1}^{b}\left[\left(\chi_{i0}-\chi\right)^2-\left(\chi_{i0}-\bar{\chi}_0\right)^2\right]\right\} \\
\propto & \left(\omega_0^2\right)^{\frac{1}{2}} exp\left\{-\frac{1}{2\omega_0^2} b\left(\chi-\bar{\chi}_0\right)^2\right\} \\
\propto & N\left(\bar{\chi}_0, \frac{\omega_0^2}{b}\right)
\end{aligned}
$$

k. The proof of

$$
\begin{aligned}
& \left[\beta_{0i} \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\beta}_{0i(j)}, \alpha_1, \boldsymbol{\theta}_{i(j)}, \sigma_1^2, \beta_0, \sigma^2\right] \\
\propto & N\left(\frac{b_i\sigma^2\left(\bar{\beta}_{0i}-\alpha_1\bar{\theta}_i\right)+\sigma_1^2\beta_0}{b_i\sigma^2+\sigma_1^2}, \frac{\sigma^2\sigma_1^2}{b_i\sigma^2+\sigma_1^2}\right)
\end{aligned}
$$

where

$$\bar{\theta}_i = \frac{1}{b_i}\sum_{j=1}^{b_i}\theta_{ij}$$

and

$$\bar{\beta}_{0i} = \frac{1}{b_i}\sum_{j=1}^{b_i}\beta_{0ij}$$

$$\left[\bar{\beta}_{0i} \mid \beta_{0i}, \alpha_1, \sigma_1^2\right] \sim N\left(\beta_{0i} + \alpha_1\bar{\theta}_i, \frac{\sigma_1^2}{b_i}\right).$$

$$\begin{aligned}
&P\left(\beta_{0i} \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\beta}_{0i(j)}, \alpha_1, \boldsymbol{\theta}_{i(j)}, \sigma_1^2\right) \\
\propto &P\left(\beta_{0i(j)} \mid \beta_{0i}, \alpha_1, \boldsymbol{\theta}_{i(j)}, \sigma_1^2\right) \times P\left(\beta_{0i} \mid \beta_0, \sigma^2\right) \\
\propto &P\left(\bar{\beta}_{0i} \mid \beta_{0i}, \alpha_1, \bar{\theta}_i, \sigma_1^2\right) \times P\left(\beta_{0i} \mid \beta_0, \sigma^2\right) \\
\propto &exp\left\{-\frac{b_i}{2\sigma_1^2}\left(\bar{\beta}_{0i} - \left(\beta_{0i} + \alpha_1\bar{\theta}_i\right)\right)^2\right\} \times exp\left\{-\frac{1}{2\sigma^2}\left(\beta_{0i} - \beta_0\right)^2\right\} \\
\propto &exp\left\{-\frac{1}{2}\left[\left(\frac{b_i}{\sigma_1^2} + \frac{1}{\sigma^2}\right)\beta_{0i}^2 - 2\left(\frac{b_i}{\sigma_1^2}\left(\bar{\beta}_{0i} - \alpha_1\bar{\theta}_i\right) + \frac{1}{\sigma^2}\beta_0\right)\beta_{0i}\right]\right\} \\
\propto &N\left(\frac{\frac{b_i}{\sigma_1^2}\left(\bar{\beta}_{0i} - \alpha_1\bar{\theta}_i\right) + \frac{1}{\sigma^2}\beta_0}{\frac{b_i}{\sigma_1^2} + \frac{1}{\sigma^2}}, \frac{1}{\frac{b_i}{\sigma_1^2} + \frac{1}{\sigma^2}}\right) \\
\propto &N\left(\frac{b_i\sigma^2\left(\bar{\beta}_{0i} - \alpha_1\bar{\theta}_i\right) + \sigma_1^2\beta_0}{b_i\sigma^2 + \sigma_1^2}, \frac{\sigma^2\sigma_1^2}{b_i\sigma^2 + \sigma_1^2}\right)
\end{aligned}$$

l. The proof of

$$\begin{aligned}
&\left[\sigma_1^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\beta}_{0(ij)}, \boldsymbol{\beta}_{0(i)}, \boldsymbol{\theta}_{(ij)}\right] \\
\propto &IG\left(a_{11} + \frac{1}{2}\left(\sum_{i=1}^{b} b_i - 1\right), b_{11} + \frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{b_i}\left(\beta_{0ij} - \left(\beta_{0i} + \hat{\alpha}_1\theta_{ij}\right)\right)^2\right)
\end{aligned}$$

and

$$\left[\alpha_1 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\beta}_{0(ij)}, \boldsymbol{\beta}_{0(i)}, \boldsymbol{\theta}_{(ij)}, \sigma_1^2\right] \propto N\left(\hat{\alpha}_1, \frac{\sigma_1^2}{\sum_{i=1}^{b}\sum_{j=1}^{b_i}\theta_{ij}^2}\right)$$

where

$$\hat{\alpha}_1 = \frac{\sum_{i=1}^{b}\sum_{j=1}^{b_i}\theta_{ij}\left(\beta_{0ij} - \beta_{0i}\right)}{\sum_{i=1}^{b}\sum_{j=1}^{b_i}\theta_{ij}^2}$$

and

$\beta_{0(ij)}$ denote the vector of β_{0ij}, for $i = 1, \cdots, b, j = 1, \cdots, b_i$.

$\beta_{0(i)}$ denote the vector of β_{0i}, for $i = 1, \cdots, b$.

$\boldsymbol{\theta}_{(ij)}$ denote the vector of θ_{ij}, for $i = 1, \cdots, b, j = 1, \cdots, b_i$.

Since the joint distribution of α_1 and σ_1^2 is

$$\begin{aligned}
&P\left(\alpha_1, \sigma_1^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\beta}_{0(ij)}, \boldsymbol{\beta}_{0(i)}, \boldsymbol{\theta}_{(ij)}\right) \\
\propto &P\left(\boldsymbol{\beta}_{0(ij)} \mid \boldsymbol{\beta}_{0(i)}, \alpha_1, \boldsymbol{\theta}_{(ij)}, \sigma_1^2\right) \times P\left(\alpha_1, \sigma_1^2\right) \\
\propto &\left(\sigma_1^2\right)^{-\frac{\sum_{i=1}^{b} b_i}{2}} exp\left\{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{b}\sum_{j=1}^{b_i}\left(\beta_{0ij} - \left(\beta_{0i} + \alpha_1\theta_{ij}\right)\right)^2\right\} \times \left(\sigma_1^2\right)^{-(a_{11}+1)} exp\left\{-\frac{b_{11}}{\sigma_1^2}\right\} \\
\propto &\left(\sigma_1^2\right)^{-\left(a_{11}+\frac{1}{2}\sum_{i=1}^{b} b_i + 1\right)} exp\left\{-\frac{1}{\sigma_1^2}\left[b_{11} + \frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{b_i}\left(\beta_{0ij} - \left(\beta_{0i} + \alpha_1\theta_{ij}\right)\right)^2\right]\right\}
\end{aligned}$$

Then integrating α_1 out to get the marginal distribution of σ_1^2

$$
\begin{aligned}
&P\left(\sigma_1^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\beta}_{0(ij)}, \boldsymbol{\beta}_{0(i)}, \boldsymbol{\theta}_{(ij)}\right) \\
\propto &\int_{-\infty}^{+\infty} exp\left\{-\frac{1}{2\sigma_1^2}\left[\left(\sum_{i=1}^{b}\sum_{j=1}^{b_i}\theta_{ij}^2\right)\alpha_1^2 - 2\left(\sum_{i=1}^{b}\sum_{j=1}^{b_i}\theta_{ij}\left(\beta_{0ij}-\beta_{0i}\right)\right)\alpha_1\right]\right\} d\alpha_1 \\
\times &\left(\sigma_1^2\right)^{-\left(a_{11}+\frac{1}{2}\sum_{i=1}^{b} b_i+1\right)} exp\left\{-\frac{b_{11}}{\sigma_1^2}\right\} exp\left\{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{b}\sum_{j=1}^{b_i}\left(\beta_{0ij}-\beta_{0i}\right)^2\right\} \\
\propto &\int_{-\infty}^{+\infty} exp\left\{-\frac{1}{2\sigma_1^2}\left(\sum_{i=1}^{b}\sum_{j=1}^{b_i}\theta_{ij}^2\right)\left(\alpha_1 - \frac{\sum_{i=1}^{b}\sum_{j=1}^{b_i}\theta_{ij}\left(\beta_{0ij}-\beta_{0i}\right)}{\sum_{i=1}^{b}\sum_{j=1}^{b_i}\theta_{ij}^2}\right)^2\right\} d\alpha_1 \\
\times &\left(\sigma_1^2\right)^{-\left(a_{11}+\frac{1}{2}\sum_{i=1}^{b} b_i+1\right)} exp\left\{-\frac{b_{11}}{\sigma_1^2}\right\} exp\left\{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{b}\sum_{j=1}^{b_i}\left(\beta_{0ij}-\beta_{0i}\right)^2\right\} \\
\times &exp\left\{+\frac{1}{2\sigma_1^2}\frac{\left(\sum_{i=1}^{b}\sum_{j=1}^{b_i}\theta_{ij}\left(\beta_{0ij}-\beta_{0i}\right)\right)^2}{\sum_{i=1}^{b}\sum_{j=1}^{b_i}\theta_{ij}^2}\right\} \\
\propto &\left(\sigma_1^2\right)^{-\left(a_{11}+\frac{1}{2}\left(\sum_{i=1}^{b} b_i+1\right)\right)} exp\left\{-\frac{b_{11}}{\sigma_1^2}\right\} \\
\times &exp\left\{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{b}\sum_{j=1}^{b_i}\left(\beta_{0ij}-\beta_{0i}\right)\left(\beta_{0ij}-\left(\beta_{0i}+\hat{\alpha}_1\theta_{ij}\right)\right)\right\} \\
\propto &\left(\sigma_1^2\right)^{-\left(a_{11}+\frac{1}{2}\left(\sum_{i=1}^{b} b_i+1\right)\right)} exp\left\{-\frac{1}{\sigma_1^2}\left[b_{11}+\frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{b_i}\left(\beta_{0ij}-\left(\beta_{0i}+\hat{\alpha}_1\theta_{ij}\right)\right)^2\right]\right\} \\
\propto &IG\left(a_{11}+\frac{1}{2}\left(\sum_{i=1}^{b} b_i - 1\right), b_{11}+\frac{1}{2}\sum_{i=1}^{b}\sum_{j=1}^{b_i}\left(\beta_{0ij}-\left(\beta_{0i}+\hat{\alpha}_1\theta_{ij}\right)\right)^2\right)
\end{aligned}
$$

Then the conditional probability of α_1

$$
\begin{aligned}
&P\left(\alpha_1 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\beta}_{0(ij)}, \boldsymbol{\beta}_{0(i)}, \boldsymbol{\theta}_{(ij)}, \sigma_1^2\right) \\
= &\frac{P\left(\alpha_1, \sigma_1^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\beta}_{0(ij)}, \boldsymbol{\beta}_{0(i)}, \boldsymbol{\theta}_{(ij)}\right)}{P\left(\sigma_1^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\beta}_{0(ij)}, \boldsymbol{\beta}_{0(i)}, \boldsymbol{\theta}_{(ij)}\right)} \\
\propto &\left(\sigma_1^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{\sigma_1^2}\left(\sum_{i=1}^{b}\sum_{j=1}^{b_i}\theta_{ij}^2\right)\left(\alpha_1 - \frac{\sum_{i=1}^{b}\sum_{j=1}^{b_i}\theta_{ij}\left(\beta_{0ij}-\beta_{0i}\right)}{\sum_{i=1}^{b}\sum_{j=1}^{b_i}\theta_{ij}^2}\right)^2\right\} \\
\propto &N\left(\hat{\alpha}_1, \frac{\sigma_1^2}{\sum_{i=1}^{b}\sum_{j=1}^{b_i}\theta_{ij}^2}\right)
\end{aligned}
$$

m. The proof of

$$\left[\chi_i \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\theta}_{i(j)}, \chi, \omega^2\right] \propto N\left(\frac{\omega^2 b_i \bar{\theta}_i + \chi}{b_i\omega^2 + 1}, \frac{\omega^2}{b_i\omega^2 + 1}\right)$$

where

$$\bar{\theta}_i = \frac{1}{b_i}\sum_{j=1}^{b_i}\theta_{ij}$$
$$\left[\bar{\theta}_i \mid \chi_i\right] \sim N\left(\chi_i, \frac{1}{b_i}\right).$$

$$\begin{aligned}
&P\left(\chi_i \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\theta}_{i(j)}, \chi, \omega^2\right)\\
\propto &P\left(\boldsymbol{\theta}_{i(j)} \mid \chi_i\right) \times P\left(\chi_i \mid \chi, \omega^2\right)\\
\propto &P\left(\bar{\theta}_i \mid \chi_i\right) \times P\left(\chi_i \mid \chi, \omega^2\right)\\
\propto &exp\left\{-\frac{b_i}{2}\left(\bar{\theta}_i - \chi_i\right)^2\right\} \times exp\left\{-\frac{1}{2\omega^2}\left(\chi_i - \chi\right)^2\right\}\\
\propto &exp\left\{-\frac{1}{2}\left[\left(b_i + \frac{1}{\omega^2}\right)\chi_i^2 - 2\left(b_i\bar{\theta}_i + \frac{\chi}{\omega^2}\right)\chi_i\right]\right\}\\
\propto &N\left(\frac{b_i\bar{\theta}_i + \frac{\chi}{\omega^2}}{b_i + \frac{1}{\omega^2}}, \frac{1}{b_i + \frac{1}{\omega^2}}\right)\\
\propto &N\left(\frac{\omega^2 b_i\bar{\theta}_i + \chi}{b_i\omega^2 + 1}, \frac{\omega^2}{b_i\omega^2 + 1}\right)
\end{aligned}$$

n. The proof of

$$\left[\omega^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\chi}_{(i)}\right] \propto IG\left(a_{10} + \frac{b-1}{2}, b_{10} + \frac{1}{2}\sum_{i=1}^{b}\left(\chi_i - \bar{\chi}\right)^2\right)$$

and

$$\left[\chi \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\chi}_{(i)}, \omega^2\right] \propto N\left(\bar{\chi}, \frac{\omega^2}{b}\right)$$

where

$$\chi_{(i)} \text{ denote the vector of } \chi_i, \text{ for } i = 1, \cdots, b.$$

$$\bar{\chi} = \frac{1}{b}\sum_{i=1}^{b}\chi_i$$

Since the joint distribution of χ and ω^2 is

$$\begin{aligned}
&P\left(\chi, \omega^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\chi}_{(i)}\right) \\
\propto &P\left(\boldsymbol{\chi}_i \mid \chi, \omega^2\right) \times P\left(\chi, \omega^2\right) \\
\propto &\left(\omega^2\right)^{-\frac{b}{2}} exp\left\{-\frac{1}{2\omega^2}\sum_{i=1}^{b}\left(\chi_i - \chi\right)^2\right\} \times \left(\omega^2\right)^{-(a_{10}+1)} exp\left\{-\frac{b_{10}}{\omega^2}\right\} \\
\propto &\left(\omega^2\right)^{-(a_{10}+\frac{b}{2}+1)} exp\left\{-\frac{1}{\omega^2}\left[b_{10} + \frac{1}{2}\sum_{i=1}^{b}\left(\chi_i - \chi\right)^2\right]\right\}
\end{aligned}$$

Integrating χ out to get the marginal distribution of ω^2

$$\begin{aligned}
&P\left(\omega^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\chi}_{(i)}\right) \\
\propto &\left(\omega^2\right)^{-(a_{10}+\frac{b}{2}+1)} exp\left\{-\frac{b_{10}}{\omega^2}\right\} \times \int_{-\infty}^{+\infty} exp\left\{-\frac{1}{2\omega^2}\left[b\left(\chi - \bar{\chi}\right)^2 + \sum_{i=1}^{b}\chi_i^2 - b\bar{\chi}^2\right]\right\} d\chi \\
\propto &\left(\omega^2\right)^{-(a_{10}+\frac{b+1}{2})} exp\left\{-\frac{b_{10}}{\omega^2}\right\} \times exp\left\{-\frac{1}{2\omega^2}\sum_{i=1}^{b}\left(\chi_i^2 - \bar{\chi}\right)^2\right\} \\
\propto &IG\left(a_{10} + \frac{b-1}{2}, b_{10} + \frac{1}{2}\sum_{i=1}^{b}\left(\chi_i - \bar{\chi}\right)^2\right)
\end{aligned}$$

Then the conditional probability of χ is

$$
\begin{aligned}
& P\left(\chi \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\chi}_i, \omega^2\right) \\
=& \frac{P\left(\chi, \omega^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\chi}_{(i)}\right)}{P\left(\omega^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\chi}_{(i)}\right)} \\
\propto & exp\left\{-\frac{1}{2\omega^2}\left[\sum_{i=1}^{b}\left(\chi_i-\chi\right)^2-\sum_{i=1}^{b}\left(\chi_i-\bar{\chi}\right)^2\right]\right\} \\
\propto & exp\left\{-\frac{1}{2\omega^2} b\left(\chi-\bar{\chi}\right)^2\right\} \\
\propto & N\left(\bar{\chi}, \frac{\omega^2}{b}\right)
\end{aligned}
$$

o. The proof of

$$
\begin{aligned}
& \left[\sigma^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\beta}_{0(i)}\right] \\
\propto & IG\left(a_1+\frac{b-1}{2}, b_1+\frac{1}{2}\sum_{i=1}^{b}\left(\beta_{0i}-\bar{\beta_0}\right)^2\right)
\end{aligned}
$$

and

$$
\left[\beta_0 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\beta}_{0(i)}, \sigma^2\right] \propto N\left(\bar{\beta_0}, \frac{\sigma^2}{b}\right)
$$

where

$\boldsymbol{\beta}_{0i}$ denote the vector of $\beta_0 i,$ for $i = 1, \cdots, b$

$$
\bar{\beta_0} = \frac{1}{b}\sum_{i=1}^{b}\beta_{0i}
$$

Since the joint distribution of β_0 and σ^2 is

$$
\begin{aligned}
& P\left(\beta_0, \sigma^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\beta}_{0(i)}\right) \\
\propto & P\left(\boldsymbol{\beta}_{0(i)} \mid \beta_0, \sigma^2\right) \times P\left(\beta_0, \sigma^2\right) \\
\propto & \left(\sigma^2\right)^{-\frac{b}{2}} exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{b}\left(\beta_{0i}-\beta_0\right)^2\right\} \times \left(\sigma^2\right)^{-(a_1+1)} exp\left\{-\frac{b_1}{\sigma^2}\right\} \\
\propto & \left(\sigma^2\right)^{-\left(a_1+\frac{b}{2}+1\right)} exp\left\{-\frac{1}{\sigma^2}\left[b_1+\frac{1}{2}\sum_{i=1}^{b}\left(\beta_{0i}-\beta_0\right)^2\right]\right\}
\end{aligned}
$$

Integrating β_0 out to get the marginal distribution of σ^2

$$
\begin{aligned}
& P\left(\sigma^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\beta}_{0(i)}\right) \\
\propto & \left(\sigma^2\right)^{-\left(a_1+\frac{b}{2}+1\right)} exp\left\{-\frac{b_1}{\sigma^2}\right\} \int_{-\infty}^{+\infty} exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{b}\left(\beta_{0i}-\beta_0\right)^2\right\} d\beta_0 \\
\propto & \left(\sigma^2\right)^{-\left(a_1+\frac{b}{2}+1\right)} exp\left\{-\frac{b_1}{\sigma^2}\right\} \int_{-\infty}^{+\infty} exp\left\{-\frac{1}{2\sigma^2} b\left(\beta_0-\bar{\beta}_0\right)^2\right\} d\beta_0 \\
\times & exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{b}\left(\beta_{0i}-\bar{\beta}_0\right)^2\right]\right\} \\
\propto & IG\left(a_1+\frac{b-1}{2}, b_1+\frac{1}{2}\sum_{i=1}^{b}\left(\beta_{0i}-\bar{\beta}_0\right)^2\right)
\end{aligned}
$$

Then the conditional probability of β_0 is

$$
\begin{aligned}
& P\left(\beta_0 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\beta}_{0(i)}, \sigma^2\right) \\
= & \frac{P\left(\beta_0, \sigma^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\beta}_{0(i)}\right)}{P\left(\sigma^2 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\beta}_{0(i)}\right)} \\
\propto & \left(\sigma^2\right)^{-\frac{1}{2}} exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{b}\left(\beta_{0i}-\beta_0\right)^2\right\} \\
\propto & \left(\sigma^2\right)^{-\frac{1}{2}} exp\left\{-\frac{1}{2\sigma^2} b\left(\beta_0-\bar{\beta}_0\right)^2\right\} \\
\propto & N\left(\bar{\beta}_0, \frac{\sigma^2}{b}\right)
\end{aligned}
$$

Appendix B

Proofs

B.1 Proof of Proposition 1

We first denote the vector of the p random variables by $\boldsymbol{y}$, and the vector of missing indicators for each of the p random variables as $\boldsymbol{r}$. We define the joint density function of $\boldsymbol{y}$ and $\boldsymbol{r}$ as $f(\boldsymbol{y}, \boldsymbol{r})$. We further define $\boldsymbol{r_i}$ as the vector of missing indicators for the ith missing-pattern group, and $f_i(\boldsymbol{y})$ as the joint density function of the p variables (including the variables both observed and missing) for the ith missing-pattern group, $i = 1, ..., s$. It is clear that $f_i(\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{r}_i)$.

We first prove that, if the missingness is MCAR, $F_1 = \cdots = F_s$. Based on the definition of MCAR, the missingness does not depend on the data, which implies $f(\boldsymbol{r}|\boldsymbol{y}) = f(\boldsymbol{r})$. Therefore, $f(\boldsymbol{y}, \boldsymbol{r}) = f(\boldsymbol{r}|\boldsymbol{y})f(\boldsymbol{y}) = f(\boldsymbol{r})f(\boldsymbol{y})$. Since $f(\boldsymbol{y}, \boldsymbol{r}) = f(\boldsymbol{y}|\boldsymbol{r})f(\boldsymbol{r})$, we have $f(\boldsymbol{y}|\boldsymbol{r}) = f(\boldsymbol{y})$. This further implies that $f_1(\boldsymbol{y}) = \cdots = f_s(\boldsymbol{y})$. In other words, $F_1 = \cdots = F_s$.

Next, we prove that, if $F_1 = \cdots = F_s$, the missingness is MCAR. Since $F_1 = \cdots = F_s$, $f_1(\boldsymbol{y}) = \cdots = f_s(\boldsymbol{y})$, i.e., $f(\boldsymbol{y}|\boldsymbol{r}_1) = \cdots = f(\boldsymbol{y}|\boldsymbol{r}_s) = f(\boldsymbol{y})$. Therefore, we have

$$f(\boldsymbol{r}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{r})f(\boldsymbol{r})}{f(\boldsymbol{y})} = \frac{f(\boldsymbol{y})f(\boldsymbol{r})}{f(\boldsymbol{y})} = f(\boldsymbol{r}),$$

which suggests that the missingness is independent of the data. Therefore, the missingness is MCAR.

B.2 Proof of Theorem 2

Suppose the null hypothesis is false, say $F_{k,\mathbf{o}_{kl}} \neq F_{l,\mathbf{o}_{kl}}$ for some $k \neq l \in \{1, ..., s\}$ and $\mathbf{o}_{kl} \neq \emptyset$. Since $F = \frac{B/(s-1)}{W/(n-s)}$, $B = \sum_{\substack{1 \leq i < j \leq s \\ \mathbf{o}_{ij} \neq \emptyset}} (\frac{n_i n_j}{2n}) d(\mathbb{Y}_{i,\mathbf{o}_{ij}}, \mathbb{Y}_{j,\mathbf{o}_{ij}})$, and $d(\mathbb{Y}_{i,\mathbf{o}_{ij}}, \mathbb{Y}_{j,\mathbf{o}_{ij}})$ is always nonnegative, we have

$$F \geq \frac{n_k n_l}{2n} \cdot \frac{d(\mathbb{Y}_{k,\mathbf{o}_{kl}}, \mathbb{Y}_{l,\mathbf{o}_{kl}})}{s-1} \cdot \frac{n-s}{W}.$$

Therefore,

$$\begin{aligned} P(F > c_\alpha) &\geq P\left(\frac{n_k n_l}{2n} \cdot \frac{d(\mathbb{Y}_{k,\mathbf{o}_{kl}}, \mathbb{Y}_{l,\mathbf{o}_{kl}})}{s-1} \cdot \frac{n-s}{W} > c_\alpha \right) \\ &= P\left(d(\mathbb{Y}_{k,\mathbf{o}_{kl}}, \mathbb{Y}_{l,\mathbf{o}_{kl}}) > \frac{2c_\alpha n(s-1)W}{n_k n_l (n-s)} \right). \end{aligned}$$

Since $W = \sum_{i=1}^{s} n_i g(\mathbb{Y}_{i,\mathbf{o}_i}, \mathbb{Y}_{i,\mathbf{o}_i})/2$ and $n_i g(\mathbb{Y}_{i,\mathbf{o}_i}, \mathbb{Y}_{i,\mathbf{o}_i})/(n_i - 1)$ a U-statistic, based on the properties of U-statistics,

$$n_i g(\mathbb{Y}_{i,\mathbf{o}_i}, \mathbb{Y}_{i,\mathbf{o}_i})/(n_i - 1) \to \eta_i, \text{ a.s.}$$

where η_i is a constant. This implies that

$$\begin{aligned} W/(n-s) &= \sum_{i=1}^{s} \frac{n_i - 1}{n-s} \cdot \frac{1}{2} \cdot \frac{n_i}{n_i - 1} g(\mathbb{Y}_{i,\mathbf{o}_i}, \mathbb{Y}_{i,\mathbf{o}_i}) \\ &\to \frac{1}{2} \sum_{i=1}^{s} \lambda_i \eta_i, \text{ a.s.} \end{aligned}$$

where $\lambda_i = \lim_{n\to\infty} \frac{n_i}{n_1+\cdots+n_s}$. Therefore,

$$\begin{aligned} \lim_{n\to\infty} P(F > c_\alpha) &\geq \lim_{n\to\infty} P\left(d(\mathbb{Y}_{k,\mathbf{o}_{kl}}, \mathbb{Y}_{l,\mathbf{o}_{kl}}) > \frac{2c_\alpha n(s-1)W}{n_k n_l (n-s)} \right) \\ &= \lim_{n\to\infty} P\left(d(\mathbb{Y}_{k,\mathbf{o}_{kl}}, \mathbb{Y}_{l,\mathbf{o}_{kl}}) > \frac{c_\alpha (s-1) \sum_{i=1}^{s} \lambda_i \eta_i}{n \lambda_k \lambda_l} \right). \end{aligned} \tag{B.1}$$

Next we show that c_α is bounded above by a constant which does not depend on n. Recall that $F = \frac{B/(s-1)}{W/(n-s)}$, $B = \sum_{\substack{1\leq i<j\leq s \\ \mathbf{o}_{ij}\neq\emptyset}} (\frac{n_i n_j}{2n}) d(\mathbb{Y}_{i,\mathbf{o}_{ij}}, \mathbb{Y}_{j,\mathbf{o}_{ij}})$. Denote the number of the pairs (i,j) satisfying $1 \leq i < j \leq s$ and $\mathbf{o}_{ij} \neq \emptyset$ by t. In other words, there are t terms in B. Clearly, $t \leq s(s-1)/2$. Therefore, for any k,

$$\begin{aligned} P(F > k) &\leq P\left(\text{at least one of the } \frac{n_i n_j}{2n} \cdot \frac{d(\mathbb{Y}_{i,\mathbf{o}_{ij}}, \mathbb{Y}_{j,\mathbf{o}_{ij}})}{s-1} \cdot \frac{n-s}{W} > k/t \right) \\ &\leq \sum_{\substack{1\leq i<j\leq s \\ \mathbf{o}_{ij}\neq\emptyset}} P\left(\frac{n_i n_j}{n_i + n_j} d(\mathbb{Y}_{i,\mathbf{o}_{ij}}, \mathbb{Y}_{j,\mathbf{o}_{ij}}) > \frac{2kn(s-1)W}{t(n_i+n_j)(n-s)} \right), \end{aligned}$$

and

$$\lim_{n\to\infty} P(F > k) \leq \lim_{n\to\infty} \sum_{\substack{1\leq i<j\leq s \\ \mathbf{o}_{ij}\neq\emptyset}} P\left(\frac{n_i n_j}{n_i + n_j} d(\mathbb{Y}_{i,\mathbf{o}_{ij}}, \mathbb{Y}_{j,\mathbf{o}_{ij}}) > \frac{k(s-1)\sum_{i=1}^{s} \lambda_i \eta_i}{t(\lambda_i + \lambda_j)} \right). \tag{B.2}$$

Based on Székely and Rizzo (2005), under the null hypothesis of equal distributions, $n_i n_j d(\mathbb{Y}_{i,\mathbf{o}_{ij}}, \mathbb{Y}_{j,\mathbf{o}_{ij}})/(n_i + n_j)$ converges in distribution to a quadratic form

$$Q_{i,j} = \sum_{l=1}^{\infty} \omega_l Z_l^2,$$

where the Z_l are independent standard normal random variables and the ω_l are positive constants and do not depend on n. Therefore, we can choose $k = k_\alpha$, a constant which does not depend on n, such that

$$\sum_{\substack{1 \le i < j \le s \\ \mathbf{o}_{ij} \neq \emptyset}} P\left(Q_{i,j} > \frac{k_\alpha (s-1) \sum_{i=1}^{s} \lambda_i \eta_i}{t(\lambda_i + \lambda_j)} \right) = \alpha.$$

For such a k_α, we have $\lim_{n\to\infty} P(F > k_\alpha) \le \alpha$ under H_0 based on (B.2). Since $\lim_{n\to\infty} P(F > c_\alpha) = \alpha$ under H_0, $\lim_{n\to\infty} c_\alpha \le k_\alpha$. Therefore, we have shown that c_α bounded above by k_α, a constant which does not depend on n.

Applying this result to (B.1), we have $c_\alpha (s-1) \sum_{i=1}^{s} \lambda_i \eta_i / (n \lambda_k \lambda_l) \to 0$, as $n \to \infty$. Since $d(\mathbb{Y}_{k,\mathbf{o}_{kl}}, \mathbb{Y}_{l,\mathbf{o}_{kl}})$ is a V-statistic, $d(\mathbb{Y}_{k,\mathbf{o}_{kl}}, \mathbb{Y}_{l,\mathbf{o}_{kl}})$ converges in probability to 0 if $F_{k,\mathbf{o}_{kl}} = F_{l,\mathbf{o}_{kl}}$, and to some nonzero constant if $F_{k,\mathbf{o}_{kl}} \neq F_{l,\mathbf{o}_{kl}}$. Therefore,

$$\lim_{n\to\infty} P\left(d(\mathbb{Y}_{k,\mathbf{o}_{kl}}, \mathbb{Y}_{l,\mathbf{o}_{kl}}) > \frac{c_\alpha (s-1) \sum_{i=1}^{s} \lambda_i \eta_i}{n \lambda_k \lambda_l} \right) = 1,$$

which implies that $\lim_{n\to\infty} P(F > c_\alpha) = 1$. As a result, our F test is consistent. This completes the proof.