2016

Development in Normal Mixture and Mixture of Experts Modeling

Meng Qi
University of Kentucky, maria.mqq88@gmail.com

Digital Object Identifier: http://dx.doi.org/10.13023/ETD.2016.042

STUDENT AGREEMENT:

REVIEW, APPROVAL AND ACCEPTANCE

Meng Qi, Student

Dr. Richard Charnigo, Major Professor

Dr. Constance Wood, Director of Graduate Studies

DEVELOPMENT IN NORMAL MIXTURE AND MIXTURE OF EXPERTS MODELING

DISSERTATION

A dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor of
Philosophy in the College of Arts and Sciences
at the University of Kentucky

By

Meng Qi
Lexington, Kentucky

Director: Dr. Richard Charnigo, Professor of Statistics
Lexington, Kentucky

2016

ABSTRACT OF DISSERTATION

DEVELOPMENT IN NORMAL MIXTURE AND MIXTURE OF EXPERTS MODELING

In this dissertation, first we consider the problem of testing homogeneity and order in a contaminated normal model, when the data is correlated under some known covariance structure. To address this problem, we developed a moment based homogeneity and order test, and design weights for test statistics to increase power for homogeneity test. We applied our test to microarray about Down's syndrome data. This dissertation also studies a singular Bayesian information criterion (sBIC) for a bivariate hierarchical normal mixture model with varying weights, and develops a new data dependent information criterion (sFLIC). We apply our model and criteria to birthweight and gestational age data for the same model, whose purposes are to select model complexity from data.

KEYWORDS: Finite Mixture Models, Micro-array Analysis, Homogeneity Test, Information Criterion, Hierarchical Mixture Model

Author's signature: Meng Qi

Date: April 6, 2016

DEVELOPMENT IN NORMAL MIXTURE AND MIXTURE OF EXPERTS MODELING

By

Meng Qi

Director of Dissertation: Dr. Richard Charnigo

Director of Graduate Studies: Dr. Constance Wood

Date: April 6, 2016

ACKNOWLEDGMENTS

First of all, I will express my deep appreciation and gratitude to my advisor Dr. Richard Charnigo, for his persistent encouragement and detailed guidance. Still remember three years ago, when I first started to working on my dissertation, it was a hard time because of my weakness in English writing and lack of knowledge in research. Dr. Charnigo is an excellent advisor, he gave me detailed guidance and considerable encouragement in every steps I made in my research. I benefit a lot from his profound knowledge and profession skills. I sincerely appreciate for his help and instruction to finish my dissertation. In addition, Dr. Charnigo also helped me to develop my research ability, writing skills and analytical thinkings. Without him, I will not get my thesis done.

Secondly, I want to express my sincere gratitude to my dissertation committee members: Dr. Arnold Stromberg, Dr. Ruriko Yoshida, Dr. William Griffith, Dr. David Fardo, Dr. Chi Wang and Dr. Dereck Young for their devotion to improving my dissertation. It is such a honor to have you all in my committee.

I am very grateful to my friends Yifan Yang, Qian Fan, Hongyuan Wang and all my classmates for making my 5-year PhD study a wonderful experience. Also, I would appreciate all faculty and staff in the Department of Statistics for their help.

Finally, I would like to thank my parents for their love, support and understanding. I will never achieve my goal without them.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

1.1 Finite Mixture models

Finite mixture models provided an reasonable approach to modeling many phenomena which can not be accurately described via a commonly encountered distribution. And it can be used in many areas like: epidemiology, genetics, marketing(Jedidi et al. [1997]), finance(Lamoureux and Lastrapes [1994]) and agriculture(Atkinson et al. [1997]).

Example 1: In epidemiologic area, the mixture model can describe the birthweight distribution and guide inference about the relation between birthweight and infant mortality(see Charnigo et al. [2010], Wilcox and T Russell [1986], Gage and Therriault [1998]). Simple bell curves are not adequate to describe the birthweight distribution, especially for 'low birthweight group(less than 2500g) which is uninformative and seldom justified. Also for different types of population, the data may be shown in different feature. In such a scenario, mixture models are potentially interpretable, since the components of the mixture may correspond to subpopulation with biologically meaningful characteristics.

Example 2: Mixture models can also be used in genetics. Ott [1999] proposed using a mixture of Binomials to describe the distribution of recombination between genetic traits and markers. Since in a homogeneous population, the probability of recombination θ between a gene and a marker inherited by a child from a parent is $\frac{1}{2}$ if the gene and the marker are independent and is in $[0, \frac{1}{2})$ if they are linked, the model can be written as

$$\lambda B(n,\theta) + (1-\lambda)B(n,\frac{1}{2}) \; where \; \lambda \in [0,1] \; \theta \in [0,\frac{1}{2}] \; and \; n \in \{2,3,...\} \; .$$

Also,λ is the unknown proportion of families with linkage and n is the family size.

Now we describe the model. We discuss two cases: with and without nuisance parameters.Consider $\{f(y;\theta,\beta) : \theta \in \Theta, \beta \in \mathbb{R}\}$ or $\{f(y;\theta) : \theta \in \Theta,\}$ to be a family of probability density(or mass) functions. Let $\mathbb{B}$ be the space of probability measures on Θ.

Case one (with nuisance parameter β) For a finite mixture model with k components, the mixture density(or mass) function can be written as

$$\sum_{j=1}^{k} \lambda_j f(x, \theta_j, \beta) \ where \ \lambda_j \in [0,1] \ and \sum_{j=1}^{k} \lambda_j = 1. \tag{1.1}$$

And more generally speaking, the model can be written as

$$g(y; Q, \beta) = \int f(y; \theta, \beta) dQ(\theta) \ for \ Q \in \mathbb{B} \tag{1.2}$$

a mixture density(or mass function) with mixing distribution Q. Note that, for a finite mixture model, Q is a finitely supported discrete distribution.

Case two (without nuisance parameters) For a finite mixture , the mixture density(or mass) function is

$$\sum_{j=1}^{k} \lambda_j f(x, \theta_j) \ where \ \lambda_j \in [0,1] \ and \sum_{j=1}^{k} \lambda_j = 1. \tag{1.3}$$

The general model is defined by Charnigo and Pilla [2007]

$$g(y; Q) = \int f(y; \theta) dQ(\theta) \ for \ Q \in \mathbb{B} \tag{1.4}$$

a mixture density(or mass function) with mixing distribution Q.

As noted above, Q is always modeled parametrically for a finite mixture model, although Q can be modeled parametrically in the more general setting; for instance, Q can be a normal distribution. If Q is modeled non-parametrically, g is referred to as a semi-parametric mixture model(Charnigo and Pilla [2007]).

1.2 EM approach

A general description of the EM algorithm is provided by Dempster et al. [1977]. EM approach is a general strategy to iteratively compute maximum-likelihood estimates and has two general steps each iteration: expectation step(E step) and maximum step(M step). For mixture models, we usually apply EM algorithm to estimate the unknown parameters(here we assume, perhaps only temporarily that the number of components is known), due to the following reasons. Firstly, EM algorithm provides an approximation to MLE(maximum likelihood estimation) without requiring numerical solutions to the difficult high-dimensional optimization problems. Secondly, we can consider which component an individual belongs to as a latent variable, then the mixture model can be expressed in terms of incomplete data. As mentioned in Dempster, Laird and Rubin's paper, EM approach has a more natural interpretation than MLE in the context of incomplete data.

In what follows, we describe application of the EM algorithm to mixture model; Suppose we have $X_1, X_2, ...X_n$ be iid random variables from a finite mixture model with k components

$$\sum_{j=1}^{k} \lambda_j f(x|\mu_j, \beta) \; where \; \lambda_j \in [0,1] \; and \sum_{j=1}^{k} \lambda_j = 1. \tag{1.5}$$

Let $z_{ij} = I$[individual i belongs to the j^{th} component], then the complete data log-likelihood function can be written as

$$l(\mathbf{x}, \mathbf{z}|\theta) = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij}[\log \lambda_{ij} + \log f(x_i|\mu_j, \beta)]. \tag{1.6}$$

At the t^{th} iteration, perform the following two steps:

E-step

Put

$$Q(\theta|\theta^{(\mathbf{t})}) = \mathbf{E}[l(\mathbf{x},\mathbf{z}|\theta)|\mathbf{x},\lambda_1^{(t)}...\lambda_j^{(t)},\mu_1^{(t)}...\mu_j^{(t)},\beta^{(t)}] \tag{1.7}$$

where quantities labeled (t) are estimates after iteration t.

Let

$$w_{ij}^{(t)} = \frac{\lambda_j f(x_i|\mu_j^{(t)},\beta^{(t)})}{\sum_{j'=1}^{k}\lambda_{j'} f(x_i|\mu_{j'}^{(t)},\beta^{(t)})} \tag{1.8}$$

Then equation 1.7 becomes:

$$\sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}^{(t)}\log\lambda_j + \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}^{(t)} f(x_i;\mu_j,\beta) \tag{1.9}$$

For example, if $f(x;\mu_j,\beta)$ is normal with mean μ_j and variance β, then we obtain

$$\sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}^{(t)}\log\lambda_j + \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}^{(t)}[-\frac{1}{2}\log\beta - \frac{1}{2}\frac{(x_i-\mu_j)^2}{\beta}] \tag{1.10}$$

M-step

Maximizing the function 1.10 gives the estimator after the next iteration. Continuing our normal example, we have

$$\lambda_j^{(t+1)} = \frac{\sum_{i=1}^{n} w_{ij}^{(t)}}{n},$$

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{n} w_{ij}^{(t)} x_i}{\sum_{i=1}^{n} w_{ij}^{(t)}},$$

and

$$\beta^{(t+1)} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n} w_{ij}^{(t)}(x_i-\mu_j^{(t+1)})^2}{\sum_{j=1}^{k}\sum_{i=1}^{n} w_{ij}^{(t)}}.$$

Then, update 1.10 with the new parameter estimates, and iterate until

$$|l(\mu_1^{(t+1)}, ...\mu_k^{(t+1)}, \beta^{(t+1)}, \lambda_1^{(t+1)}, ...\lambda_k^{(t+1)}) - l(\mu_1^{(t)}, ...\mu_k^{(t)}, \beta^{(t)}, \lambda_1^{(t)}, ...\lambda_k^{(t)})| < \varepsilon$$

, for some small $\varepsilon > 0$; that is iterate until the likelihood converges. And if the model is correct, then under regularity conditions, the MLE's as approximated by the EM approach are $\sqrt{\text{n}}$-consistent(Redner and Walker [1984]). Note, $\sqrt{n}$-consistency dose not hold if the model is not correct(i.e. there are either too few or two many components).

1.3 Estimate the number of mixture component

Testing

Let m := |supp Q| an important but challenging problem in mixture model is determining m, which is the order of the mixture distribution.

Even for the simplest case, testing for homogeneity in finite mixture models without nuisances parameters(1 component vs. 2 component) is not that easy.i.e

$$(1-\gamma)f(x;\theta_1) + \gamma f(x;\theta_2), \theta_1, \theta_2 \in \Theta, \gamma \in [0,1] \tag{1.11}$$

where $\{f(x;\theta), \theta \in \Theta\}$ is family of pdfs(or pmfs).

Testing hypothesis:

$$H_0 : \gamma(1-\gamma)(\theta_2 - \theta_1) = 0 \; vs. \; H_1 : \gamma(1-\gamma)(\theta_2 - \theta_1) \neq 0$$

(If H_0 is true, then equation(1.11) simplifies to $f(x, \theta_0)$ for some θ_0.) is not easy.

LRT(likelihood ratio test) seems to be the best choice since it locally ,most powerful test(Chen et al. [2001]). But Hartigan [1985] showed that the LRT statistics will diverges to ∞ under the null hypothesis(homogeneity) when the family is normal, which has known mean θ_1, unknown mean θ_2, standard deviation equals to 1 and

$\Theta = \mathbb{R}$. To solve this problem, much research has been done (see McLachlan and Basford [1988], Lindsay [1995], and reference therein).

More recently, Chen et al. [2001], Dacunha-Castelle et al. [1999] showed that if Θ is compact, θ_1, θ_2, γ are unknown and $\{f(x, \theta), : \theta \in \Theta\}$ satisfied some condition, then the LRT statistic will converges in law to a random variable $sup_{\theta \in \Theta}\{(max(0, W(\theta)))^2\}$, where $\{W(\theta), \theta \in \Theta\}$ is a Gaussian process. Even though we have the limit distribution of the test statistic, there are still some problems. Since the Gaussian process depends on the parameter space Θ, it may vary as the space is changed. Even we fixed a space, the critical value is hard to calculate.Chen et al. [2001] developed a new method by introducing the MLRT(Modified likelihood ratio test). Different from the LRT, MLRT adds a penalty term into the likelihood function, in order to force the estimator of γ away from 0 and 1. Then under basically the same condition, MLRT statistic converges in law to $(max(0, W(\theta_0))^2$. Then under null hypothesis, $\theta_1, \theta_2, \gamma$ unknown, test statistic will converge in law to $\frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_0^2$ (where χ_0^2 is a degenerate random variable at value 0). The MLRT is also a locally most powerful test and has asymptotic tractable distribution under the null. Further more, Dai and Charnigo [2008] showed that forθ_1 known,θ_2, γ unknown, the MLRT statistic will converge in law to χ_k^2 under the null hypothesis, where the $k \in \mathbb{N}$ is known, and is the dimension of parameter space.

The MLRT also has some drawbacks, like the formal expressions for the test statistic in terms of parameter estimators are not only complicated, but also still require the exact value of $x_1 ... x_n$. Charnigo and Sun [2004] proposed a new test, D-test, which is based on the L^2 distance between a fitted homogeneous model and a fitted heterogeneous model. This test may be applied for mixture from a parametric family of continuous distribution and has a greater advantage than MLRT if the full dataset is not readily accessible, since D-test statistics depend on data only through mixture parameter estimators and has a simple form. Also, if θ_1, θ_2,γ are estimated using modified likelihood, D-test statistic has a tractable null distribution which is provided in Charnigo and Sun [2010]. Charnigo and Sun also developed a test for mixtures from a parametric family of discrete distributions in 2008(see Charnigo and Sun [2008]), called W-test. This test is competitive with MLRT in terms of power, and it also depends on data solely through the parameter estimators. Moreover, in many situation, W-test can use large sample critical value with small to moderate size samples.

It seems that the problem of choosing the order of the mixture model is solved by those tests. For more complex cases, however, like testing 2 components vs. 3 components or even a test of homogeneity under more general circumstances, those tests can not be easily be applied, since the asymptotic distribution is unclear. To solve this problem, Chen and Li [2009] proposed the EM-test, which has a simple limiting distribution. This test based on the EM algorithm using a small number of iterates to estimate parameters before construction the test statistics. The limiting distribution of the EM-test statistic has a nice asymptotic distribution, for testing 1 component vs. 2 components, it has the same asymptotic distribution as the MLRT. While for 2 components vs. 3 or 4 components, under the null hypothesis, the EM-test statistic converges in law to $\alpha_0\chi_0^2 + \alpha_1\chi_1^2 + \alpha_2\chi_2^2$ where $\alpha_0, \alpha_1, \alpha_2$ sum up to 1 and can be calculated based on the parametric family $\{f(x, \theta) : \theta \in \Theta\}$. And for testing 3 components vs. 4,5 or 6 components, the asymptotic form of the test statistic under the null hypothesis is $\alpha_0\chi_0^2 + \alpha_1\chi_1^2 + \alpha_2\chi_2^2 + \alpha_3\chi_3^2$ where $\alpha_0, \alpha_1, \alpha_2, \alpha_3$ sum up to 1 and can be calculated based on the parametric family $\{f(x, \theta) : \theta \in \Theta\}$. These are useful results, but the EM-test does have some complications: Firstly, for testing 1 vs. 2 components, the weight for the chi-square distributions are immediately apparent. But it seems less easy to solve α_0 to α_3 for testing 3 components vs. 4,5 or 6 components. Secondly, the level of significance is associated with sequential testing, which makes its ascertainment complex. What's more, the test works beautifully if θ is a scalar, but above conclusions do not hold for the case in which θ is a vector nor for hierarchical mixture models. There is much work remaining to be done.

Information criterion

Besides testing, another major contribution to selecting the proper number of mixture components is the development of model selection criteria, such as AIC(Akaike information criterion) and BIC(Bayesian information criterion). These two criteria belong to the family of penalized likelihood criteria, as well as the algorithm provided by Figueiredo and Leitã [1993] for estimating a mixture model. Suppose we generalized our model selection problem wit h models indexed by $m \in \{1, 2...M\}$. Denote the complexities of those models to be $C_1 < C_2 < ... < C_M$.Then the AIC(Akaike [1973]) and BIC(Schwarz et al. [1978]) have penalty terms $2C_m$ and $C_m \log n$ respectively. Figueiredo and Leitã [1993] showed an criterion named MMDL(mixture minimum description length type criterion) which is based on the identication of an equivalent sample size, for each component. MMDL introduced a lower penalty than BIC, but still compatible with BIC. Lahiri [2001] showed that the AIC is an inconsistent estimator when n is large, while the BIC will underestimate the number of components when n is small. So some researchers have defined new criteria which could have the consistency of BIC, while also retaining the small sample performance of AIC. Keribin [2000] proposed an almost surely consistent penalized likelihood estimator for given appropriate penalization sequence, based on locally conic parameterizations(Dacunha-Castelle et al. [1999]).However, Kerbin's estimator did not have a data-dependent penalty.

Thus, Pilla and Charnigo [2006] proposed a new model selection criteria named FLIC(Flexible information criterion). FLIC performs better than BIC when components are poorly separated and n is small, while also performing better than AIC when components are well separated and n is large. Importantly, FLIC takes into account the structure of the data to determine the strength of the penalty term. More specifically, when the dimension of θ is $3m$, the penalty term for FLIC is $2(\log \sqrt{n})^{B(n,\delta)}(3m - 1)$, where 3m-1 is the number of free parameters in a m-component mixture model. n is the sample size, δ denotes the fraction of within-component variability to the total variability, and $B(n,\gamma) = \frac{\Phi[(\log\sqrt{n})^{\gamma}]-\Phi(1)}{1-\Phi(1)}$ is a bivariate function taking a value between 0 and 1. Since the penalty term is not only determined by simple size n, but also by the data configuration, it tends to select more components if the data suggest greater heterogeneity.

Drton and Plummer [2013] proposed a new information criterion called sBIC(singular Bayesian information criterion), which is a Bayesian information criterion in context of a singular model selection problem. Note that the singular model refers to the models whose Fisher information matrices may be singular and fail to be invertible. For BIC, the large sample quadratic approximation to the log-likelihood function is not possible when the Fisher-information matrix is singular(Watanabe [2009a]). However sBIC can circumvent this singularity problem. The sBIC agrees with the BIC for regular model, and can be calculated without using the Monte Carlo computations. Since the sBIC makes use of the information about the learning coefficients that capture the large sample behavior of the concerned marginal likelihood integrals, the sBIC is not only consistent, but also enjoys some nice properties about Bayesian model choice in singular settings, which normal BIC is not applicable due to the invertible of Fisher information matrix. Note that sBIC penalty is no more stronger than ordinary BIC penalty. Hence sBIC will select a equal or greater complexity than BIC.

While the FLIC and sBIC have been applied to univariate mixture models, to our knowledge, neither has been applied to multivariate mixture models expect in one special case: a bivariate normal mixture(Fan [2014]). Fan [2014], Chapter 4 introduces a hierarchical normal mixture model with nuisance parameters(HNP+NP model) and applied the sBIC to it. Chapter 5 shows a new data dependent information criterion inspired by Pilla and Charnigos FLIC (see Charnigo and Pilla [2007]) for HNP+NP model.

Chapter 2 Two component normal model under correlation

2.1 Modeling contamination under correlation

In microarray data analysis and large-scale hypothesis testing, p-values of multiple tests can be modeled as a mixture of Beta distributions(Allison et al. [2002], Dai and Charnigo [2008]). Furthermore, Dai and Charnigo [2010] examined a different approach of using a contaminated normal model to describe the distribution of Z statistics from such tests.

For iid Z statistics, the homogeneity testing problem was solved by Dai and Charnigo [2010]. They indicated that, under the null hypothesis, both the MLRT statistic (modified likelihood ratio test, proposed by Chen et al. [2001]) and empirical D-test statistic(proposed by Charnigo and Sun [2004]) will converge in law to χ^2_1. Under any fixed alternative, MLRT statistic and D-test statistic will converge to positive constant. Thus these tests are consistent. However, this theory can not be applied to correlated Z statistics.

The assumption about iid may sometimes not be applicable, though, for example, when 2 genes are in a common biological pathway, their expression levels that defined two Z statistics may be correlated with each other. This motivates the development of methodology for analyzing the distribution of correlated Z statistics, specified within known clusters by biological pathways. Even though the assumptions of known clusters may not be perfect, it is more satisfactory in practice than assumptions of iid data. In any event, the development of such methodology at least permits a rigorous investigation of robustness of scientific conclusions to the assumption of iid data. Therefore, in this chapter, we show a methodology to deal with correlated data, with focus on homogeneity testing: Z statistics arising from tests of differential expression on some genes constitute a second component with non-zero mean in a contaminated normal model.

The problem description:

We start with paired correlated data. This is a special case which indicates that there are only two genes in each biological pathway. The data is constructed as following: let $X \in \{1, 2\}$ be a variable identifying component membership, $P(X = 2) = \lambda$, where $0 \leq \lambda \leq 1$. Consider $Y|X = 1 \sim N(0, 1)$, $Y|X = 2 \sim N(\mu, 1)$. So

$$Y \sim (1 - \lambda)N(0, 1) + \lambda N(\mu, 1),$$

a contaminated normal without nuisance parameters, the variance is known to be 1 for each component.

Suppose Y_1 and Y_2 may be correlated. Let Z_1, Z_2, Z_3 be independent, $Z_1, Z_2 \sim N(0, \sigma^2)$, $Z_3 \sim N(0, \tau^2)$, put $Y_1 = Z_1 + \mu(X_1 - 1) + Z_3$, $Y_2 = Z_2 + \mu(X_2 - 1) + Z_3$, here we assume $\sigma^2 + \tau^2 = 1$ for now and that σ^2, τ^2 are known.

Now, we will show the calculations of conditional and marginal moments of a contaminated normal distribution, which are necessary for establishing a homogeneity test for whether some genes are differentially expressed.There are some questions:

Question 1

If $X_1 = X_2 = X$(perfectly dependent component membership), then the correlation between Y_1 and Y_2 conditionally on X and marginally follow:

Conditionally, we can find

$$\mathbb{E}(Y_1|X) = \mathbb{E}(Y_2|X) = \mu(X - 1)$$

$$cov(Y_1, Y_2|X) = \mathbb{E}[(Y_1 - \mathbb{E}(Y_1|X))(Y_2 - \mathbb{E}(Y_2|X)] = \tau^2$$

$$var(Y_1|X) = var(Z_1) + var(Z_3) = \sigma^2 + \tau^2 = 1$$

Then the covariance matrix of Y_1 and Y_2 is

$$\Sigma = \begin{pmatrix} 1 & \tau^2 \\ \tau^2 & 1 \end{pmatrix}.$$

Marginally, we can find

$$\mathbb{E}(X_1) = \mathbb{E}(X_2) = 1 + \lambda$$

$$\mathbb{E}(Y_1) = \mathbb{E}(Y_2) = \mu(1 + \lambda) - \mu = \mu\lambda$$

$$cov(Y_1, Y_2) = \mathbb{E}[(Y_1 - \mathbb{E}(Y_1))(Y_2 - \mathbb{E}(Y_2)] = \mu^2\lambda(1 - \lambda) + \tau^2$$

$$var(Y_1) = var(Y_2) = \sigma^2 + \tau^2 + \mu^2\lambda(1 - \lambda) = 1 + \mu^2\lambda(1 - \lambda)$$

Then, the covariance matrix of Y_1 and Y_2 is

$$\Sigma = \begin{pmatrix} 1 + \mu^2\lambda(1 - \lambda) & \tau^2 + \mu^2\lambda(1 - \lambda) \\ \tau^2 + \mu^2\lambda(1 - \lambda) & 1 + \mu^2\lambda(1 - \lambda) \end{pmatrix}.$$

Question 2

If X_1, X_2 are independent then the correlations between Y_1 and Y_2, conditionally on (X_1, X_2) follow:

Conditionally, we can find:

$$\mathbb{E}(Y_1|X_1) = \mu(X_1 - 1) \quad \text{and} \quad \mathbb{E}(Y_2|X_2) = \mu(X_2 - 1)$$

$$cov(Y_1, Y_2|X_1, X_2) = \tau^2$$

$$var(Y_1|X_1) = var(Y_2|X_2) = \sigma^2 + \tau^2 = 1$$

Then, the covariance matrix of Y_1 and Y_2 is

$$\Sigma = \begin{pmatrix} 1 & \tau^2 \\ \tau^2 & 1 \end{pmatrix}$$

. Marginally, we can find:

$$\mathbb{E}(Y_1) = \mathbb{E}(Y_2) = \mu\lambda$$

$$cov(Y_1, Y_2) = \tau^2$$

$$var(Y_1) = var(Y_2) = 1 + \mu^2\lambda(1 - \lambda)$$

Then, the covariance matrix of Y_1 and Y_2 is

$$\Sigma = \begin{pmatrix} 1 + \mu^2\lambda(1 - \lambda) & \tau^2 \\ \tau^2 & 1 + \mu^2\lambda(1 - \lambda) \end{pmatrix}.$$

Question 3

Suppose that $X_1 \neq X_2$ but X_1 and X_2 are correlated.

$$var \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix} \lambda(1-\lambda)$$

for $\theta \in (0,1)$

Marginally, we can find:

$$\mathbb{E}(X_1) = \mathbb{E}(X_2) = 1 + \lambda$$

$$\mathbb{E}(Y_1) = \mathbb{E}(Y_2) = \mu\lambda$$

$$cov(Y_1, Y_2 | (X_1, X_2)) = \tau^2 + \mu^2\theta\lambda(1-\lambda)$$

$$var(Y_1) = var(Y_2) = 1 + \mu^2\lambda(1-\lambda)$$

Conditionally, we have

$$\mathbb{E}(Y_1|X_1) = \mu(X_1 - 1)$$

$$\mathbb{E}(Y_2|X_2) = \mu(X_2 - 1)$$

$$cov(Y_1, Y_2|X_1, X_2) = \tau^2$$

$$var(Y_1|X_1) = var(Y_2|X_2) = \sigma^2 + \tau^2 = 1$$

2.2 Define moment estimators

Toward developing a homogeneity test, we will define moment estimators:

$$\widehat{m_1} = n^{-1}\sum_{i=1}^{n}(Y_{1i} + Y_{2i}) \quad \text{and} \quad \widehat{m_2} = n^{-1}\sum_{i=1}^{n}(Y_{1i} + Y_{2i})^2$$

since

$$\mathbb{E}(Y_{1i} + Y_{2i}) = 2\mu\lambda$$

$$\mathbb{E}(Y_{1i} + Y_{2i})^2 = 4\mu^2\lambda^2 + 2\sigma^2 + 2\mu^2\lambda(1-\lambda) + 4\tau^2 + 2\mu^2\theta\lambda(1-\lambda),$$

Then we define $\widehat{\mu}$ and $\widehat{\lambda}$ by

$$\begin{cases} \widehat{m_1} = 2\widehat{\mu}\widehat{\lambda} \\ \widehat{m_2} = 4\widehat{\mu}^2\widehat{\lambda}^2 + 2\sigma^2 + 2\widehat{\mu}^2\widehat{\lambda}(1-\widehat{\lambda}) + 4\tau^2 + 2\widehat{\mu}^2\theta\hat{\lambda}(1-\widehat{\lambda}). \end{cases}$$

which may be solved as follows:

$$\begin{cases} \widehat{\lambda} &= \frac{(1+\theta)m_1^2}{2(m_2-m_1^2-2-2\tau^2)+(1+\theta)m_1^2} \\ \widehat{\mu} &= \frac{m_1}{2\widehat{\lambda}} \end{cases} \tag{2.1}$$

2.3 Homogeneity hypothesis Testing

Testing null hypothesis $\mu\lambda = 0$ is the homogeneity testing: some genes differently expressed as represented by a second component with non-zero mean in a contaminated normal model.

We consider a moment-based approach for testing the hypothesis since for likelihood ratio test, the assumption of regularity conditions is violated, such as identifiability(Chen et al. [2001]; Dai and Charnigo [2008]). We first show a special case which is paired correlated data. Then we show a more general case for different size of clusters.

Only with paired data per group

Description

As before, $\kappa_1, kappa_2$ are discrete random variables taking values 1 or 2 and are defined as follows: $\mathbb{P}(kappa_i = 1) = 1 - \lambda$, $\mathbb{P}(kappa_i = 2) = \lambda$, where $i = 1, 2$. For calculation convenience, we further define $X_i = \kappa_i - 1$, thus

$\mathbb{P}(X_i = 0) = 1 - \lambda$, $\mathbb{P}(X_i = 1) = \lambda$, where $i = 1, 2$.

$$var\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix}\lambda(1-\lambda),$$

where $\theta \in [0, 1]$

Let Z_1, Z_2, Z_3 be iid Normal(0,1) random variables,

$$Y_1|X_1 = x \sim \mu x + \sigma Z_1 + \tau Z_3$$

$$Y_1|X_2 = x \sim \mu x + \sigma Z_2 + \tau Z_3$$

Testing

For testing

$$H_0 : \mu\lambda = 0 \; vs. \; H_a : \mu\lambda \neq 0,$$

define $m = (m_1, m_2)$, $\widehat{m} = (\widehat{m_1}, \widehat{m_2})$; where

$$\begin{array}{rcl} \widehat{m_1} & = & n^{-1}\sum_{i=1}^{n}(Y_{1i} + Y_{2i}) \\ \widehat{m_2} & = & n^{-1}\sum_{i=1}^{n}(Y_{1i} + Y_{2i})^2 \end{array}$$

and

$$\begin{array}{rcccl} m_1 & = & \mathbb{E}(Y_{1i} + Y_{2i}) & = & 2\mu\lambda \\ m_2 & = & \mathbb{E}(Y_{1i} + Y_{2i})^2 & = & 2 + 2\tau^2 + 2\mu^2\lambda(1-\lambda)(1+\theta) + 4\mu^2\lambda^2 \end{array}$$

Then for testing the hypothesis, under H_0 $m_1 = 0$, while under H_a $m_1 \neq 0$, then define

$$\begin{array}{rcccl} \mathbb{V}(\widehat{m}) & = & var(\widehat{m_1}) & = & \mathbb{E}\widehat{m_2} - (\mathbb{E}\widehat{m_1})^2 \\ \widehat{\mathbb{V}(m)} & = & \widehat{var(m_1)} & = & \widehat{m_2} - \widehat{m_1}^2 \end{array}$$

Then under the null hypothesis, $m_1 = 0$, thus by central limit theorem and slutsky's theorem:

$$\sqrt{n}\frac{\widehat{m_1}}{\sqrt{\widehat{\mathbb{V}(m)}}} \xrightarrow{d} N(0,1)$$

Consider

$$T = \sqrt{\frac{n}{\widehat{\mathbb{V}(m)}}}\widehat{m_1}, \tag{2.2}$$

and let z_p denote the p quantile of standard normal distribution. Then we have proved

Theorem 1: Under null hypothesis:

$$\lim_{n\to\infty} P(|T| > z_{1-\alpha/2}) = \alpha$$

for any $\alpha \in (0,1)$.

Theorem 2 Under any fixed alternative, $m_1 \neq 0$

$$\lim_{n\to\infty} P(|T| > z_{1-\alpha/2}) = 1$$

for any $\alpha \in (0,1)$. Theorem 2 is a corollary of Theorem 4.

Generalized case

Description

Suppose the data is clustered in known groups of size m_i, $X_{1i}, X_{2i}, ...X_{m_ii}$ are discrete random variables taking values 1 or 2 and are defined as follows $\mathbb{P}(X_i = 0) = 1 - \lambda$, $\mathbb{P}(X_i = 1) = \lambda$, where $i = 1, 2, ...m_i$;

$$var \begin{pmatrix} X_{1i} \\ X_{2i} \\ ... \\ X_{m_ii} \end{pmatrix} = \begin{pmatrix} 1 & \theta & ... & \theta \\ \theta & 1 & ... & \theta \\ ... & ... & ... & \\ \theta & \theta & ... & 1 \end{pmatrix} \lambda(1-\lambda),$$

where $\theta \in [0,1]$

Then define $Z_{1i}, Z_{1i}, ...Z_{(m_i+1)i}$ be iid standard normal, then

$$\begin{aligned} Y_{1i}|X_{1i} = x &= \mu x + \sigma Z_{1i} + \tau Z_{(m_i+1)i} \\ Y_{2i}|X_{2i} = x &= \mu x + \sigma Z_{2i} + \tau Z_{(m_i+1)i} \\ ... \quad & ... \quad ... \\ Y_{m_ii}|X_{m_ii} = x &= \mu x + \sigma Z_{m_ii} + \tau Z_{(m_i+1)i} \end{aligned}$$

We divide analysis into three cases:

Case 1 When n is finite but $\min m_i$ goes to infinity. In real case study, this indicates that the number of biological pathways is finite, but size of each biological pathway approaches infinity.

For any fixed i, there is no constant limit of $\frac{1}{m_i}\sum_{j=1}^{m_i} Y_{ji}$, just use a simple example to illustrate

Without loss of generality, suppose $\min_{1\leq i\leq n} m_i$ is m_1.Suppose $\theta = 1$, if $X_{11} = X_{21} = ... = X_{m_1 1} = 0$, then $Y_{11}, Y_{21}, ...Y_{m_1 1}$ are $N(0, \sigma^2)$ random variables; While if $X_{11} = X_{21} = ... = X_{m_1 1} = 1$, then $Y_{11}, Y_{21}, ...Y_{m_1 1}$ are $N(\mu, \sigma^2)$ random variables. Thus $m_i^{-1}\sum_{j=1}^{m_i} Y_{ji}$ cannot converge to $\lambda\mu$, precluding a test based on such convergence.

Case 2 When all $m_i's$ and n are all goes to infinity, which indicates that both the number of subjects and the biological pathways become infinity large;

Here we assume that all $m_i's$ are equal to m, then

$$\frac{Y_{1i}+Y_{2i}+...+Y_{mi}}{m} = \frac{\sigma(Z_{1i}+Z_{2i}+...Z_{mi})}{m} + \mu\frac{X_{1i}+X_{2i}+...X_{mi}}{m} + \tau Z_{(m+1)i} \quad (2.3)$$

Since

$$\begin{array}{ll} \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\sigma Z_{ji} & \text{is} \quad N(0, \sigma^2/nm) \\ \frac{1}{n}\sum_{i=1}^{n}\tau Z_{(m+1)i} & \text{is} \quad N(0, \tau^2/n), \end{array}$$

since $\mu\mathbb{E}(\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\mu X_{ji}) = \lambda$ and $var(\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\mu X_{ji}) = \frac{\mu^2}{nm}\lambda(1-\lambda)(1-\theta) + \frac{\mu^2}{n}\theta\lambda(1-\lambda)$,

$$\sqrt{n}(\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\mu X_{ji} - \mu\lambda) \xrightarrow{d} N(0, \mu^2\theta\lambda(1-\lambda)/m).$$

Then

$$\sqrt{n}(\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}Y_{ji} - \mu\lambda) \xrightarrow{d} N(0, \mu^2\theta\lambda(1-\lambda) + \tau^2)$$

Case 3 When $m_i's$ are all bounded, but n goes to infinity;

This case is the one we focus on. First, define:

$$\begin{array}{rcl} \widehat{m_1^c} & = & n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m_i} m_i^{-1}Y_{ji} \\ \widehat{m_2^c} & = & n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m_i} m_i^{-1}Y_{ji}^2 \end{array} \quad (2.4)$$

and

$$\begin{array}{rcccl} m_1^c & = & \mathbb{E}(\widehat{m_1^c}) & = & \mu\lambda \\ m_2^c & = & \mathbb{E}(\widehat{m_2^c}) & = & 1+\mu^2\lambda \end{array}$$

In order to get a test with the best local power , we consider adding a weight to each cluster. For a weight w_i, define

$$\widehat{m^*} = n^{-1}\sum_{i=1}^{n} w_i(Y_{1i}+...+Y_{m_i i}). \tag{2.5}$$

We have

$$m^* = \mathbb{E}(\widehat{m^*}) = n^{-1}\sum_{i=1}^{n} w_i\mu\lambda \tag{2.6}$$

and

$$\begin{aligned} var(\widehat{m^*}) &= var(n^{-1}\textstyle\sum_{i=1}^n w_i(Y_{1i}+...+Y_{m_i i})) \\ &= n^{-2}\textstyle\sum_{i=1}^n w_i^2[m_i var(Y_{11}) + m_i(m_i-1)cov(Y_{11},Y_{12})] \\ &= n^{-2}\textstyle\sum_{i=1}^n w_i^2 m_i(\mu^2\lambda(1-\lambda)+1) + n^{-2}\sum_{i=1}^n w_i^2 m_i(m_i-1)(\mu^2\theta\lambda(1-\lambda)+\tau^2) \\ &= n^{-2}\textstyle\sum_{i=1}^n w_i^2 P_i \end{aligned} \tag{2.7}$$

where $P_i = m_i(\mu^2\lambda(1-\lambda)+1) + m_i(m_i-1)(\mu^2\theta\lambda(1-\lambda)+\tau^2)$.

We may estimate thus as

$$\begin{aligned} \widehat{var(\widehat{m^*})} &= n^{-2}\textstyle\sum_{i=1}^n \widehat{w}_i{}^2 m_i(\widehat{\mu}^2\widehat{\lambda}(1-\widehat{\lambda})+1) + n^{-2}\sum_{i=1}^n \widehat{w}_i{}^2 m_i(m_i-1)(\widehat{\mu}^2\theta\widehat{\lambda}(1-\widehat{\lambda})+\tau^2) \\ &= n^{-2}\textstyle\sum_{i=1}^n \widehat{w}_i{}^2\widehat{P}_i, \end{aligned} \tag{2.8}$$

where $\widehat{P}_i = m_i(\widehat{\mu}^2\widehat{\lambda}(1-\widehat{\lambda})+1) + m_i(m_i-1)(\widehat{\mu}^2\theta\widehat{\lambda}(1-\widehat{\lambda})+\tau^2)$ and $\widehat{w}_i$ is to be determined.

We write $\widehat{w}_i$ instead of w_i since as we will see, $\widehat{w}_i$ will depend on $\widehat{P}_i$.

How to choose the weight w_i, first we need to calculate the local power for testing $H_0 : \mu\lambda = 0$ vs. $H_a : \mu\lambda \neq 0$. We set the local alternative to be $H_a^* : \mu\lambda = cn^{-1/2}$, where c is a constant and $c \in (0,\infty)$. Let $T = \frac{\widehat{m^*}}{\sqrt{var(\widehat{m^*})}}$ Then under the locally alternative,

$$\begin{aligned}
\mathbb{P}(|T| > 1.96) \quad &= \mathbb{P}(T - \tfrac{cn^{-1/2} * n^{-1}\sum_{i=1}^n w_i m_i}{\sqrt{var(\widehat{m^*})}} > 1.96 - \tfrac{cn^{-1/2} * n^{-1}\sum_{i=1}^n w_i m_i}{\sqrt{var(\widehat{m^*})}}) \\
&+\mathbb{P}(T - \tfrac{cn^{-1/2} * n^{-1}\sum_{i=1}^n w_i m_i}{var(\sqrt{m^*})} < -1.96 - \tfrac{cn^{-1/2} * n^{-1}\sum_{i=1}^n w_i m_i}{\sqrt{var(\widehat{m^*})}}) \\
&\approx \Phi(-1.96 - \tfrac{cn^{-1/2} * n^{-1}\sum_{i=1}^n w_i m_i}{\sqrt{var(\widehat{m^*})}}) + 1 - \Phi(1.96 - \tfrac{cn^{-1/2} * n^{-1}\sum_{i=1}^n w_i m_i}{\sqrt{var(\widehat{m^*})}}).
\end{aligned} \tag{2.9}$$

Next, choose w_i to maximize

$$\frac{n^{-1}\sum_{i=1}^n w_i m_i}{\sqrt{var(\widehat{m^*})}}.$$

Since

$$\frac{n^{-1}\sum_{i=1}^n w_i m_i}{\sqrt{var(m^*)}} = \frac{\sum_{i=1}^n w_i m_i}{\sqrt{\sum_{i=1}^n w_i^2 P_i}}$$

, then, if $0 \leq w_i \leq 1$ and $\sum_{i=1}^n w_i = 1$, define

$$f(\underset{\sim}{w}) = \frac{\sum_{j=1}^n w_j^2 P_j}{(\sum_{j=1}^n w_j m_j)^2},$$

$$F(\underset{\sim}{w}) = \frac{\sum_{j=1}^n w_j^2 P_j}{(\sum_{j=1}^n w_j m_j)^2} + \lambda(1 - \sum_{i=1}^n w_i)$$

Here $F(\underset{\sim}{w})$ is defined for optimization by Lagrange multiplier. Next, try to minimize $F(\underset{\sim}{w})$ with respect to w_i. Take derivative with respect to w_i and setting to zero:

$$\begin{aligned}
\frac{\partial F}{\partial w_i} \quad &= \quad \tfrac{2w_i P_i}{(\sum_{j=1}^n w_j m_j)^2} - \tfrac{2\sum_{i=1}^n w_j^2 P_j m_i}{(\sum_{j=1}^n w_j m_j)^3} - \lambda = 0 \\
\Rightarrow w_i \quad &= \quad \tfrac{m_i}{P_i}\tfrac{\sum_{j=1}^n w_j^2 P_j}{\sum_{j=1}^n w_j m_j} + \tfrac{\lambda(\sum_{j=1}^n w_j m_j)^2}{2P_i}
\end{aligned} \quad . \tag{2.10}$$

Since $\sum_{i=1}^n w_i = 1$, then

$$\Rightarrow \lambda = \frac{1 - \frac{\sum_{j=1}^n w_j^2 P_j}{\sum_{j=1}^n w_j m_j} \sum_{j=1}^n \frac{m_j}{P_j}}{(\sum_{j=1}^n w_j m_j)^2 \sum_{j=1}^n \frac{1}{2P_j}}. \tag{2.11}$$

If set $w_i = \frac{m_i}{P_i \sum \frac{m_i}{P_i}}$ take back to equation 2.11, we have

$$\lambda = \frac{1 - \frac{1}{\sum_{j=1}^n \frac{m_i}{P_i}} \sum_{j=1}^n \frac{m_i}{P_i}}{(\sum_{j=1}^n \frac{m_j^2}{P_j \sum \frac{m_i}{P_i}})^2 \sum_{j=1}^n \frac{1}{2P_j}} = 0.$$

Then equation 2.10 satisfying:

$$\frac{\partial F}{\partial w_i} = \frac{2 w_i P_i}{(\sum_{j=1}^n w_j m_j)^2} - \frac{2 \sum_{i=1}^n w_j^2 P_j m_i}{(\sum_{j=1}^n w_j m_j)^3} = 0.$$

Thus, the solution $w_i = \frac{m_i}{P_i \sum \frac{m_i}{P_i}}$ minimizes the equation $F(\underset{\sim}{w})$.

We could use the moment estimator of the unknown parameters,

$$\widehat{\mu} = \frac{\widehat{m_2^c} - 1}{\widehat{m_1^c}} \quad \text{and} \quad \widehat{\lambda} = \frac{\widehat{m_1^c}^2}{\widehat{m_2^c} - 1}$$

Then define

$$\widehat{w_i} = \frac{m_i}{\widehat{P_i} \sum \frac{m_i}{\widehat{P_i}}}, \tag{2.12}$$

where $\widehat{P_i} := m_i(\widehat{\mu^2}\widehat{\lambda}(1 - \widehat{\lambda}) + 1) + m_i(m_i - 1)(\widehat{\mu^2}\theta\widehat{\lambda}(1 - \widehat{\lambda}) + \tau^2)$, and

$$\widehat{var(\widehat{m^*})} := n^{-2} \sum_{i=1}^n \widehat{w_i^2} m_i(\widehat{\mu}^2\widehat{\lambda}(1 - \widehat{\lambda}) + 1) + n^{-2} \sum_{i=1}^n \widehat{w_i^2} m_i(m_i - 1)(\widehat{\mu}^2\theta\widehat{\lambda}(1 - \widehat{\lambda}) + \tau^2). \tag{2.13}$$

Next we discuss the size, power and local power of the test.

To test

$$H_0 : \mu\lambda = 0 \ vs. \ H_a : \mu\lambda \neq 0,$$

we set the test statistic:

$$T = \frac{\widehat{m^*}}{\sqrt{var(\widehat{m^*})}}. \tag{2.14}$$

Since the denominator contains unknown parameters, if we further assume that m_i's are bounded above and have some discrete distribution, we could use $\widehat{var(\widehat{m^*})}$ approximate $var(\widehat{m^*})$, then

$$T \approx \frac{\widehat{m^*}}{\sqrt{\widehat{var(\widehat{m^*})}}}. \tag{2.15}$$

Under null hypothesis $H_0 : \mu\lambda = 0$, If we assume that m_i are all bounded above and have some discrete distribution. Then we have

$$w_i = \frac{1}{1+(m_i-1)\tau^2} / \sum_{i=1}^{n} \frac{1}{1+(m_i-1)\tau^2},$$

$$var(\widehat{m^*}) = n^{-2} \sum_{i=1}^{n} \frac{m_i}{1+(m_i-1)\tau^2} / (\sum_{i=1}^{n} \frac{1}{1+(m_i-1)\tau^2})^2,$$

and

$$\widehat{m^c} = \widehat{\mu}\widehat{\lambda} \xrightarrow{p} 0 = \mu\lambda.$$

By Continuous Mapping Theorem and Slutsky's Theorem,

$$\widehat{\mu^2}\widehat{\lambda}(1-\widehat{\lambda}) + 1 \xrightarrow{p} 1 \quad \text{and} \quad \widehat{\mu^2}\theta\widehat{\lambda}(1-\widehat{\lambda}) + \tau^2 \xrightarrow{p} \tau^2,$$

thus, $\widehat{P_i} \xrightarrow{p} m_i + m_i(m_i+1)\tau^2$ uniformly over i, then

$$\frac{w_i}{\widehat{w_i}} = \frac{\frac{1}{1+(m_i-1)\tau^2} / \sum_{i=1}^{n} \frac{1}{1+(m_i-1)\tau^2}}{m_i / (\widehat{P_i} \sum_{j=1}^{n} \frac{m_j}{\widehat{P_j}})} \xrightarrow{p} 1$$

uniformly over i.

Moreover,

$$\frac{var(\widehat{m^*})}{\widehat{var(\widehat{m^*})}} = \frac{n^{-2}\sum_{i=1}^{n} \widehat{w_i}^2 \widehat{P_i}}{n^{-2}\sum_{i=1}^{n} w_i^2 P_i} \xrightarrow{p} 1,$$

then by Slutsky's Theorem,

$$T \xrightarrow{d} N(0,1). \quad \text{under} \quad H_0.$$

We have thus established

Theorem 3: Under the null hypothesis:

$$\lim_{n\to\infty} \mathbb{P}(|T| > z_{1-\alpha/2}) = \alpha,$$

where $\alpha \in (0,1)$.

Under local alternative $H_a^* : \mu\lambda = cn^{-1/2}$ where $c > 0$.

Note that $\widehat{\mu\lambda} \xrightarrow{p} 0$, then by CMT(Continuous Mapping Theorem) and Slutsky's Theorem,

$$\frac{\widehat{P}_i}{P_i} \xrightarrow{p} 1,$$

and

$$\frac{\sum_{i=1}^n \widehat{w}_i^{\,2}}{\sum_{i=1}^n w_i^2} \xrightarrow{p} 1.$$

note that w_i here are slightly different than on last page.

Furthermore,

$$\frac{\widehat{var(\widehat{m^*})}}{var(\widehat{m^*})} \xrightarrow{p} 1.$$

Therefore,

$$\begin{aligned}
\mathbb{P}(|T| > z_{1-\alpha/2}) \quad &= \mathbb{P}(\frac{\widehat{m^*}}{\sqrt{\widehat{var(\widehat{m^*})}}} - \frac{cn^{-1/2}n^{-1}\sum_{i=1}^n w_i m_i}{\sqrt{var(\widehat{m^*})}} \sqrt{\frac{var(\widehat{m^*})}{\widehat{var(\widehat{m^*})}}} \\
&> z_{1-\alpha/2} + \frac{cn^{-1/2}n^{-1}\sum_{i=1}^n w_i m_i}{\sqrt{var(\widehat{m^*})}} \sqrt{\frac{var(\widehat{m^*})}{\widehat{var(\widehat{m^*})}}}) \\
&+\mathbb{P}(\frac{\widehat{m^*}}{\sqrt{\widehat{var(\widehat{m^*})}}} - \frac{cn^{-1/2}n^{-1}\sum_{i=1}^n w_i m_i}{\sqrt{var(\widehat{m^*})}} \sqrt{\frac{var(\widehat{m^*})}{\widehat{var(\widehat{m^*})}}} \\
&< -z_{1-\alpha/2} + \frac{cn^{-1/2}n^{-1}\sum_{i=1}^n w_i m_i}{\sqrt{var(\widehat{m^*})}} \sqrt{\frac{var(\widehat{m^*})}{\widehat{var(\widehat{m^*})}}}).
\end{aligned} \tag{2.16}$$

Moreover,

$$\begin{aligned}\frac{cn^{-1/2}n^{-1}\sum_{i=1}^{n} w_i m_i}{\sqrt{var(\widehat{m^*})}} &= \frac{cn^{-1/2}n^{-1}\sum_{i=1}^{n} w_i m_i}{\sqrt{n^{-2}\sum_{i=1}^{n} w_i^2[m_i(\mu^2\lambda(1-\lambda)+1)+m_i(m_i-1)(\mu^2\theta\lambda(1-\lambda)+\tau^2)]}} \\ &= \frac{c\sum_{i=1}^{n}\frac{m_i^2}{(\sum_{j=1}^{n}\frac{m_j}{P_j})P_i}}{\sqrt{n\sum_{i=1}^{n}\frac{m_i^2}{(\sum_{j=1}^{n}\frac{m_j}{P_j})^2 P_i}}}.\end{aligned} \tag{2.17}$$

Put $M := \sum_{j=1}^{n}\frac{m_j}{P_j}$, we have

$$\begin{aligned}\frac{cn^{-1/2}n^{-1}\sum_{i=1}^{n} w_i m_i}{\sqrt{var(\widehat{m^*})}} &= \frac{c\frac{1}{M}\sum_{i=1}^{n}\frac{m_i^2}{P_i}}{\frac{1}{M}\sqrt{n\sum_{i=1}^{n}\frac{m_i^2}{P_i}}} \\ &= c\sqrt{n^{-1}\sum_{i=1}^{n}\frac{m_i^2}{P_i}}.\end{aligned} \tag{2.18}$$

where $P_i = m_i(\mu^2\lambda(1-\lambda)+1) + m_i(m_i-1)(\mu^2\theta\lambda(1-\lambda)+\tau^2)$.

Since we assume $m_i's$ are bound above for each i, and if we further assume that m_i has some distribution: for example

$$\mathbb{P}(m_i = 2) = \mathbb{P}(m_i = 3) = \frac{1}{2}.$$

Thus, $m_i's$ are iid random variables and $\mathbb{E}(m_i) = \frac{5}{2}$. Then, under the assumption about m_i,

$$n^{-1}\sum_{i=1}^{n}\frac{m_i^2}{P_i} \xrightarrow{p} \frac{1}{2}\frac{4}{2+2\tau^2} + \frac{1}{2}\frac{9}{3+6\tau^2}$$

some finite number. Thus establish

Theorem 4: Under the local alternative and the assumption about m_i above:

$$\begin{aligned}\lim_{n\to\infty}\mathbb{P}(|T| > z_{1-\alpha/2}) &= \Phi(-z_{1-\alpha/2} - c\sqrt{\tfrac{1}{2}\tfrac{4}{2+2\tau^2} + \tfrac{1}{2}\tfrac{9}{3+6\tau^2}}) \\ &\quad +1 - \Phi(z_{1-\alpha/2} - c\sqrt{\tfrac{1}{2}\tfrac{4}{2+2\tau^2} + \tfrac{1}{2}\tfrac{9}{3+6\tau^2}}) \\ &\approx 1 - \Phi(z_{1-\alpha/2} - c\sqrt{\tfrac{1}{2}\tfrac{4}{2+2\tau^2} + \tfrac{1}{2}\tfrac{9}{3+6\tau^2}}),\end{aligned} \tag{2.19}$$

which is between 0 and 1. Here $\alpha \in (0,1)$.

Under a fixed alternative $H_1 : \mu\lambda = \mu_1\lambda_1$, where $\lambda_1 \in (0,1]$ and $\mu_1 > 0$

Since under our assumption about m_i,

$$n^{-1}\textstyle\sum_{i=1}^{n}\frac{m_i^2}{P_i} \quad \xrightarrow{p} \frac{1}{2}\frac{4}{2(\mu_1^2\lambda_1(1-\lambda_1)+1)+2(\mu_1^2\theta\lambda_1(1-\lambda_1)+\tau^2)}$$
$$+\frac{1}{2}\frac{9}{3(\mu_1^2\lambda_1(1-\lambda_1)+1)+6(\mu_1^2\theta\lambda_1(1-\lambda_1)+\tau^2)}$$

Then

$$\sum_{i=1}^{n}\frac{m_i^2}{P_i}\xrightarrow{p}\infty$$

Thus we proved :

Theorem 5: Under a fixed alternative, since $\sqrt{\frac{var(\widehat{m^*})}{\widehat{var(\widehat{m^*})}}} \to 1$.

$$\begin{aligned}
\mathbb{P}(|T|>z_{1-\alpha/2}) \quad &= \mathbb{P}(\frac{m^*}{\sqrt{\widehat{var(m^*)}}} - \frac{\mu_1\lambda_1 n^{-1}\sum_{i=1}^{n} w_i}{\sqrt{var(m^*)}}\sqrt{\frac{var(m^*)}{\widehat{var(m^*)}}} \\
&> z_{1-\alpha/2} - \frac{\mu_1\lambda_1 n^{-1}\sum_{i=1}^{n} w_i}{\sqrt{var(m^*)}}\sqrt{\frac{var(m^*)}{\widehat{var(m^*)}}}) \\
&+\mathbb{P}(\frac{m^*}{\sqrt{\widehat{var(\widehat{m^*})}}} - \frac{\mu_1\lambda_1 n^{-1}\sum_{i=1}^{n} w_i}{\sqrt{\widehat{var(\widehat{m^*})}}}\sqrt{\frac{var(m^*)}{\widehat{var(\widehat{m^*})}}} \\
&< -z_{1-\alpha/2} - \frac{\mu_1\lambda_1 n^{-1}\sum_{i=1}^{n} w_i}{\sqrt{var(\widehat{m^*})}}\sqrt{\frac{var(\widehat{m^*})}{\widehat{var(\widehat{m^*})}}}) \qquad (2.20)\\
&= \Phi(-z_{1-\alpha/2} - \mu_1\lambda_1\sqrt{\textstyle\sum_{i=1}^{n}\frac{m_i^2}{P_i}}\sqrt{\frac{var(\widehat{m^*})}{\widehat{var(\widehat{m^*})}}}) \\
&+1-\Phi(z_{1-\alpha/2} - \mu_1\lambda_1\sqrt{\textstyle\sum_{i=1}^{n}\frac{m_i^2}{P_i}}\sqrt{\frac{var(\widehat{m^*})}{\widehat{var(\widehat{m^*})}}}) \\
&\to \Phi(-\infty)+1-\Phi(-\infty) = 0+1-0 = 1
\end{aligned}$$

Some special cases

1. If $m_i's$ are equal to m

Then,

$$w_1 = w_2 = ... = w_n = \frac{1}{n}$$

$$\widehat{m^*} = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ji}$$

$$\widehat{var(\widehat{m^*})} = n^{-3}(m(\widehat{\mu}^2\widehat{\lambda}(1-\widehat{\lambda})+1)+m(m-1)(\widehat{\mu}^2\theta\widehat{\lambda}(1-\widehat{\lambda})+\tau^2))$$

$$T = \frac{m^*}{\sqrt{\widehat{var(\widehat{m^*})}}} = \frac{n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ji}}{\sqrt{m(\widehat{\mu}^2\widehat{\lambda}(1-\widehat{\lambda})+1)+m(m-1)(\widehat{\mu}^2\theta\widehat{\lambda}(1-\widehat{\lambda})+\tau^2)}} \qquad (2.21)$$

2. If there are no correlations, i.e: $\theta = 0$, $\tau = 0$

Then,

$$w_i = \widehat{w}_i = \frac{m_i}{m_i(\widehat{\mu}^2\widehat{\lambda}(1-\widehat{\lambda})+1)\sum_{i=1}^{n}\frac{m_i}{m_i(\widehat{\mu}^2\widehat{\lambda}(1-\widehat{\lambda})+1)}} = \frac{1}{n}$$

$$\widehat{m^*} = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ji}$$

$$\widehat{var(\widehat{m^*})} = n^{-4}\sum_{i=1}^{n} m_i(\widehat{\mu}^2\widehat{\lambda}(1-\widehat{\lambda})+1)$$

$$T = \frac{m^*}{\sqrt{\widehat{var(\widehat{m^*})}}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ji}}{\sqrt{(\widehat{\mu}^2\widehat{\lambda}(1-\widehat{\lambda})+1)\sum_{i=1}^{n} m_i}} \tag{2.22}$$

2.4 Simulations

We did size and power simulations for general case, and special cases when $m_i's$ are all equal to m(here we set m=5).

For special case:

For size simulation, we estimate the rejection rate under the null hypothesis. We take $Y_{1i}, Y_{2i}, ...Y_{mi}$, $i = 1, 2, .., n$ from normal distribution with

$$var\begin{pmatrix} Y_{1i} \\ Y_{2i} \\ ... \\ Y_{mi} \end{pmatrix} = \begin{pmatrix} 1 & \tau^2 & ... & \tau^2 \\ \tau^2 & 1 & ... & \tau^2 \\ ... & & & \\ \tau^2 & \tau^2 & ... & 1 \end{pmatrix}.$$

Here we take $\tau^2 = 0.3$ and $\tau^2 = 0.6$ respectively, sample size n from 20 to 1500.For each of various sample size, we generate 2000 sets of normal data. Next, we calculate how many times out of 2000, we reject H_0 based on theoretical critical value as estimated size of the test. As shown in figure 2.1 and figure 2.2, the estimated size all fall in the band of 0.05(±0.01)(the nominal rejection rate is 0.05), this result is satisfactory.

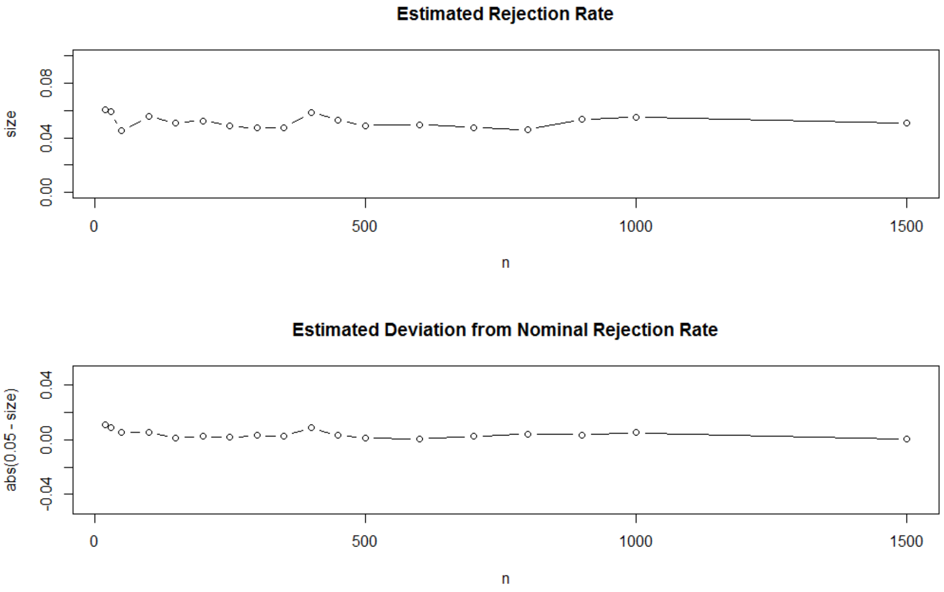

Figure 2.1: Simulation Size 2000, special case,$\tau^2 = 0.3$,$\sigma^2 = 0.7$

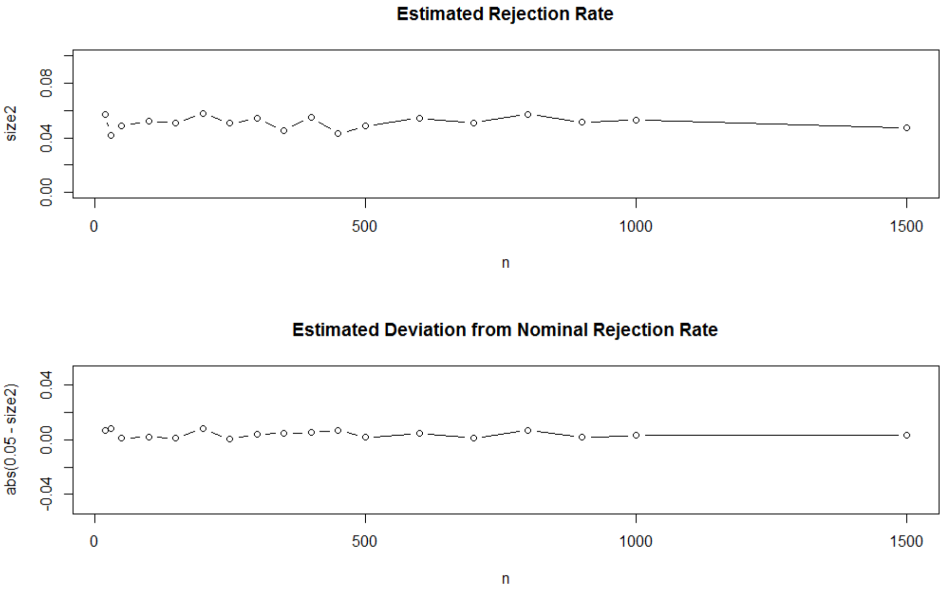

Figure 2.2: Simulation Size 2000, special case,$\tau^2 = 0.6$,$\sigma^2 = 0.4$

The power simulation concerns the behavior of the test under a fixed alternative. We generate data from a contaminated normal mixture distribution. We take sample size $n = 500$, μ varies from 0.01 to 1.5 and λ is various increments from 0.01 to 1 by

0.1.

First we generate $X_{1i}, ...X_{mi}$ as correlated binary data with correlation $\theta = 0.3$, next generate $Z_{1i}...Z_{mi}$ from $N(0, \sigma^2)$ where $\sigma^2 = 0.4$, and $Z_{(m+1),i}$ from $N(0, \tau^2)$ where $\tau^2 = 0.6$. Thus let $Y_{ji} = \mu X_{ji} + Z_{ji} + Z_{(m+1)i}$ where $i = 1, 2, ..., n$, $j = 1, 2, ..., m$, $Y_{1i}, Y_{2i}, ...Y_{mi}$ are from $(1 - \lambda)N(0, 1) + \lambda N(\mu, 1)$

We estimate the rates that we reject H_0 based on theoretical critical value as estimated power of the test. As shown in figure 2.3, for a fixed λ away from 0, when we increase μ, or for a fixed $\mu > 0.3$ when we increase λ, the power goes to 1. This is believable since for μ and λ away from 0, the probability of rejecting H_0 should go to 1.

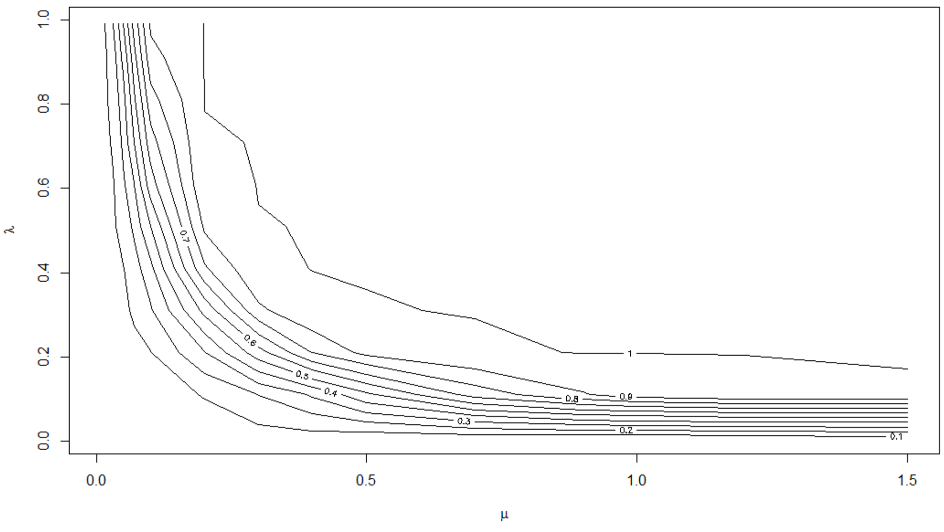

Figure 2.3: Simulation Size 2000, special case,$\tau^2 = 0.3$, $\sigma^2 = 0.7$

Next, we take $\tau^2 = 0.6$ and $\sigma^2 = 0.4$, other settings remain the same, the power simulation is shown in figure 2.4, which is also satisfactory.

For general case:

For size simulation, $m_i's$ are chosen to be 2 or 3 each with probability $\frac{1}{2}$. We take n from 100 to 1500, for each of various sample sizes, we generate 2000 sets of normal data. We take $\lambda = 0.3$, $\theta = 0.3$ $\sigma^2 = 0.4$ and $\tau^2 = 0.6$.

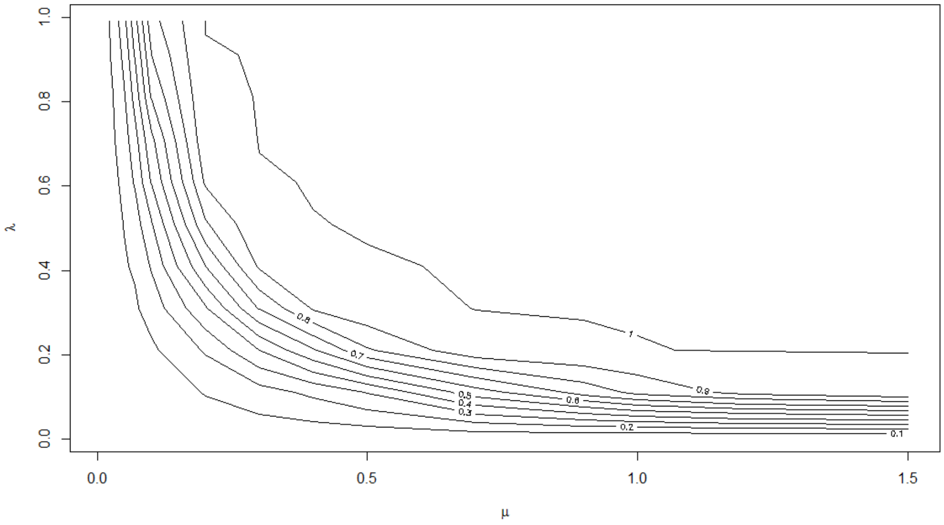

Figure 2.4: Simulation Size 2000, special case,$\tau^2 = 0.6$, $\sigma^2 = 0.4$

As show in figure 2.5, the sizes fall in the band of 0.05(±0.01)(the nominal rejection rate is 0.05), which is satisfactory.

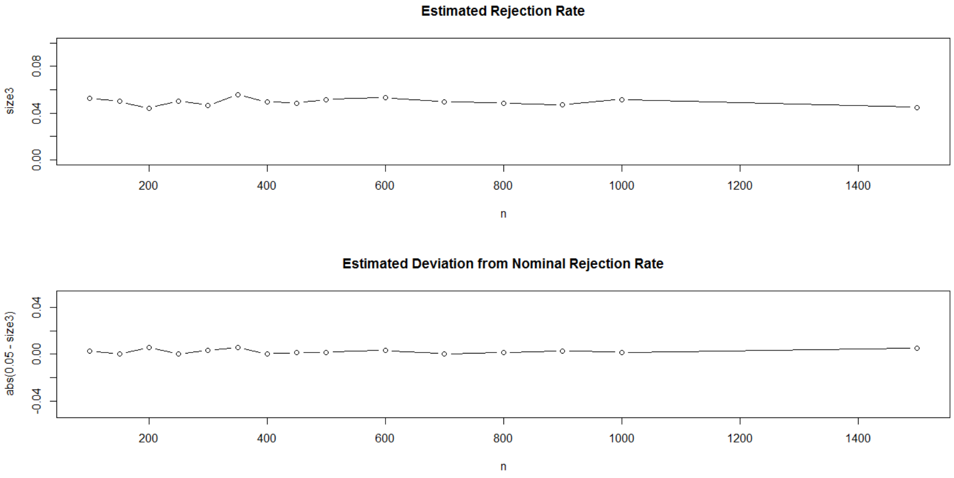

Figure 2.5: Simulation Size 2000, general case,$\tau^2 = 0.3$, $\sigma^2 = 0.7$

Again, we tried $\tau^2 = 0.6$, $\sigma^2 = 0.4$, while other settings remain the same. The plot is shown in figure 2.6.

For power simulation, $m_i's$ are chosen to be either 2 or 3 with probability $\frac{1}{2}$. We take sample size n to be 500, μ from 0.01 to 1.3 in various increments and λ from 0

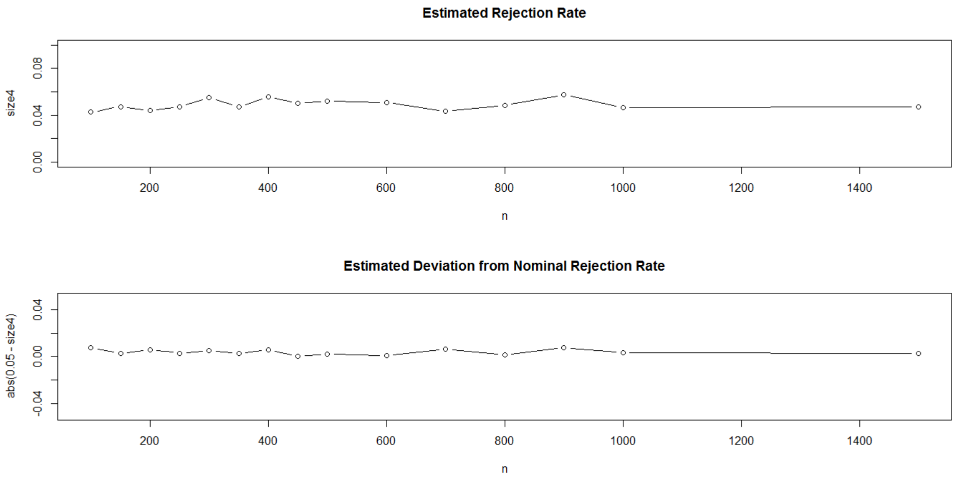

Figure 2.6: Simulation Size 2000, general case,$\tau^2 = 0.6$, $\sigma^2 = 0.4$

to 1 by 0.1.

The result shown in figure 2.7 is satisfactory, since the power of the test goes to 1 when $\mu\lambda$ is away from 0.

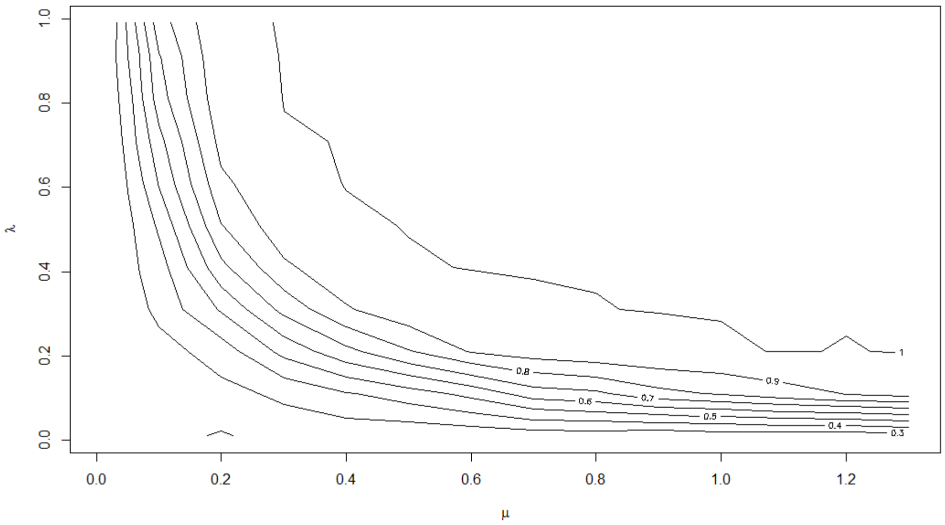

Figure 2.7: Simulation Size 2000, general case,$\tau^2 = 0.3$, $\sigma^2 = 0.7$

Next, we change $\tau^2 = 0.6$ and $\sigma^2 = 0.4$, the contour plot is shown in figure 2.8. As we can see, similarly the power will go to 1 when $\mu\lambda$ is away from 0, and since

σ^2 is smaller, which means that there is less between subject correlations, then the effective sample size becomes larger than 2.7, thus the power goes to 1 faster.

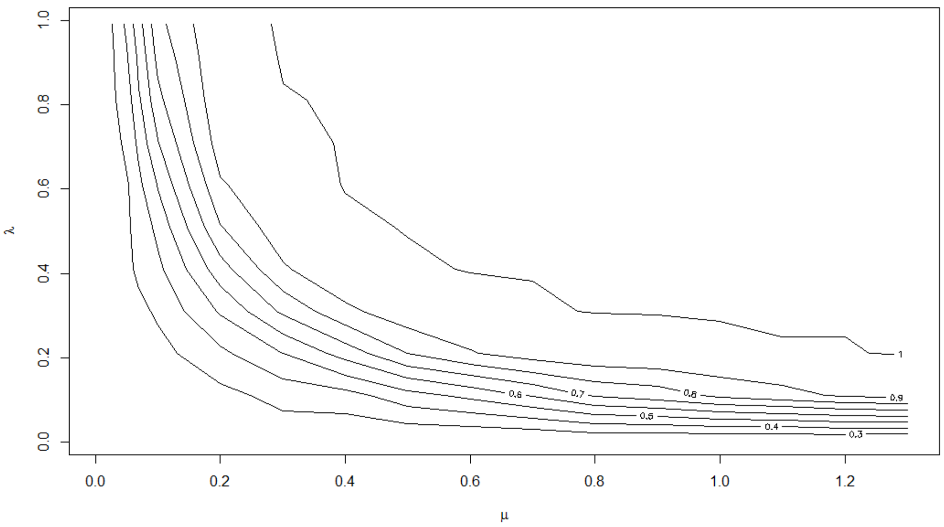

Figure 2.8: Simulation Size 2000, general case,$\tau^2 = 0.6$, $\sigma^2 = 0.4$

2.5 Real data application

For this section, we analyze microarray data from Mao et al. [2005], data can be download from http://www.partek.com. According to Mao's description, Down's syndrome is caused by an extra copy of chromosome 21, so we examine chromosome 21 only. In the data set, there are four samples from four human subjects with Down's syndrome from cerebral tissue, as well as seven samples without Down's syndrome from four human subjects. There are in total 251 genes of interest.

For each of the 251 genes, we apply method mentioned in Charnigo et al. [2013]. First, we fit a linear mixed model $Y_{ij} = \beta_0+\beta_1 x_i+\alpha_i+\varepsilon_{ij}$ where $i = 1, 2, ..., 8; j = 1, 2$ for each of the 251 genes. Here Y_{ij} denotes the gene expression level in sample j from subject i, and x_i is the indicator of Down's syndrome: $x_i = 1$ if subject i has Down's syndrome, otherwise, $x_i = 0$. By using the R function lme in nlme package, we can get 251 T-statistics of testing $\beta_1 = 0$. For each of the 251 T-statistics, we transformed them through T cumulative distribution function(cdf) and the inverse standard normal cdf to get 251 Z-statistics.

Next, we use EM algorithm get the fitted contaminated normal mixture model $(1-\widehat{\lambda})N(0,1)+\widehat{\lambda}N(\widehat{\mu},1)$ with $(\widehat{\lambda},\widehat{\mu})=(0.29,2.41)$. Figure 3.16 shows the histogram of Z-statistics, fitted standard normal curve, and fitted contaminated normal mixture model with $(\widehat{\lambda},\widehat{\mu})=(0.29,2.41)$ when we assume that there is no correlation among the Y's.

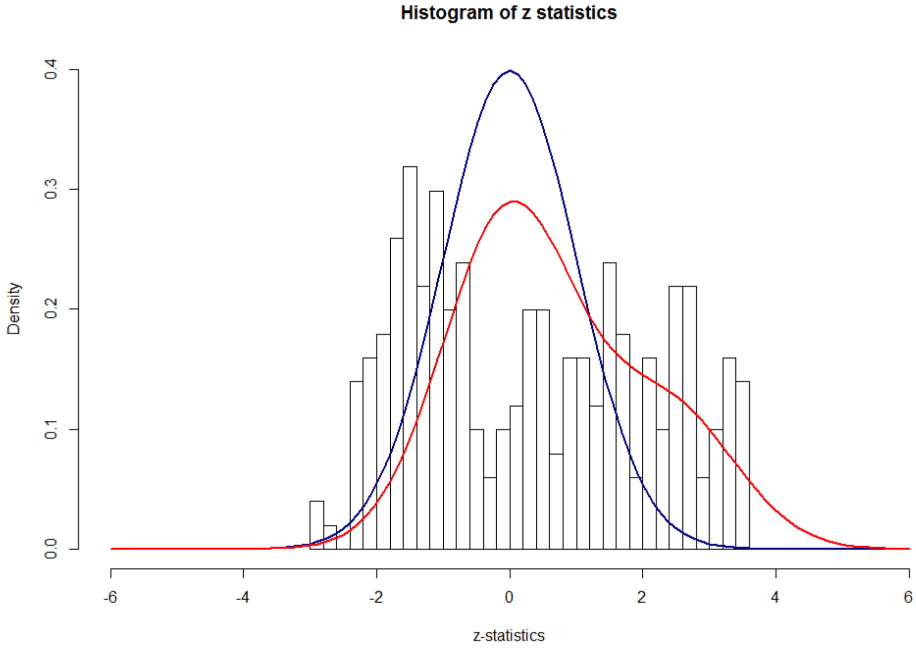

Figure 2.9: Histogram of Z-statistics, blue line is fitted standard normal curve, red line is fitted contaminated normal mixture model with $(\widehat{\lambda},\widehat{\mu})=(0.29,2.41)$

To construct the test, we first need to group the 251 Z-statistics. We separate those Z-statistics into 6 groups according to different Chromosomal locations of Chromosome 21(q21,q22.1,q22.2,q22.3,q23,other locations). The group size are $(m_1,m_2,m_3,m_4,m_5,m_6)=(20,54,23,122,23,9)$ and $n=6$. Here we assume that the correlation structure within each group is known as compound symmetric with known θ and τ^2. Figure 2.10 shows the contour plots of P-values. As we can see in the plot, the contour are close to the straight lines, and as we increase θ and τ^2, the p-value also increases. We use a red dashed line to separate the region of accepting and rejecting $H_0 : \mu\lambda = 0$ at $\alpha = 0.05$.

The result is reasonable since when we increase θ and τ, the effective sample size decreases, so the power of the test is decreasing. This also indicates that if we ignore the correlation structure of some correlated data, we may obtain some wrong inferences. This result is consistent with many published articles, for example, Goeman and Bühlmann [2007] show by simulation that, for some gene expression models which are based on independence assumption between genes, the P-values derived can be wildly anti-conservative. Moreover, as we can see in Figure 2.9, the contaminated normal mixture model is also not a good fit for the data. Note that $\widehat{\lambda}$ and $\widehat{\mu}$ as obtained previously do assume independence: $\tau^2 = 0$ and $\theta = 0$. But for the test, when we increase τ^2 and θ, $\widehat{\lambda}$ and $\widehat{\mu}$ will change. To update $\widehat{\lambda}$ and $\widehat{\mu}$ for different τ^2 and θ, we use optim in R to minimize the negative likelihood, and using $(\widehat{\lambda}, \widehat{\mu}) = (0.29, 2.41)$ as initial values. The rationale is that $\widehat{\mu}\widehat{\lambda}$ should be consistently estimated permitting its substitution for $\mu\lambda$ in the denominator of a moment based test statistics. As we can see if we update $\widehat{\mu}$ and $\widehat{\lambda}$, the shape of the contour does not change noticeably, only the slope changes slightly. Comparing Figure 2.10 and Figure 2.11, we see that failing to take into account correlation may massively understate the p-value, but failing to adjust for correlation in estimation of μ and λ(even when taking correlation into account for the test statistic) may slightly understate the p-value.

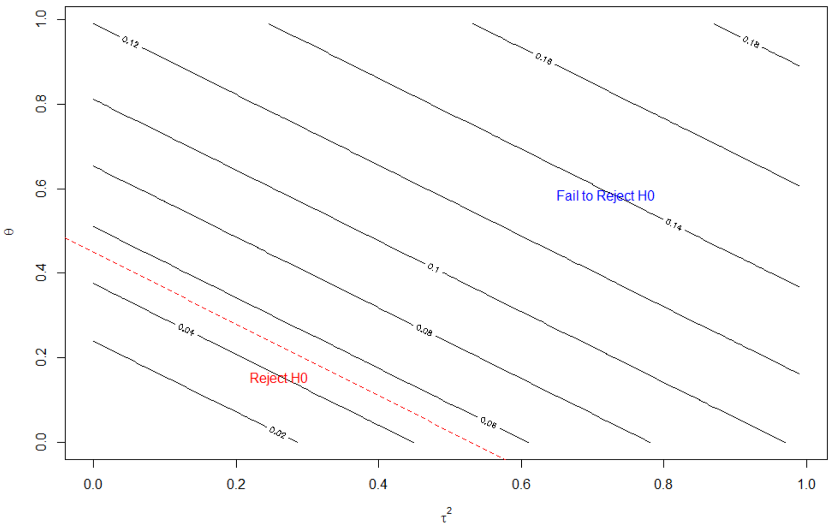

Figure 2.10: Contour plot of P-values with $\widehat{\tau^2}$ and $\widehat{\theta}$ fixed

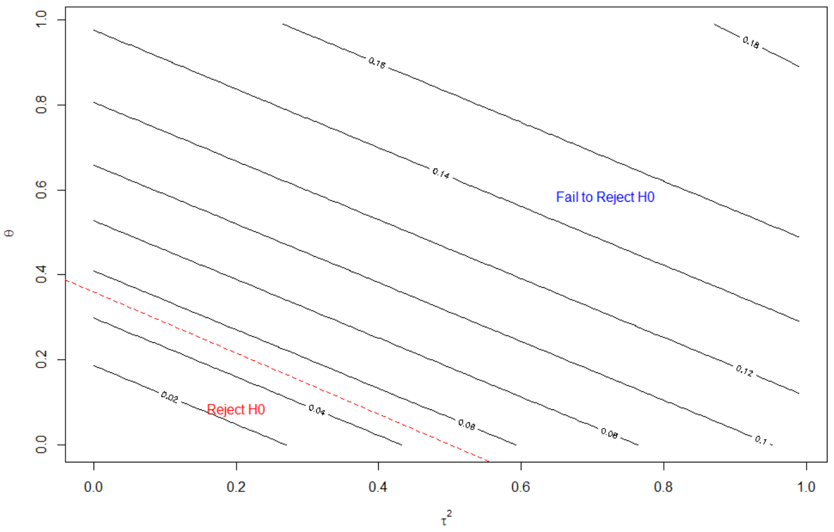

Figure 2.11: Contour plot of P-values with $\widehat{\tau^2}$ and $\widehat{\theta}$ changing with τ^2 and θ

Chapter 3 Three component normal model under correlation

In chapter 2, we discussed contaminated mixture modeling under correlation. As we mentioned in the application section, a two components normal mixture is not a good fit of the presumably correlated Z-values, this motivates us to expand the method to testing 2 versus 3 component when data is correlated under some known correlation structure.

3.1 Three component normal model under correlation for paired data

First, we show the structure of three component normal model under correlated paired data. Let X_{1j}, X_{2j}, j=1,2,...,n, be random variables taking values from {0, 1, -1} and with probability

$$\mathbb{P}(X_{ij}=1)=\lambda_1, \quad \mathbb{P}(X_{ij}=-1)=\lambda_2, \quad \mathbb{P}(X_{ij}=0)=1-\lambda_1-\lambda_2,$$

where i=1, 2 and j=1, 2, ..., n. And the correlation between X_{1j}, X_{2j} is θ which is treated as known.

Next, consider the joint probability distributions of X_{1j} and X_{2j}. Define

$$\begin{aligned}
a=\mathbb{P}(X_{1j}=0,X_{2j}=0) \quad b=\mathbb{P}(X_{1j}=0,X_{2j}=1) \quad c=\mathbb{P}(X_{1j}=0,X_{2j}=-1)\\
b=\mathbb{P}(X_{1j}=1,X_{2j}=0) \quad d=\mathbb{P}(X_{1j}=1,X_{2j}=1) \quad f=\mathbb{P}(X_{1j}=1,X_{2j}=-1)\\
c=\mathbb{P}(X_{1j}=-1,X_{2j}=0) \quad f=\mathbb{P}(X_{1j}=-1,X_{2j}=1) \quad e=\mathbb{P}(X_{1j}=-1,X_{2j}=-1).
\end{aligned}$$

We have the equations:

$$a+b+c=1-\lambda_1-\lambda_2, \quad b+d+f=\lambda_1, \quad c+f+e=\lambda_2$$

The specific probabilities are

$$\begin{gathered}
a=(1-\lambda_1-\lambda_2)-(1-\theta)(1-\lambda_1-\lambda_2)(\lambda_1+\lambda_2)\\
b=(1-\theta)(1-\lambda_1-\lambda_2)\lambda_1 \quad c=(1-\theta)(1-\lambda_1-\lambda_2)\lambda_2\\
d=\lambda_1^2(1-\theta)+\theta\lambda_1 \quad e=\lambda_2^2(1-\theta)+\theta\lambda_2 \quad f=\lambda_1\lambda_2(1-\theta).
\end{gathered}$$

Define random variables

$$Y_{1j} = X_{1j}(I_{[X_{1j}=1]}\mu_1 + I_{[X_{1j}=-1]}\mu_2) + \sigma Z_{1i} + \tau Z_i^*, \quad Y_{2j} = X_{2j}(I_{[X_{2j}=1]}\mu_1 + I_{[X_{2j}=-1]}\mu_2) + \sigma Z_{2i} + \tau Z_i^*,$$

where Z_{1i}, Z_{2i}, Z_i^* are iid standard normal random variables, μ_1, μ_2, σ, τ are all nonnegative and σ, τ are known with $\sigma^2 + \tau^2 = 1$. Then Y_{1j}, Y_{2j} are correlated random variables distributed as 3 component normal mixture: $(1 - \lambda_1 - \lambda_2)N(0,1) + \lambda_1 N(\mu_1, 1) + \lambda_2 N(-\mu_2, 1)$ and have marginal density function:

$$f(Y_{ij} = y) = \frac{1}{\sqrt{2\pi}}(1 - \lambda_1 - \lambda_2)\exp(-y^2/2) + \lambda_1 \exp(-(y - \mu_1)^2/2) + \lambda_2 \exp(-(y + \mu_2)^2/2) \tag{3.1}$$

3.2 Define moments estimator

To develop a moment based test of 2 vs. 3 components with correlated data, we first need to define the first six moments of the data. Define:

$$\widehat{m_1^c} = n^{-1}\sum_{j=1}^{n}(Y_{1j} + Y_{2j})/2, \quad \widehat{m_2^c} = n^{-1}\sum_{j=1}^{n}(Y_{1j}^2 + Y_{2j}^2)/2, \quad ..., \quad \widehat{m_6^c} = n^{-1}\sum_{j=1}^{n}(Y_{1j}^6 + Y_{2j}^6)/2.$$

Let m_1^c through m_6^c be the expected value of the moments respectively. In the following calculation, we define $X_{ij}I_i := I_{[X_{ij}=1]}\mu_1 + I_{[X_{ij}=-1]}(-\mu_2)$, $X_{ij}^2 I_i^2 := I_{[X_{ij}=1]}\mu_1^2 + I_{[X_{ij}=-1]}(-\mu_2)^2$, $X_{ij}^3 I_i^3 := I_{[X_{ij}=1]}\mu_1^3 + I_{[X_{ij}=-1]}(-\mu_2)^3$ and so forth. Then we have

$$
\begin{aligned}
m_1^c &= n^{-1}\textstyle\sum_{j=1}^n \mathbb{E}((Y_{1j}+Y_{2j})/2) = \mu_1\lambda_1 - \mu_2\lambda_2 \\
m_2^c &= n^{-1}\textstyle\sum_{j=1}^n \mathbb{E}((Y_{1j}^2+Y_{2j}^2)/2) = \mathbb{E}(X_{11}^2 I_1^2 + \sigma^2 Z_{11}^2 + \tau^2 (Z_1^*)^2) = 1 + \mu_1^2\lambda_1 + \mu_2^2\lambda_2 \\
m_3^c &= n^{-1}\textstyle\sum_{j=1}^n \mathbb{E}((Y_{1j}^3+Y_{2j}^3)/2) = \lambda_1(\mu_1^3 + 3\mu_1) - \lambda_2(\mu_2^3 + 3\mu_2) \\
m_4^c &= n^{-1}\textstyle\sum_{j=1}^n \mathbb{E}((Y_{1j}^4+Y_{2j}^4)/2) \\
&= \mathbb{E}(X_{11}^4 I_1^4 + \sigma^4 Z_{11}^4 + \tau^4 (Z_1^*)^4 + 6X_{11}^2 I_1^2 \sigma^2 Z_{11}^2 + 6X_{11}^2 I_1^2 \tau^2 (Z_1^*)^2 + 6\sigma^2 Z_{11}^2 \tau^2 (Z_1^*)^2) \\
&\quad \textit{other terms expected value equal zero} \\
&= \lambda_1\mu_1^4 + \lambda_2\mu_2^4 + 3(\sigma^2+\tau^2)^2 + 6(\sigma^2+\tau^2)(\mu_1^2\lambda_1 + \mu_2^2\lambda_2) \\
&= \lambda_1\mu_1^4 + \lambda_2\mu_2^4 + 6m_2^c - 3 \\
m_5^c &= n^{-1}\textstyle\sum_{j=1}^n \mathbb{E}((Y_{1j}^5+Y_{2j}^5)/2) \\
&= \mathbb{E}(X_{11}^5 I_1^5 + 3X_{11}^3 I_1^3(\sigma^2 Z_{11}^2 + \tau^2 (Z_1^*)^2) + X_{11}^3 I_1^3(\sigma^2 Z_{11}^2 + \tau^2 (Z_1^*)^2) + 3X_{11} I_1(\sigma^4 Z_{11}^4 \\
&+ \tau^4 (Z_1^*)^4) + 30X_{11} I_1 \sigma^2 Z_{11}^2 \tau^2 (Z_1^*)^2 + 2X_{11} I_1(\sigma^4 Z_{11}^4 + \tau^4 (Z_1^*)^4) \\
&+ 6X_{11}^3 I_1^3(\sigma^2 Z_{11}^2 + \tau^2 (Z_1^*)^2)) \\
&\quad \textit{other terms expected value equal zero} \\
&= \mu_1^5\lambda_1 - \mu_2^5\lambda_2 + 10(\mu_1^3\lambda_1 - \mu_2^3\lambda_2) + (15\tau^4 + 30\sigma^2\tau^2) \\
&= \mu_1^5\lambda_1 - \mu_2^5\lambda_2 + 10m_3^c - 15m_1^c \\
m_6^c &= n^{-1}\textstyle\sum_{j=1}^n \mathbb{E}((Y_{1j}^6+Y_{2j}^6)/2) \\
&= \mathbb{E}(X_{11}^6 I_1^6 + \sigma^6 Z_{11}^6 + \tau^6 (Z_1^*)^6 + 15(\sigma^2 Z_{11}^2 + \tau^2 (Z_1^*)^2) X_{11}^4 I_1^4 + 15(\sigma^4 Z_{11}^4 + \tau^4 (Z_1^*)^4) X_{11}^2 I_1^2 \\
&+ 90(\sigma^2 Z_{11}^2 + \tau^2 (Z_1^*)^2) X_{11}^2 I_1^2 + 15\sigma^4 Z_{11}^4 + 15\sigma^2 Z_{11}^2 + \tau^4 (Z_1^*)^4) \\
&\quad \textit{other terms expected value equal zero} \\
&= \mu_1^6\lambda_1 + \mu_2^6\lambda_2 + 15(\mu_1^4\lambda_1 + \mu_2^4\lambda_2) + 45(\sigma^2+\tau^2)^2(\mu_1^2\lambda_1 + \mu_2^2\lambda_2) + 15(\sigma^2+\tau^2)^3 \\
&= \mu_1^6\lambda_1 + \mu_2^6\lambda_2 + 15m_4^c - 45m_2^c + 15.
\end{aligned}
\tag{3.2}
$$

3.3 Hypothesis Testing

The test of 2 vs. 3 components is equivalent to testing $H_0 : \lambda_1\mu_1\lambda_2\mu_2 = 0$ vs. $H_a : \lambda_1\mu_1\lambda_2\mu_2 \neq 0$. Then if $-\lambda_1\mu_1\lambda_2\mu_2(\mu_1+\mu_2)^2 = 0$ holds, that implies that H_0 is true. Then let the test statistic be(Charnigo et al. [2013])

$$M = \widehat{m_2^c}^2 - 2\widehat{m_2^c} - 1 + 3\widehat{m_1^c}^2 - \widehat{m_3^c}\widehat{m_1^c}, \tag{3.3}$$

we have

$$\mathbb{E}(M) = -\lambda_1\mu_1\lambda_2\mu_2(\mu_1 + \mu_2)^2 = 0 \tag{3.4}$$

if H_0 holds. Note that the test statistic contains first three moments. Next, consider a variance-covariance matrix of the first three moments.Define

$$V_1 = cov\begin{pmatrix} (Y_{11}+Y_{21})(Y_{11}+Y_{21}) & (Y_{11}+Y_{21})(Y_{11}^2+Y_{21}^2) & (Y_{11}+Y_{21})(Y_{11}^3+Y_{21}^3) \\ (Y_{11}+Y_{21})(Y_{11}^2+Y_{21}^2) & (Y_{11}^2+Y_{21}^2)(Y_{11}^2+Y_{21}^2) & (Y_{11}^2+Y_{21}^2)(Y_{11}^3+Y_{21}^3) \\ (Y_{11}+Y_{21})(Y_{11}^3+Y_{21}^3) & (Y_{11}^2+Y_{21}^2)(Y_{11}^3+Y_{21}^3) & (Y_{11}^3+Y_{21}^3)(Y_{11}^3+Y_{21}^3) \end{pmatrix}.$$

Then

$$V = var\begin{pmatrix} \widehat{m_1^c} \\ \widehat{m_2^c} \\ \widehat{m_3^c} \end{pmatrix} = \frac{1}{4}V_1/n.$$

We have

$$\begin{aligned}
cov(Y_{11}+Y_{21},Y_{11}+Y_{21}) &= var(Y_{11}+Y_{21}) = 2var(Y_{11}) + 2cov(Y_{11},Y_{21})\\
&= 2[\mathbb{E}(Y_{11}^2) - (\mathbb{E}Y_{11})^2] + 2[\mathbb{E}(Y_{11}Y_{21}) - \mathbb{E}(Y_{11})\mathbb{E}(Y_{21})]\\
&= 2[m_2^c - (m_1^c)^2] + 2[d\mu_1^2 - 2\mu_1\mu_2 f + e\mu_2^2 + \tau^2 - (m_1^c)^2]\\
&= 2[(1-\theta)(\mu_1^2\lambda_1^2 + \mu_2^2\lambda_2^2 - 2\mu_1\mu_2\lambda_1\lambda_2) + \theta(\mu_1^2\lambda_1 + \mu_2^2\lambda_2)\\
&+\tau^2 - (m_1^c)^2] + 2[m_2^c - (m_1^c)^2]\\
&= 2[\theta m_2^c - \theta m_1^{c2} + \tau^2 - \theta] + 2[m_2^c - (m_1^c)^2]\\
&= 2(1+\theta)(m_2^c - (m_1^c)^2) + 2(\tau^2 - \theta)\\
cov(Y_{11}+Y_{21},Y_{11}^2+Y_{21}^2) &= \mathbb{E}((Y_{11}+Y_{21})(Y_{11}^2+Y_{21}^2)) - \mathbb{E}(Y_{11}+Y_{21})\mathbb{E}(Y_{11}^2+Y_{21}^2)\\
&= \mathbb{E}(Y_{11}^3 + Y_{11}^2Y_{21} + Y_{21}^2Y_{11} + Y_{21}^3) - 4m_2^c m_1^c\\
&= 2\mathbb{E}(Y_{11}Y_{21}^2) + 2m_3^c - 4m_2^c m_1^c\\
&= 2\mathbb{E}(X_{11}I_1X_{21}^2I_2^2 + \sigma^2 Z_{21}^2 X_{11}I_1 + \tau^2(Z^*)^2X_{11}I_1 + 2X_{21}I_2(Z^*)^2\tau^2)\\
&+2m_3^c - 4m_2^c m_1^c \; \textit{other terms expected value equal zero}\\
&= 2[\mu_1^3 d + \mu_1\mu_2^2 f - \mu_1^2\mu_2 f - e\mu_2^3 + (2\tau^2+1)m_1^c] + 2m_3^c - 4m_2^c m_1^c\\
&= 2[(1-\theta)(m_2^c - 1)m_1^c + \theta(m_3^c - 3m_1^c) + (1+2\tau^2)m_1^c]\\
&+2m_3^c - 4m_2^c m_1^c\\
&= 2(1+\theta)(m_3^c - m_2^c m_1^c) + 4m_1^c(\tau^2 - \theta)\\
cov(Y_{11}+Y_{21},Y_{11}^3+Y_{21}^3) &= \mathbb{E}((Y_{11}+Y_{21})(Y_{11}^3+Y_{21}^3)) - \mathbb{E}(Y_{11}+Y_{21})\mathbb{E}(Y_{11}^3+Y_{21}^3)\\
&= \mathbb{E}(Y_{11}^4 + Y_{21}^4 + Y_{11}Y_{21}^3 + Y_{11}^3Y_{21}) - 4m_3^c m_1^c\\
&= 2\mathbb{E}(Y_{11}Y_{21}^3) + 2m_4^c - 4m_3^c m_1^c\\
&= 2\mathbb{E}(X_{11}I_1X_{21}^3I_2^3 + 3X_{11}I_1X_{21}I_2(\sigma^2Z_{21}^2 + \tau^2(Z_1^*)^2) + \tau^4(Z_1^*)^4\\
&+3X_{21}^2I_2^2\tau^2(Z_1^*)^2 + 3\sigma^2Z_{21}^2\tau^2(Z_1^*)^2) + 2m_4^c - 4m_3^c m_1^c\\
&\;\textit{other terms expected value equal zero}
\end{aligned}$$

$$\begin{aligned}
&= 2(\mu_1^4 d - \mu_1\mu_2^3 f - \mu_1^3\mu_2 f + \mu_1^4 e \\
&+3(\sigma^2+\tau^2)(\mu_1^2 d - 2\mu_1\mu_2 f + \mu_2^2 e) + 3\tau^4 \\
&+3(\mu_1^2\lambda_1 + \mu_2^2\lambda_2)\tau^2 + 3\tau^2\sigma^2) + 2m_4^c - 4m_3^c m_1^c \\
&= 2((1-\theta)(m_3^c m_1^c - 3m_1^c) + \theta(m_4^c - 6m_2^c + 3) + 3(\theta+\tau^2)(m_2^c - 1) \\
&+3(1-\theta)(m_1^c)^2 + 3\tau^2(\tau^2+\sigma^2)) + 2m_4^c - 4m_3^c m_1^c \\
&= 2(1+\theta)(m_4^c - m_3^c m_1^c) - 6(\tau^2-\theta)m_2^c \\
cov(Y_{11}^2 + Y_{21}^2, Y_{11}^2 + Y_{21}^2) &= var(Y_{11}^2 + Y_{21}^2) = 2var(Y_{11}^2) + 2cov(Y_{11}^2, Y_{21}^2) \\
&= 2[\mathbb{E}(Y_{11}^4) - (\mathbb{E}(Y_{11}^2))^2] + 2[\mathbb{E}(Y_{11}^2 Y_{21}^2) - \mathbb{E}(Y_{11}^2)\mathbb{E}(Y_{21}^2)] \\
&= 2\mathbb{E}(Y_{11}^2 Y_{21}^2) + 2(m_4^c - (m_2^c)^2) - 4(m_2^c)^2 \\
&= 2\mathbb{E}(X_{11}^2 I_1^2 X_{21}^2 I_2^2 + (\sigma^2 Z_{21}^2 + \tau^2 (Z_1^*)^2) X_{11}^2 I_1^2 + (\sigma^2 Z_{11}^2 \\
&+\tau^2 (Z_1^*)^2) X_{21}^2 I_2^2 + 2\sigma^2\tau^2 (Z_1^*)^2 Z_{11}^2 + \sigma^4 Z_{11}^2 Z_{21}^2 + \tau^4 (Z_1^*)^4 \\
&+4X_{11} I_1 X_{21} I_2 \tau^2 (Z^*)^2 + 2(m_4^c - (m_2^c)^2) - 4(m_2^c)^2 \\
&\ \textit{other terms expected value equal zero} \\
&= 2[\mu_1^4 d + 2\mu_1^2\mu_2^2 f + \mu_2^4 e + 2(\sigma^2+\tau^2)(\mu_1^2\lambda_1 + \mu_2^2\lambda_2) + \sigma^4 + 2\sigma^2\tau^2 \\
&+3\tau^4 + 4\tau^2(\mu_1^2 d - 2\mu_1\mu_2 f + \mu_2^2 e) + 2(m_4^c - (m_2^c)^2) - 4(m_2^c)^2 \\
&= 2(\theta m_4^c - \theta(m_2^c)^2 + 2(\tau^2-\theta)(\tau^2 - 1 + 2m_2^c) + 2\tau^2 \\
&+(\theta-1)(2m_2^c - 2(m_1^c)^2 - 1)) + 2(m_4^c - (m_2^c)^2) \\
&= 2(\theta+1)(m_4^c - (m_2^c)^2) + 4(\tau^2-\theta)(\tau^2 - 1 + 2m_2^c) \\
&+4\tau^2(\theta-1)(2m_2^c - 2(m_1^c)^2 - 1) \\
cov(Y_{11}^2 + Y_{21}^2, Y_{11}^3 + Y_{21}^3) &= \mathbb{E}((Y_{11}^2 + Y_{21}^2)(Y_{11}^3 + Y_{21}^3)) - \mathbb{E}(Y_{11}^2 + Y_{21}^2)\mathbb{E}(Y_{11}^3 + Y_{21}^3) \\
&= \mathbb{E}(Y_{11}^5 + Y_{11}^2 Y_{21}^3 + Y21^2 Y_{11}^3 + Y_{21}^5) + 4m_3^c m_2^c \\
&= 2\mathbb{E}(Y_{11}^2 Y_{21}^3) + 2m_5^c + 4m_3^c m_2^c
\end{aligned}$$

$$
\begin{aligned}
&= 2\mathbb{E}(X_{11}I_1X_{21}^3I_2^3 + 3X_{11}^2I_1^2X_{21}I_2(\sigma^2Z_{21}^2 + \tau^2(Z_1^*)^2) \\
&+X_{11}^3I_1^3(\sigma^2Z_{11}^2 + \tau^2(Z_1^*)^2) + 3X_{21}I_2(\sigma^4Z_{11}^2Z_{21}^2 + \tau^2\sigma^2Z_{11}^2(Z_1^*)^2 \\
&+\tau^2\sigma^2Z_{21}^2(Z_1^*)^2 + \tau^4(Z_1^*)^4) + 2X_{11}I_1\tau^4(Z_1^*)^4 + 6X_{11}I_1X_{21}^2I_2^2\tau^2(Z_1^*)^2 \\
&+6X_{11}I_1\tau^2(Z_1^*)^2\sigma^2Z_{21}^2) + 2m_5^c + 4m_3^cm_2^c \\
&\ other\ terms\ expected\ value\ equal\ zero \\
&= 2[\mu_1^5d - \mu_1^2\mu_2^3f + \mu_1^3\mu_2^2f - \mu_2^5e + (3+6\tau^2)(\mu_1^3d \\
&-\mu_1^2\mu_2f + \mu_1\mu_2^2f - \mu_2^3e + \mu_1^3\lambda + 1 \\
&-\mu_2^3\lambda_2 + 3(\sigma^4 + 4\sigma^2\tau^2 + 5\tau^4)(\mu_1\lambda_1 - \mu_2\lambda_2)] + 2m_5^c + 4m_3^cm_2^c \\
&= 2(\theta+1)(m_5^c - m_3^cm_2^c) + 12\theta(\tau^2-1)(m_3^c - m_1^c) + 12\tau^2(1 \\
&-\theta)m_1^cm_2^c + 12\tau^2(\tau^2-\theta)m_1^c \\
cov(Y_{11}^3 + Y_{21}^3, Y_{11}^3 + Y_{21}^3) &= var(Y_{11}^3 + Y_{21}^3) = 2var(Y_{11}^3) + 2cov(Y_{11}^3, Y_{21}^3) \\
&= 2[\mathbb{E}(Y_{11}^6) - (\mathbb{E}(Y_{11}^3))^2] + 2[\mathbb{E}(Y_{11}^3Y_{21}^3) - \mathbb{E}(Y_{11}^3)\mathbb{E}(Y_{21}^3)] \\
&= 2\mathbb{E}(Y_{11}^3Y_{21}^3) + 2(m_6^c - (m_3^c)^2) - 2(m_3^c)^2 \\
&= 2\mathbb{E}(X_{11}^3I_3X_{21}^3I_2^3 + 3(\sigma^2Z_{11}^2 + \tau^2Z_1^{*2})X_{11}^3I_1^3X_{21}I_2 + \tau^6Z_1^{*6} \\
&+3X_{21}^2I_2^2\tau^4Z_1^{*4} + 3\sigma^2Z_{21}^2\tau^4Z_1^{*4} \\
&+3X_{11}^2I_1^2\tau^4Z_1^{*4} + 9X_{11}^2I_1^2X_{21}^2I_2^2\tau^4Z_1^{*4} \\
&+9X_{11}^2I_1^2\sigma^2Z_{21}^2\tau^2Z_1^{*2} + 3(\sigma^2Z_{11}^2 \\
&+\tau^2Z_1^{*2})X_{11}I_1X_{21}^3I_2^3 + 9X_{11}I_1X_{21}I_2(\sigma^4Z_{11}^2Z_{21}^2 + 2\sigma^2Z_{21}^2\tau^2Z_1^{*2} \\
&+\tau^4Z_1^{*4}) + 3X_{21}^2I_2^2\tau^4Z_1^{*4}\sigma^2Z_{11}^2 + 3\tau^4Z_1^{*4}\sigma^2Z_{11}^2 \\
&+9\sigma^4Z_{11}^2Z_{21}^2\tau^2Z_1^{*2}) + 2(m_6^c - (m_3^c)^2) - 4(m_3^c)^2 \\
&\ other\ terms\ expected\ value\ equal\ zero \\
&= 2(\theta+1)(m_6^c - (m_3^c)^2) + 18\theta(\tau^2-1)m_4^c + 18\tau^2(1-\theta)(m_2^c)^2
\end{aligned}
$$

$$
\begin{aligned}
&+18m_2^c\theta(\tau^4 + 2 - 3\tau^2) \\
&+18\tau^2(1-\theta)(\tau^2+1)(m_1^c)^2 + 12\tau^6 - 12\theta \\
&-18\tau^4\theta + 18\tau^2\theta
\end{aligned}
$$

Then, we could apply multivariate delta method to get the variance of test statistics M: $var(M) = G^{'}VG$ where G is the matrix of the partial derivative, and defined as:

$$G = \begin{pmatrix} \frac{\partial \mathbb{E}M}{\partial m_1^c} \\ \frac{\partial \mathbb{E}M}{\partial m_2^c} \\ \frac{\partial \mathbb{E}M}{\partial m_3^c} \end{pmatrix} = \begin{pmatrix} 6m_1^c - m_3^c \\ 2m_2^c - 2 \\ -m_1^c \end{pmatrix}.$$

Then, plug in the 3.2, we have

$$\begin{aligned} var(M) = G'VG &= \begin{pmatrix} \frac{\partial \mathbb{E}M}{\partial m_1^c} \\ \frac{\partial \mathbb{E}M}{\partial m_2^c} \\ \frac{\partial \mathbb{E}M}{\partial m_3^c} \end{pmatrix} \\ &= \begin{pmatrix} 6m_1^c - m_3^c \\ 2m_2^c - 2 \\ -m_1^c \end{pmatrix}' V \begin{pmatrix} 6m_1^c - m_3^c \\ 2m_2^c - 2 \\ -m_1^c \end{pmatrix}. \end{aligned}$$

Then the estimated variance is

$$\widehat{var(M)} = \begin{pmatrix} 6\widehat{m_1^c} - \widehat{m_3^c} \\ 2\widehat{m_2^c} - 2 \\ -\widehat{m_1^c} \end{pmatrix}' \widehat{V} \begin{pmatrix} 6\widehat{m_1^c} - \widehat{m_3^c} \\ 2\widehat{m_2^c} - 2 \\ -\widehat{m_1^c} \end{pmatrix}$$

Define the test statistic as

$$T = \frac{M}{\sqrt{\widehat{var(M)}}} = \frac{\widehat{m_2^c}^2 - 2\widehat{m_2^c} - 1 + 3\widehat{m_1^c}^2 - \widehat{m_3^c}\widehat{m_1^c}}{\sqrt{\begin{pmatrix} 6\widehat{m_1^c} - \widehat{m_3^c} \\ 2\widehat{m_2^c} - 2 \\ -\widehat{m_1^c} \end{pmatrix}' \widehat{V} \begin{pmatrix} 6\widehat{m_1^c} - \widehat{m_3^c} \\ 2\widehat{m_2^c} - 2 \\ -\widehat{m_1^c} \end{pmatrix}}}$$

Under the null hypothesis

By multivariate central limit theory and Cramer Theorem (Ferguson, 1996, p.45),

$$\frac{M - \mathbb{E}M}{\sqrt{var(M)}} \xrightarrow{d} N(0,1).$$

Then by slutsky's theorem,

$$T = \frac{M - \mathbb{E}M}{\sqrt{\widehat{var(M)}}} = \frac{M}{\sqrt{var(M)}} \sqrt{\frac{var(M)}{\widehat{var(M)}}} \rightarrow N(0,1).$$

Thus, we have proved:

Theorem 3.1:Under the null hypothesis,

$$\lim_{n\to\infty} \mathbb{P}(T < -z_{1-\alpha}) = \alpha,$$

where $\alpha \in (0, 1)$ and $z_{1-\alpha}$ is the $1-\alpha$ quantile of a standard normal distribution, then testing procedure is approximately level α. Notice, here we use one sided rejection region, since as we define T makes it no probability to be positive.

Under the local alternative hypothesis

Define the hypothesis as $H_0 : \mu_1\mu_2\lambda_1\lambda_2 = 0$ vs $H_{ak} : \mu_1\mu_2\lambda_1\lambda_2 = \mu_1^*\lambda_1^*\lambda_2^*\mu_{2n}$, where μ_1^*, λ_1^*, λ_2^*, are fixed positive numbers, while $\mu_{2n} = kn^{-1/2}$. Then under the local alternative,

$$\begin{aligned}
\lim\nolimits_{n\to\infty} \mathbb{P}(T < -Z_{1-\alpha/2}) &= \lim\nolimits_{n\to\infty}[\mathbb{P}(T - \frac{kn^{-1/2}\mu_1^*\lambda_1^*\lambda_2^*(\mu_1^*+kn^{-1/2})^2}{\sqrt{\widehat{var(M)}}}\sqrt{\frac{\widehat{var(M)}}{var(M)}} \\
&< -Z_{1-\alpha/2} - \frac{kn^{-1/2}\mu_1^*\lambda_1^*\lambda_2^*(\mu_1^*+kn^{-1/2})^2}{\sqrt{\widehat{var(M)}}}\sqrt{\frac{\widehat{var(M)}}{var(M)}})] \\
&= \lim\nolimits_{n\to\infty} \mathbb{P}[(T - \frac{kn^{-1/2}\mu_1^*\lambda_1^*\lambda_2^*(\mu_1^*+kn^{-1/2})^2}{n^{-1/2}\sqrt{\frac{1}{4}\widehat{G}`\widehat{V_1}\widehat{G}}}\sqrt{\frac{\widehat{var(M)}}{var(M)}} \\
&< -Z_{1-\alpha/2} - \frac{kn^{-1/2}\mu_1^*\lambda_1^*\lambda_2^*(\mu_1^*+kn^{-1/2})^2}{n^{-1/2}\sqrt{\frac{1}{4}\widehat{G}`\widehat{V_1}\widehat{G}}}\sqrt{\frac{\widehat{var(M)}}{var(M)}})] \\
&= \Phi(-Z_{1-\alpha/2}.
\end{aligned}$$

Since $\widehat{G} \xrightarrow{P} G$ and $\widehat{V_1} \xrightarrow{P} V_1$. If $k \to \infty$, then $\lim_{n\to\infty} \mathbb{P}(T < -Z_{1-\alpha/2}) = 1$; If $k \to 0$, then $\lim_{n\to\infty} \mathbb{P}(T < -Z_{1-\alpha/2}) = \Phi(-Z_{1-\alpha/2})) = \alpha$. Thus, the test is asymptotically locally unbiased, and we have proved

Theorem 3.2:Under the local alternative hypothesis $H_{ak} : \mu_1\mu_2\lambda_1\lambda_2 = \mu_1^*\lambda_1^*\lambda_2^*\mu_{2n}$,

$$\lim_{n\to\infty} \mathbb{P}(T < -Z_{1-\alpha/2}) = \Phi(-Z_{1-\alpha/2} + \frac{k\mu_1^*\lambda_1^*\lambda_2^*\mu_1^{*2}}{\sqrt{\frac{1}{4}G^TV_1G}}).$$

A fixed number between α and 1.

Under a fixed alternative hypothesis

Consider a fixed alternative $H_a : \mu_1\mu_2\lambda_1\lambda_2 = \mu_1^*\mu_2^*\lambda_1^*\lambda_2^*$, where μ_1^*,μ_2^*, λ_1^*, λ_2^* are fixed positive numbers. Then

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}(|T| > Z_{1-\alpha/2}) &= \lim_{n\to\infty} \mathbb{P}(T - \frac{\mu_1^*\mu_2^*\lambda_1^*\lambda_2^*(\mu_1^*+\mu_2^*)^2}{\sqrt{\widehat{var(M)}}}\sqrt{\frac{\widehat{var(M)}}{var(M)}} \\
&> Z_{1-\alpha/2} - \frac{\mu_1^*\mu_2^*\lambda_1^*\lambda_2^*(\mu_1^*+\mu_2^*)^2}{\sqrt{\widehat{var(M)}}}\sqrt{\frac{\widehat{var(M)}}{var(M)}}) \\
&+ \lim_{n\to\infty} \mathbb{P}(T - \frac{\mu_1^*\mu_2^*\lambda_1^*\lambda_2^*(\mu_1+\mu_2^*)^2}{\sqrt{\widehat{var(M)}}}\sqrt{\frac{\widehat{var(M)}}{var(M)}} \\
&< -Z_{1-\alpha/2} - \frac{\mu_1^*\mu_2^*\lambda_1^*\lambda_2^*(\mu_1+\mu_2^*)^2}{\sqrt{\widehat{var(M)}}}\sqrt{\frac{\widehat{var(M)}}{var(M)}}) \\
&= \lim_{n\to\infty} \mathbb{P}(T - \frac{\mu_1^*\mu_2^*\lambda_1^*\lambda_2^*(\mu_1+\mu_2^*)^2}{n^{-1/2}\sqrt{\frac{1}{4}\widehat{G`}\widehat{V_1}\widehat{G}}}\sqrt{\frac{\widehat{var(M)}}{var(M)}} \\
&> Z_{1-\alpha/2} - \frac{\mu_1^*\mu_2^*\lambda_1^*\lambda_2^*(\mu_1+\mu_2^*)^2}{n^{-1/2}\sqrt{\frac{1}{4}\widehat{G`}\widehat{V_1}\widehat{G}}}\sqrt{\frac{\widehat{var(M)}}{var(M)}}) \\
&+ \lim_{n\to\infty} \mathbb{P}(T - \frac{\mu_1^*\mu_2^*\lambda_1^*\lambda_2^*(\mu_1+\mu_2^*)^2}{n^{-1/2}\sqrt{\frac{1}{4}\widehat{G`}\widehat{V_1}\widehat{G}}}\sqrt{\frac{\widehat{var(M)}}{var(M)}} \\
&< -Z_{1-\alpha/2} - \frac{\mu_1^*\mu_2^*\lambda_1^*\lambda_2^*(\mu_1+\mu_2^*)^2}{n^{-1/2}\sqrt{\frac{1}{4}\widehat{G`}\widehat{V_1}\widehat{G}}}\sqrt{\frac{\widehat{var(M)}}{var(M)}}) \\
&= \lim_{n\to\infty} \Phi(-Z_{1-\alpha/2} + \tfrac{\mu_1^*\mu_2^*\lambda_1^*\lambda_2^*(\mu_1+\mu_2^*)^2 n^{1/2}}{\sqrt{\frac{1}{4}\widehat{G`}\widehat{V_1}\widehat{G}}}) + 1 \\
&- \lim_{n\to\infty} \Phi(Z_{1-\alpha/2} + \tfrac{\mu_1^*\mu_2^*\lambda_1^*\lambda_2^*(\mu_1+\mu_2^*)^2 n^{1/2}}{\sqrt{\frac{1}{4}\widehat{G`}\widehat{V_1}\widehat{G}}}) = 1.
\end{aligned}$$

Thus, we have proved

Theorem 3.3:Under a fixed alternative hypothesis $H_{ak} : \mu_1\mu_2\lambda_1\lambda_2 = \mu_1^*\mu_2^*\lambda_1^*\lambda_2^*$,

$$\lim_{n\to\infty} \mathbb{P}(|T| > Z_{1-\alpha/2}) = 1.$$

3.4 Simulation study of paired data

We did size and power simulation for this case.

Size simulation

For size simulation, we estimate the rejection rate under the null hypothesis. We take $\mu_1 = 0$, $\mu_2 = 5$, $\lambda_1 = 0.2$, $\lambda_2 = 0.3$, thus the model reduce to 2 component mixture normal: $0.7N(0,1)+0.3N(-5,1)$. take $\tau^2=0.3$ and $\tau^2=0.6$ respectively, sample size n from 20 to 1500.For each of various sample size, we generate 7000 sets of two component normal data. Next, we calculate how many times out of 7000, we reject H_0 based on theoretical critical value as estimated size of the test. As shown in figure 3.1 and figure 3.2, the estimated size all fall around 0.05(the nominal rejection rate is 0.05), this result is satisfactory.

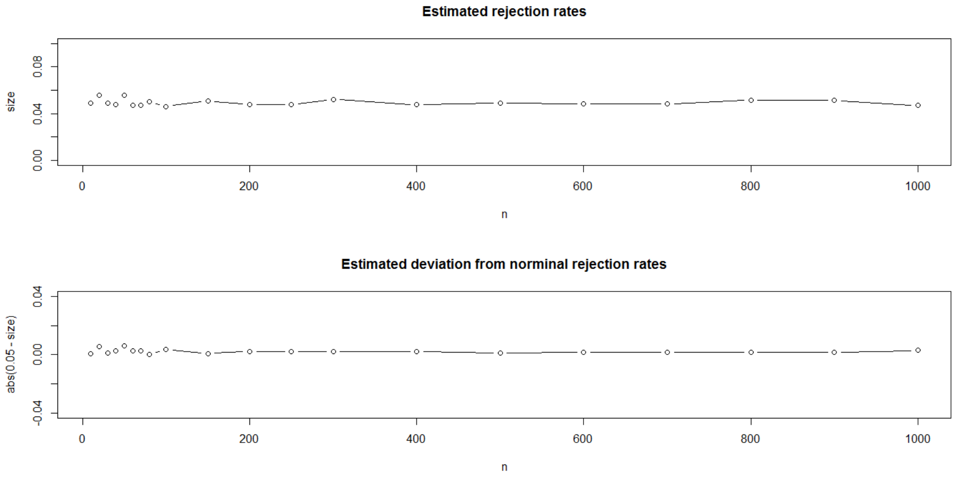

Figure 3.1: Simulation Size 7000, $\tau^2=0.3$,$\sigma^2=0.7$

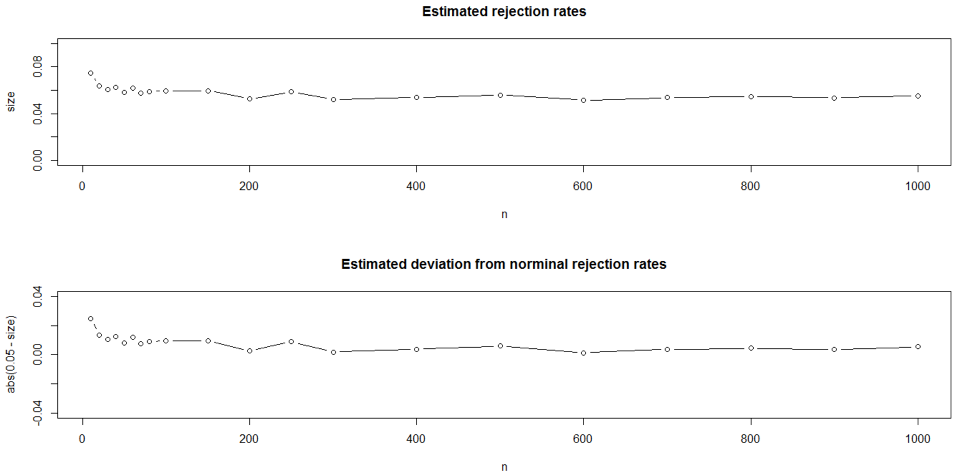

Figure 3.2: Simulation Size 7000,$\tau^2=0.6$,$\sigma^2=0.4$

Power simulation

For power simulation, we generate data from a 3 component normal mixture distribution: $\lambda_1 N(\mu_1, 1) + \lambda_2 N(-\mu_2, 1) + (1 - \lambda_1 - \lambda_2)N(0, 1)$. We take sample size $n = 100$. We simulate data with different θ and τ^2, see figure 3.3 to figure 3.6.

Under each condition, we generate X_{1i}, X_{2i} as correlated data with correlation $\theta \in \{0.2, 0.7\}$, next generate Z_{1i}, Z_{2i} from $N(0, \sigma^2)$ where $\sigma^2 \in \{0.8, 0.3\}$, and Z_i^* from $N(0, \tau^2)$ where $\tau^2 \in \{0.2, 0.7\}$. Thus let $Y_{ji} = (\mu_1 I_{[X_{ji}=1]} - \mu_2 I_{[X_{ji}=-1]}) + Z_{ji} + Z_i^*$ where $i = 1, 2, ..., n$, $j = 1, 2$, Y_{1i}, Y_{2i} are from $\lambda_1 N(\mu_1, 1) + \lambda_2 N(-\mu_2, 1) + (1 - \lambda_1 - \lambda_2)N(0, 1)$.

We estimate the rates that we reject H_0 based on theoretical critical value as estimated power of the test. As shown in figure, for a fixed λ_1(or λ_2) away from 0, when we increase μ_1(or μ_2), or for a fixed μ_1(or μ_2) when we increase λ_1(or λ_2), the power goes to 1. Also, for a fixed μ_1, if we increase μ_2, or for a fixed μ_2, we increase μ_1, the power will also go to 1. This is believable since for $\mu_1 \lambda_1 \mu_2 \lambda_2$ away from 0, the probability of rejecting H_0 should go to 1.

Then if we compare the four figures for different θ and τ^2, we could see that, the power will increase faster to 1 for smaller θ and τ^2 than larger ones. This makes sense since if we increase the correlations, the effective sample size decreases. The effect of changing τ^2, however, appears more pronounced than the effect of changing θ.

3.5 Three components normal model under correlation with group size m

Next we consider a more general case: suppose each cluster has equal size of $m(m > 2)$, then the data is constructed as following. $X_{1i}, X_{2i}, ..., X_{mi}$ are correlated random variables with values from {0, 1, -1} and with probability

$$\mathbb{P}(X_{ij} = 1) = \lambda_1, \quad \mathbb{P}(X_{ij} = -1) = \lambda_2, \quad \mathbb{P}(X_{ij} = 0) = 1 - \lambda_1 - \lambda_2,$$

where i=1, 2,...,m and j=1, 2, ..., n. And the correlation is $\theta > 0$ which is known.

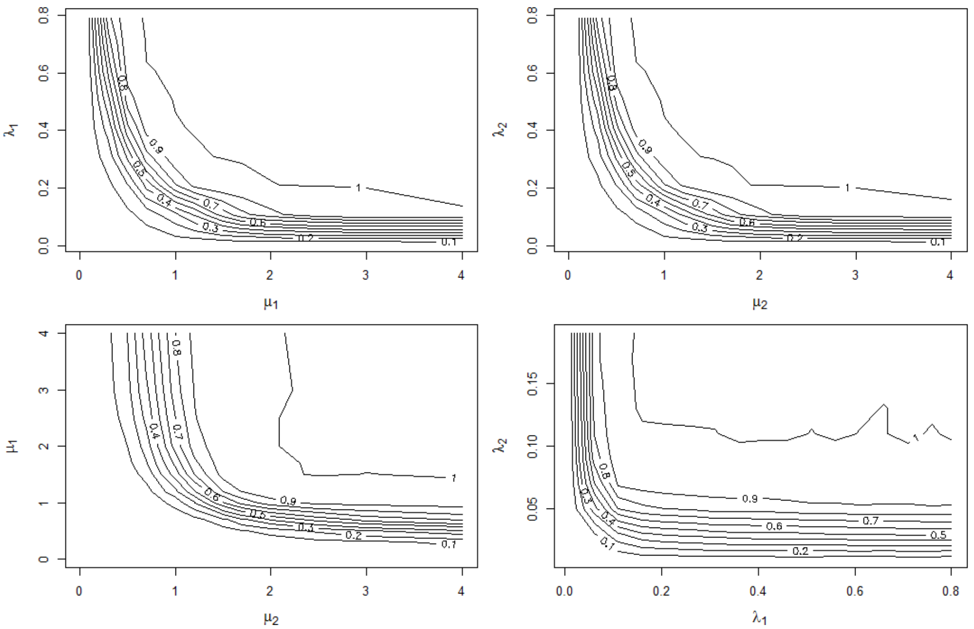

Figure 3.3: Power simulation for $\theta = 0.2$,$\tau^2 = 0.2$

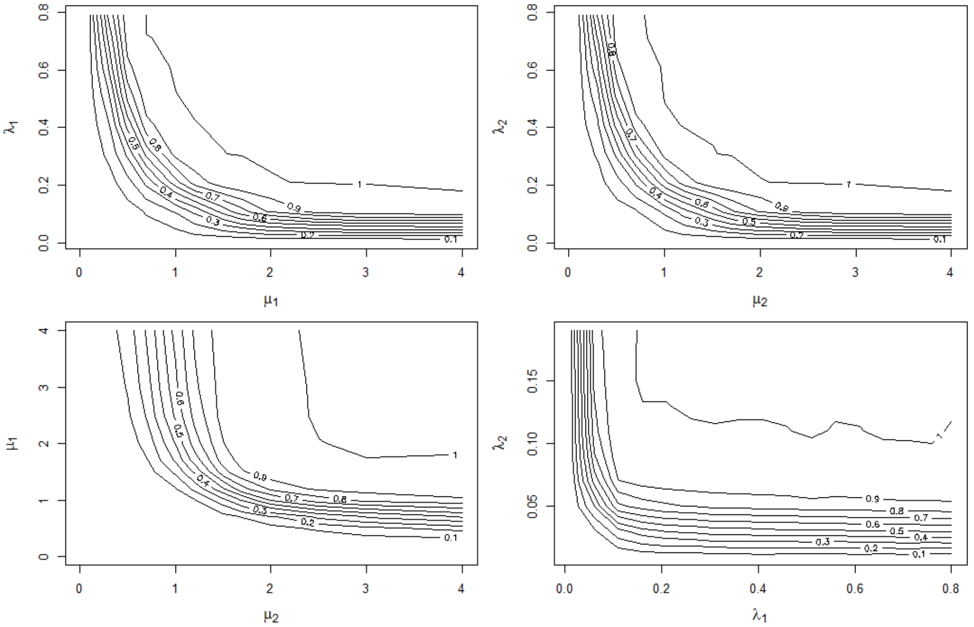

Figure 3.4: Power simulation for $\theta = 0.2$,$\tau^2 = 0.7$

For $i \neq k$, we define the joint probability distributions of X_{ij} and X_{kj} as

$$a = \mathbb{P}(X_{ij} = 0, X_{kj} = 0) \quad b = \mathbb{P}(X_{ij} = 0, X_{kj} = 1) \quad c = \mathbb{P}(X_{ij} = 0, X_{kj} = -1)$$
$$b = \mathbb{P}(X_{ij} = 1, X_{kj} = 0) \quad d = \mathbb{P}(X_{ij} = 1, X_{kj} = 1) \quad f = \mathbb{P}(X_{ij} = 1, X_{kj} = -1)$$
$$c = \mathbb{P}(X_{ij} = -1, X_{kj} = 0) \quad f = \mathbb{P}(X_{ij} = -1, X_{kj} = 1) \quad e = \mathbb{P}(X_{ij} = -1, X_{kj} = -1).$$

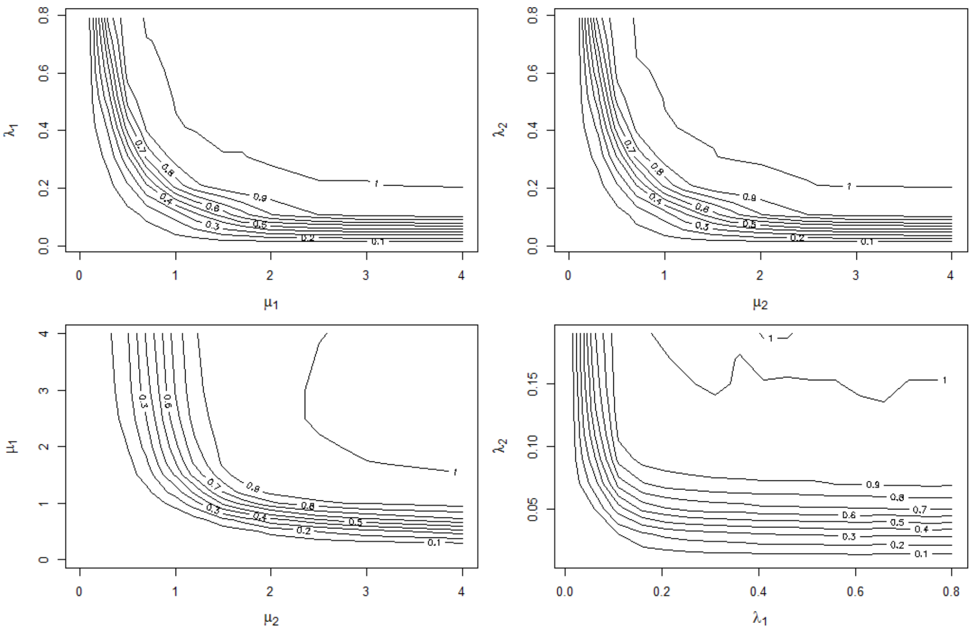

Figure 3.5: Power simulation for $\theta = 0.7$,$\tau^2 = 0.2$

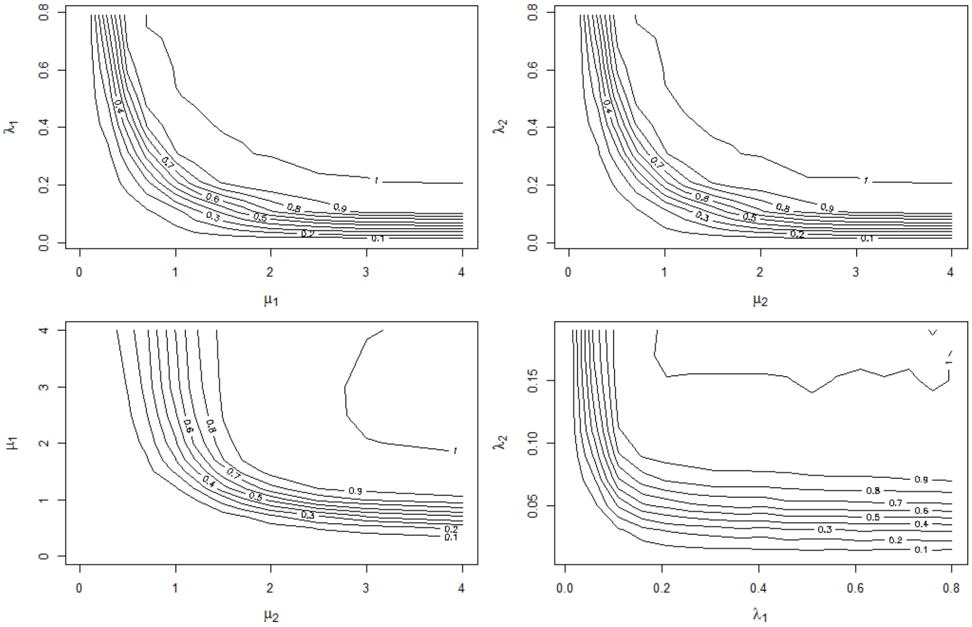

Figure 3.6: Power simulation for $\theta = 0.7$,$\tau^2 = 0.7$

Similar to $m = 2$, we can get

$$a = (1 - \lambda_1 - \lambda_2) - (1 - \theta)(1 - \lambda_1 - \lambda_2)(\lambda_1 + \lambda_2)$$

$$b = (1 - \theta)(1 - \lambda_1 - \lambda_2)\lambda_1 \quad c = (1 - \theta)(1 - \lambda_1 - \lambda_2)\lambda_2$$

$$d = \lambda_1^2(1 - \theta) + \theta\lambda_1 \quad e = \lambda_2^2(1 - \theta) + \theta\lambda_2 \quad f = \lambda_1\lambda_2(1 - \theta).$$

Next, define random variables

$$Y_{ij} = X_{1j}(I_{[X_{ij}=1]}\mu_1 + I_{[X_{ij}=-1]}\mu_2) + \sigma Z_{ij} + \tau Z_j^*,$$

where $i = 1, 2, ..., m, j = 1, 2, ..., n$. Then Y_{1j}, Y_{2j},...,Y_{mj} are correlated random variables distributed as 3 component normal mixture: $(1-\lambda_1-\lambda_2)N(0,1)+\lambda_1 N(\mu_1, 1)+\lambda_2 N(-\mu_2, 1)$.

The moments estimator can be defined as:

$$\widehat{m_1^c} = n^{-1}m^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n} Y_{ij}, \quad \widehat{m_2^c} = n^{-1}m^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n} Y_{ij}^2, \quad ..., \quad \widehat{m_6^c} = n^{-1}m^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n} Y_{ij}^6.$$

Also like the $m = 2$ case, we can get the expected moments as:

$$\begin{aligned}
m_1^c &= n^{-1}m^{-1}\textstyle\sum_{i=1}^m\sum_{j=1}^n \mathbb{E}(Y_{ij}) = \mu_1\lambda_1 - \mu_2\lambda_2 \\
m_2^c &= n^{-1}m^{-1}\textstyle\sum_{i=1}^m\sum_{j=1}^n \mathbb{E}(Y_{ij}^2) = 1 + \mu_1^2\lambda_1 + \mu_2^2\lambda_2 \\
m_3^c &= n^{-1}m^{-1}\textstyle\sum_{i=1}^m\sum_{j=1}^n \mathbb{E}(Y_{ij}^3) = \lambda_1(\mu_1^3 + 3\mu_1) - \lambda_2(\mu_2^3 + 3\mu_2) \\
m_4^c &= n^{-1}m^{-1}\textstyle\sum_{i=1}^m\sum_{j=1}^n \mathbb{E}(Y_{ij}^4) = \lambda_1\mu_1^4 + \lambda_2\mu_2^4 + 6m_2^c - 3 \\
m_5^c &= n^{-1}m^{-1}\textstyle\sum_{i=1}^m\sum_{j=1}^n \mathbb{E}(Y_{ij}^5) = \mu_1^5\lambda_1 - \mu_2^5\lambda_2 + 10m_3^c - 15m_1^c \\
m_6^c &= n^{-1}m^{-1}\textstyle\sum_{i=1}^m\sum_{j=1}^n \mathbb{E}(Y_{ij}^6) = \mu_1^6\lambda_1 + \mu_2^6\lambda_2 + 15m_4^c - 45m_2^c + 15.
\end{aligned}$$

Hypothesis Testing $H_0 : \lambda_1\mu_1\lambda_2\mu_2 = 0$ vs. $H_a : \lambda_1\mu_1\lambda_2\mu_2 \neq 0$. Similar to $m = 2$, define

$$M = \widehat{m_2^c}^2 - 2\widehat{m_2^c} - 1 + 3\widehat{m_1^c}^2 - \widehat{m_3^c}\widehat{m_1^c},$$

then

$$\mathbb{E}(M) = -\lambda_1\mu_1\lambda_2\mu_2(\mu_1 + \mu_2)^2 = 0$$

if H_0 holds. Define

$$V_1 = cov\begin{pmatrix} (\sum_{i=1}^m Y_{i1})(\sum_{i=1}^m Y_{i1}) & (\sum_{i=1}^m Y_{i1})(\sum_{i=1}^m Y_{i1}^2) & (\sum_{i=1}^m Y_{i1})(\sum_{i=1}^m Y_{i1}^3) \\ (\sum_{i=1}^m Y_{i1})(\sum_{i=1}^m Y_{i1}^2) & (\sum_{i=1}^m Y_{i1}^2)(\sum_{i=1}^m Y_{i1}^2) & (\sum_{i=1}^m Y_{i1}^2)(\sum_{i=1}^m Y_{i1}^3) \\ (\sum_{i=1}^m Y_{i1})(\sum_{i=1}^m Y_{i1}^3) & (\sum_{i=1}^m Y_{i1}^2)(\sum_{i=1}^m Y_{i1}^3) & (\sum_{i=1}^m Y_{i1}^3)(\sum_{i=1}^m Y_{i1}^3) \end{pmatrix}.$$

Then

$$V = var\begin{pmatrix} \widehat{m_1^c} \\ \widehat{m_2^c} \\ \widehat{m_3^c} \end{pmatrix} = m^{-2}n^{-1}V_1.$$

After calculation, we have

$$\begin{aligned}
cov(\textstyle\sum_{i=1}^m Y_{i1}, \sum_{i=1}^m Y_{i1}) &= mvar(Y_{11}) + m(m-1)cov(Y_{11}, Y_{21}) \\
&= (m + \theta m(m-1))(m_2^c - (m_1^c)^2) + m(m-1)(\tau^2 - \theta) \\
cov(\textstyle\sum_{i=1}^m Y_{i1}, \sum_{i=1}^m Y_{i1}^2) &= m\mathbb{E}(Y_{11}^3) + m(m-1)\mathbb{E}(Y_{11}Y_{21}^2) - m^2 m_1^c m_2^c \\
&= (m + \theta m(m-1))(m_3^c - m_2^c m_1^c) + m(m-1)m_1^c(\tau^2 - \theta) \\
cov(\textstyle\sum_{i=1}^m Y_{i1}, \sum_{i=1}^m Y_{i1}^3) &= m(m-1)\mathbb{E}(Y_{11}Y_{21}^3) + mm_4^c - m^2 m_3^c m_1^c \\
&= (m + \theta m(m-1))(m_4^c - m_3^c m_1^c) - m(m-1)3(\tau^2 - \theta)m_2^c \\
cov(\textstyle\sum_{i=1}^m Y_{i1}^2, \sum_{i=1}^m Y_{i1}^2) &= mvar(Y_{11}^2) + m(m-1)cov(Y_{11}^2, Y_{21}^2) \\
&= (m + \theta m(m-1))(m_4^c - (m_2^c)^2) + m(m-1)[2(\tau^2 - \theta)(\tau^2 - 1 \\
&\quad +2m_2^c) + 2\tau^2(\theta - 1)(2m_2^c - 2(m_1^c)^2 - 1)] \\
cov(\textstyle\sum_{i=1}^m Y_{i1}^2, \sum_{i=1}^m Y_{i1}^3) &= m(m-1)\mathbb{E}(Y_{11}^2 Y_{21}^3) + mm_5^c + m^2 m_3^c m_2^c \\
&= (m + \theta m(m-1)(m_5^c - m_3^c m_2^c) + m(m-1)6\theta(\tau^2 - 1)(m_3^c - m_1^c) \\
&\quad +m(m-1)6\tau^2(1-\theta)m_1^c m_2^c + m(m-1)6\tau^2(\tau^2 - \theta)m_1^c \\
cov(\textstyle\sum_{i=1}^m Y_{i1}^3, \sum_{i=1}^m Y_{i1}^3) &= mvar(Y_{11}^3) + m(m-1)cov(Y_{11}^3, Y_{21}^3) \\
&= (m + \theta m(m-1))(m_6^c - (m_3^c)^2) + m(m-1)[9\theta(\tau^2 - 1)m_4^c + 9\tau^2(1 \\
&\quad -\theta)(m_2^c)^2 + 9m_2^c\theta(\tau^4 + 2 - 3\tau^2) + 9\tau^2(1-\theta)(\tau^2 \\
&\quad +1)(m_1^c)^2 + 6(\tau^6 - \theta) - 9\tau^4\theta + 9\tau^2\theta]
\end{aligned}$$

$var(M) = G^{'}VG$ where G is the matrix of the partial derivative, and defined as:

$$G = \begin{pmatrix} \frac{\partial \mathbb{E}M}{\partial m_1^c} \\ \frac{\partial \mathbb{E}M}{\partial m_2^c} \\ \frac{\partial \mathbb{E}M}{\partial m_3^c} \end{pmatrix} = \begin{pmatrix} 6m_1^c - m_3^c \\ 2m_2^c - 2 \\ -m_1^c \end{pmatrix}.$$

Note that if we take $m = 2$, the equations still hold, then $m = 2$ is the special case of group size of m.

Then the estimated variance is

$$\widehat{var(M)} = \begin{pmatrix} 6\widehat{m_1^c} - \widehat{m_3^c} \\ 2\widehat{m_2^c} - 2 \\ -\widehat{m_1^c} \end{pmatrix}^{'} \widehat{V} \begin{pmatrix} 6\widehat{m_1^c} - \widehat{m_3^c} \\ 2\widehat{m_2^c} - 2 \\ -\widehat{m_1^c} \end{pmatrix}.$$

Define the test statistic as

$$T = \frac{M}{\sqrt{\widehat{var(M)}}} = \frac{\widehat{m_2^c}^2 - 2\widehat{m_2^c} - 1 + 3\widehat{m_1^c}^2 - \widehat{m_3^c}\widehat{m_1^c}}{\sqrt{\begin{pmatrix} 6\widehat{m_1^c} - \widehat{m_3^c} \\ 2\widehat{m_2^c} - 2 \\ -\widehat{m_1^c} \end{pmatrix}' \widehat{V} \begin{pmatrix} 6\widehat{m_1^c} - \widehat{m_3^c} \\ 2\widehat{m_2^c} - 2 \\ -\widehat{m_1^c} \end{pmatrix}}}$$

Under the null hypothesis

By multivariate central limit theory and Cramer Theorem (Ferguson [1996], p.45),

$$\frac{M - \mathbb{E}M}{\sqrt{var(M)}} \xrightarrow{d} N(0,1).$$

Then by slutsky's theorem,

$$T = \frac{M - \mathbb{E}M}{\sqrt{\widehat{var(M)}}} = \frac{M}{\sqrt{var(M)}}\sqrt{\frac{var(M)}{\widehat{var(M)}}} \rightarrow N(0,1).$$

Thus, we have proved:

Theorem 3.4:Under the null hypothesis,

$$\lim_{n\to\infty} \mathbb{P}(T < -z_{1-\alpha}) = \alpha,$$

where $\alpha \in (0,1)$ and $z_{1-\alpha}$ is the $1-\alpha$ quantile of a standard normal distribution.

Theorem is a more general case of Theorem 3.1.

Under the local alternative hypothesis $H_{ak} : \mu_1\mu_2\lambda_1\lambda_2 = \mu_1^*\lambda_1^*\lambda_2^*\mu_{2n}$, where μ_1^*, λ_1^*, λ_2^*, are fixed positive numbers, while $\mu_{2n} = kn^{-1/2}$.

Similar to $m = 2$ case, we have

Theorem 3.5:Under the local alternative hypothesis $H_{ak} : \mu_1\mu_2\lambda_1\lambda_2 = \mu_1^*\lambda_1^*\lambda_2^*\mu_{2n}$,

$$\lim_{n\to\infty} \mathbb{P}(T < -Z_{1-\alpha/2}) = \Phi(-Z_{1-\alpha/2} + \frac{k\mu_1^*\lambda_1^*\lambda_2^*\mu_1^{*2}}{\sqrt{m^{-2}G^T V_1 G}}).$$

A fixed number between α and 1.

The proof of Theorem 3.5 is same as Theorem 3.2.

Under a fixed alternative hypothesis $H_a : \mu_1\mu_2\lambda_1\lambda_2 = \mu_1^*\mu_2^*\lambda_1^*\lambda_2^*$, where μ_1^*,μ_2^*, λ_1^*, λ_2^* are fixed positive numbers.

Theorem 3.5:Under a fixed alternative hypothesis $H_a : \mu_1\mu_2\lambda_1\lambda_2 = \mu_1^*\mu_2^*\lambda_1^*\lambda_2^*$,

$$\lim_{n\to\infty} \mathbb{P}(|T| > Z_{1-\alpha/2}) = 1.$$

Prove of the theorem is same as Theorem 3.3.

3.6 Simulation study

To simulate data with m per group, we need first get the joint distribution of m variables. Assume $m = 3$, then define

$$\begin{aligned}
&\bar{a} = \mathbb{P}(X_1 = -1, X_2 = -1, X_3 = -1) && \bar{b} = \mathbb{P}(X_1 = -1, X_2 = -1, X_3 = 0)\\
&\bar{c} = \mathbb{P}(X_1 = -1, X_2 = -1, X_3 = 1) && \bar{d} = \mathbb{P}(X_1 = -1, X_2 = 0, X_3 = 1)\\
&\bar{e} = \mathbb{P}(X_1 = -1, X_2 = 0, X_3 = 0) && \bar{f} = \mathbb{P}(X_1 = -1, X_2 = 1, X_3 = 1)\\
&\bar{g} = \mathbb{P}(X_1 = 0, X_2 = 0, X_3 = 0) && \bar{h} = \mathbb{P}(X_1 = 0, X_2 = 0, X_3 = 1)\\
&\bar{i} = \mathbb{P}(X_1 = 0, X_2 = 1, X_3 = 1) && \bar{j} = \mathbb{P}(X_1 = 1, X_2 = 1, X_3 = 1).
\end{aligned}$$

We have the following equations:

$$\begin{aligned}
a = \bar{e} + \bar{g} + \bar{h} \quad b = \bar{h} + \bar{i} + \bar{d} \quad c = \bar{b} + \bar{d} + \bar{e}\\
d = \bar{f} + \bar{i} + \bar{j} \quad f = \bar{c} + \bar{d} + \bar{f} \quad e = \bar{a} + \bar{b} + \bar{c}.
\end{aligned} \tag{3.5}$$

Make a conjecture that

$$\begin{aligned}
&\bar{a} = e\lambda_2(1-\theta) + \lambda_2\theta && \bar{b} = c\lambda_2(1-\theta)\\
&\bar{c} = f\lambda_2(1-\theta) && \bar{d} = b\lambda_2(1-\theta)\\
&\bar{e} = a\lambda_2(1-\theta) && \bar{f} = d\lambda_2(1-\theta) - \lambda_1\lambda_2\theta(1-\theta)\\
&\bar{g} = a - (1-\theta)(\lambda_1+\lambda_2)a && \bar{h} = a\lambda_1(1-\theta)\\
&\bar{i} = b\lambda_1(1-\theta) && \bar{j} = d\lambda_1(1-\theta) + \lambda_1\theta - 2\lambda_1\lambda_2\theta(1-\theta),
\end{aligned}$$

then these satisfy the equation 3.5 , so we use them as the joint distribution of X_1, X_2, X_3.

Size Simulation

For size simulation, assume $\mu_1 = 0$, $\mu_2 = 5$, $\lambda_1 = 0.2$, $\lambda_2 = 0.3$, thus the model reduce to 2 component mixture normal: $0.7N(0,1) + 0.3N(-5,1)$. take $\tau^2 = 0.3$ and $\tau^2 = 0.6$ respectively, sample size n from 20 to 1500. For each of various sample size, generate 7000 sets of two component normal data. Next, we calculate how many times out of 7000, we reject H_0 based on theoretical critical value as estimated size of the test. As shown in figure 3.7 and figure 3.8, the estimated size all fall in the band of $0.05(\pm 0.01)$(the nominal rejection rate is 0.05).

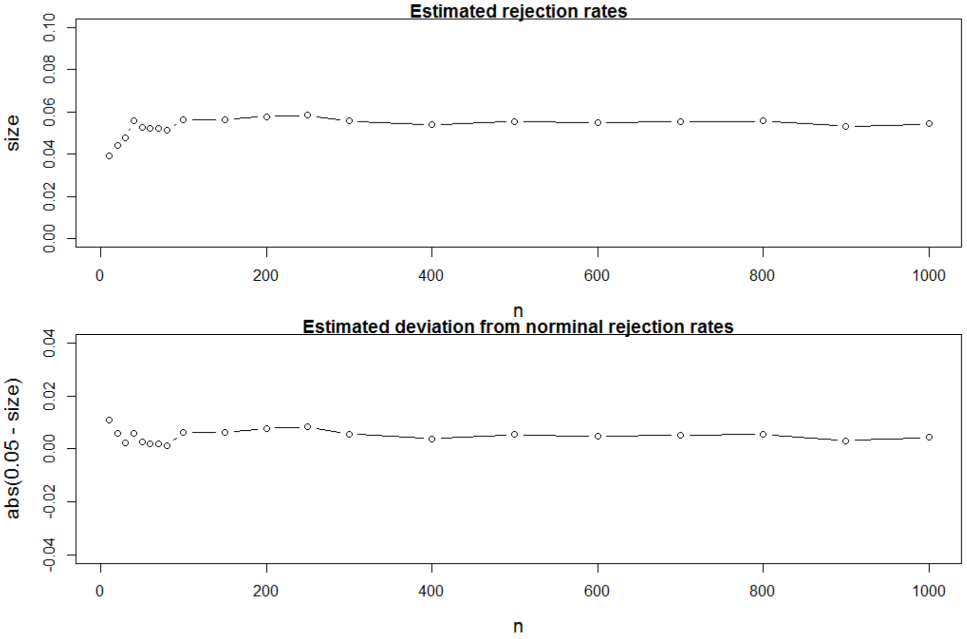

Figure 3.7: Simulation Size 7000, $\tau^2 = 0.3$,$\sigma^2 = 0.7$

Power Simulation

For power simulation, similar to $m = 2$ case, simulate data with different θ and τ^2.

Under each condition, we take sample size $n = 100$, generate $X_{1i}, X_{2i}, ..., X_{mi}$ as correlated data with correlation $\theta \in \{0.2, 0.7\}$(we here take m=3), next generate $Z_{1i}, Z_{2i}, ..., Z_{mi}$ from $N(0, \sigma^2)$ where $\sigma^2 \in \{0.8, 0.3\}$, and Z_i^* from $N(0, \tau^2)$ where $\tau^2 \in \{0.2, 0.7\}$. Thus let $Y_{ji} = (\mu_1 I_{[X_{ji}=1]} - \mu_2 I_{[X_{ji}=-1]}) + Z_{ji} + Z_i^*$ where $i = 1, 2, ..., n$, $j = 1, 2, ..., m$, $Y_{1i}, Y_{2i}, ..., Y_{mi}$ are from $\lambda_1 N(\mu_1, 1) + \lambda_2 N(-\mu_2, 1) + (1 - \lambda_1 - \lambda_2)N(0, 1)$.

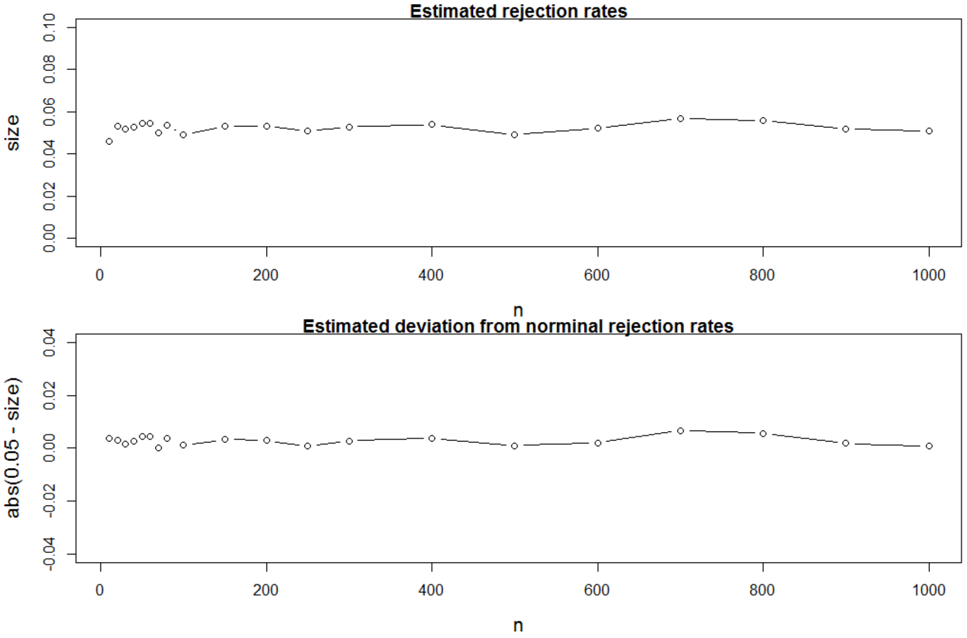

Figure 3.8: Simulation Size 7000,$\tau^2 = 0.6$,$\sigma^2 = 0.4$

We estimate the rates that we reject H_0 based on theoretical critical value as estimated power of the test. Similar to $m = 2$, for a fixed λ_1(or λ_2) away from 0, when we increase μ_1(or μ_2), or for a fixed μ_1(or μ_2) when we increase λ_1(or λ_2), the power goes to 1. Also, for a fixed μ_1, if we increase μ_2, or for a fixed μ_2, we increase μ_1, the power will also go to 1.

Also, if we compare the figure 3.9 to figure 3.12, if we increase the correlation parameters τ^2 and θ, it will need larger value of μs or λs for power to reach 1.

For $m \geq 3$, we could use NORTA method to simulate correlated multinomial data $\overline{X}$ which is a general-purpose method for generating samples of a random vector with given marginal distributions and given correlation matrix for its component random variable(Aad et al. [2015]). For example, suppose we want to simulate correlated multinomial data $\overline{X} = (X_1, X_2, X_3, X_4)$ with $m = 4$ and X_i taking value 0, 1 or -1 each with probability $1 - \lambda_1 - \lambda_2$, λ_1 and λ_2 respectively. According to Ghosh and Henderson(2002a), we could first simulate a normal vector $\overline{Z} = (Z_1, Z_2, Z_3, Z_4)$, with

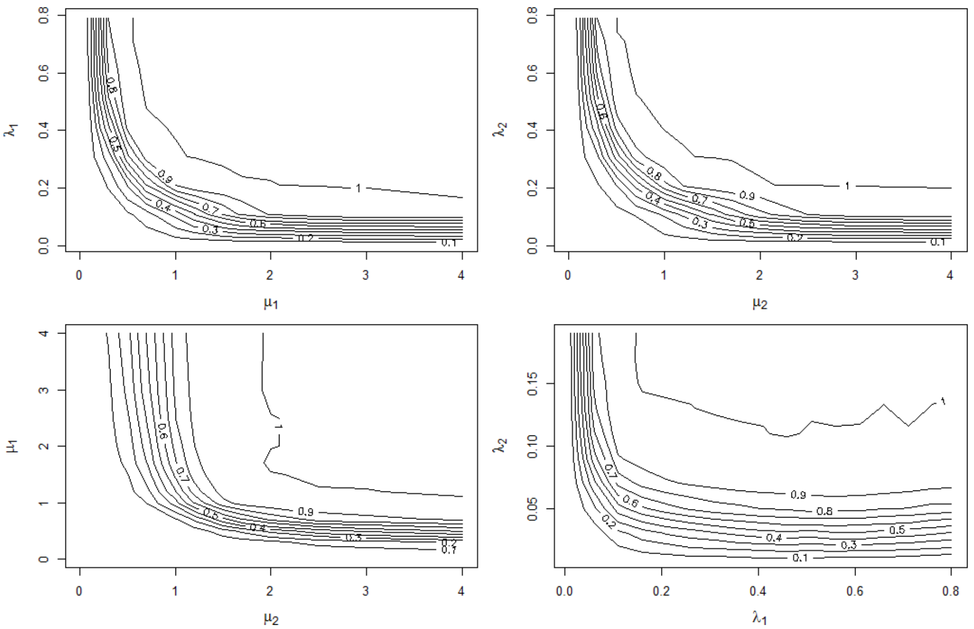

Figure 3.9: Power simulation for $\theta = 0.2$,$\tau^2 = 0.2$

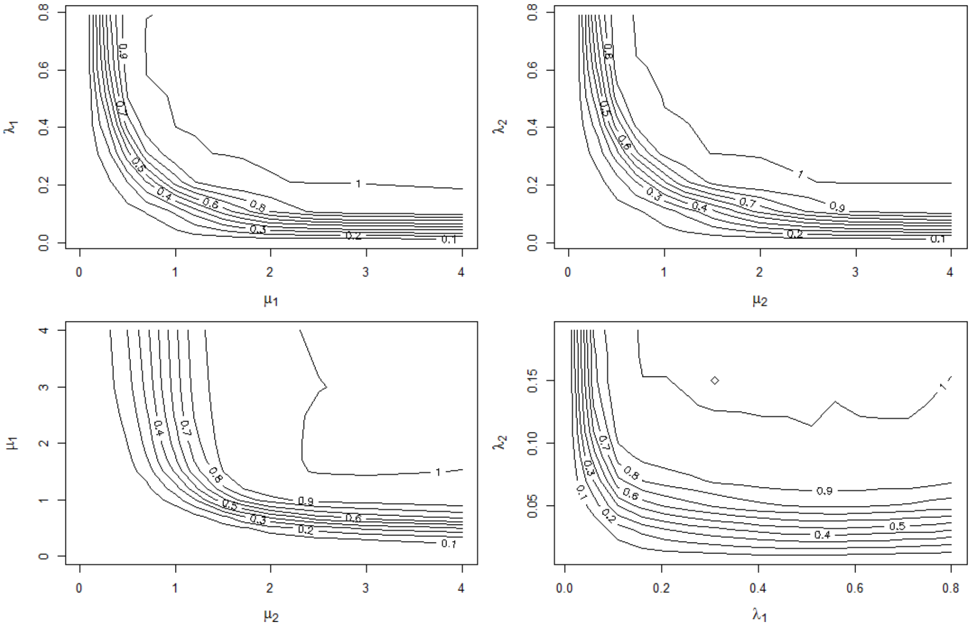

Figure 3.10: Power simulation for $\theta = 0.2$,$\tau^2 = 0.7$

covariance matrix

$$\Sigma_Z = \begin{pmatrix} 1 & a & a & a \\ a & 1 & a & a \\ a & a & 1 & a \\ a & a & a & 1 \end{pmatrix},$$

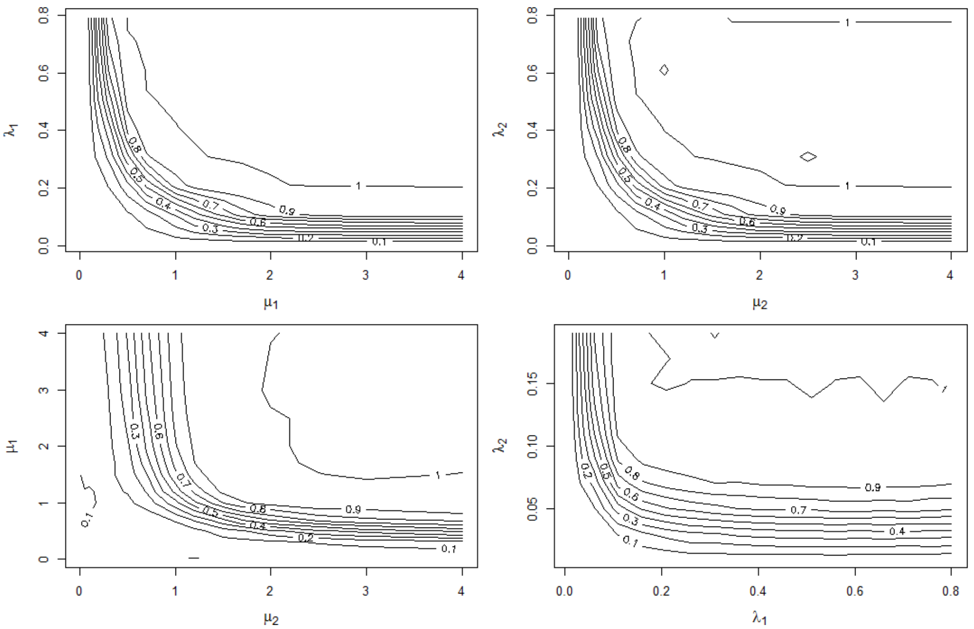

Figure 3.11: Power simulation for $\theta = 0.7$,$\tau^2 = 0.2$

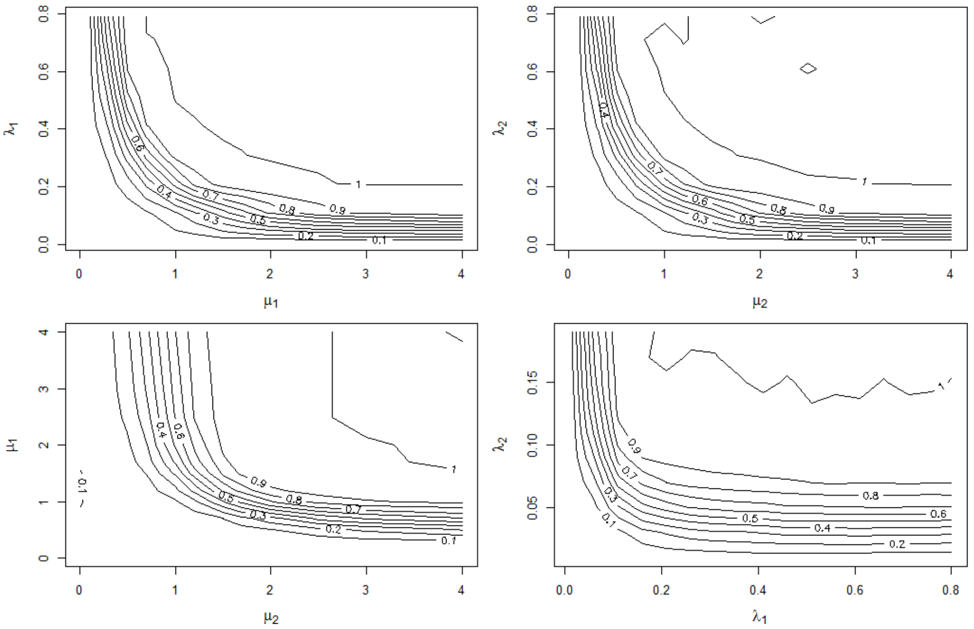

Figure 3.12: Power simulation for $\theta = 0.7$,$\tau^2 = 0.7$

then do the transformation $X_i = F^{-1}(\Phi(Z_i))$, where

$$F(x) = P(X_1 \leq x) = \begin{cases} 0 & x < -1 \\ \lambda_2 & -1 \leq x < 0 \\ 1 - \lambda_1 & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}.$$

We will get a vector X with correlation matrix

$$\Sigma_X = \begin{pmatrix} 1 & \theta & \theta & \theta \\ \theta & 1 & \theta & \theta \\ \theta & \theta & 1 & \theta \\ \theta & \theta & \theta & 1 \end{pmatrix}$$

for some θ, where θ is anticipated to be an increasing function of a. The relations between a and θ are presented in Table 1 based on numerical search.

Cor between normals(a)	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
Cor between multinormals(θ)	0.08	0.12	0.16	0.20	0.25	0.29	0.32	0.36	0.41
Cor between normals(a)	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
Cor between multinormals(θ)	0.45	0.50	0.53	0.58	0.62	0.67	0.72	0.77	0.84

Table 3.1: Relations between a and θ

Actually, we can still use the aforementioned testing procedure if the sizes of two groups are close but not the same. For example, if n_1 groups of the sample have group size 3, the other n_2 groups have group size 5, where n_1 and n_2 are close, we could then use the testing procedure as if $m = 4$ for each group. Below are simulation results to support thus assertion.

For size simulation, we take n from 100 to 1500 and half of the groups to have group size 3, while half of the groups have group size 5. Then we did the simulation using the testing procedure as if $m = 4$. We had $\tau^2 = 0.3$, $\theta = 0.2$, $\mu_1 = 0$, $\mu_2 = 5$, $\lambda_1 = 0.2$ and $\lambda_2 = 0.3$. For 7000 simulated data sets, calculated how often we rejected H_0 based on the theoretical critical value as the estimated size of the test, shown in Figure 3.13. For $n > 50$, the estimated deviation from nominal rejection rate is no more than 0.005.

For power simulation, we take $n = 100$, still half of the groups have size 3, half have size 5, and we proceed as if $m = 4$. We had $\theta = 0.2$, $\tau^2 = 0.2$. We take different values of μ_1, μ_2, λ_1 and λ_2, simulate 7000 sets of data under each parameter combination, then calculate the rate that we reject H_0 based on theoretical critical

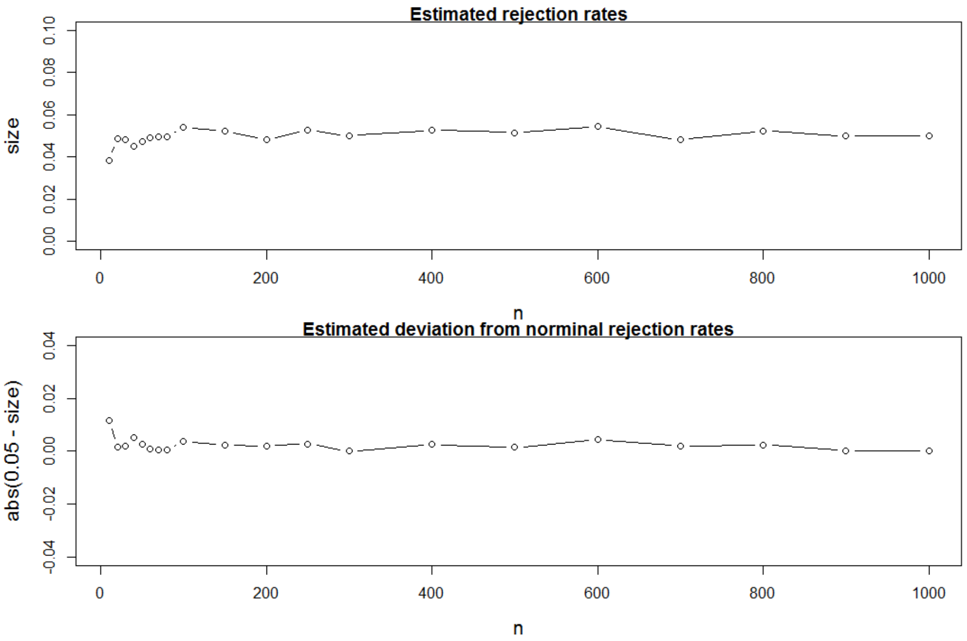

Figure 3.13: Simulation Size 7000, $\tau^2 = 0.3$,$\theta = 0.2$

value as estimated power of the test. The simulation results are shown in Fig 3.14 as contour plots.

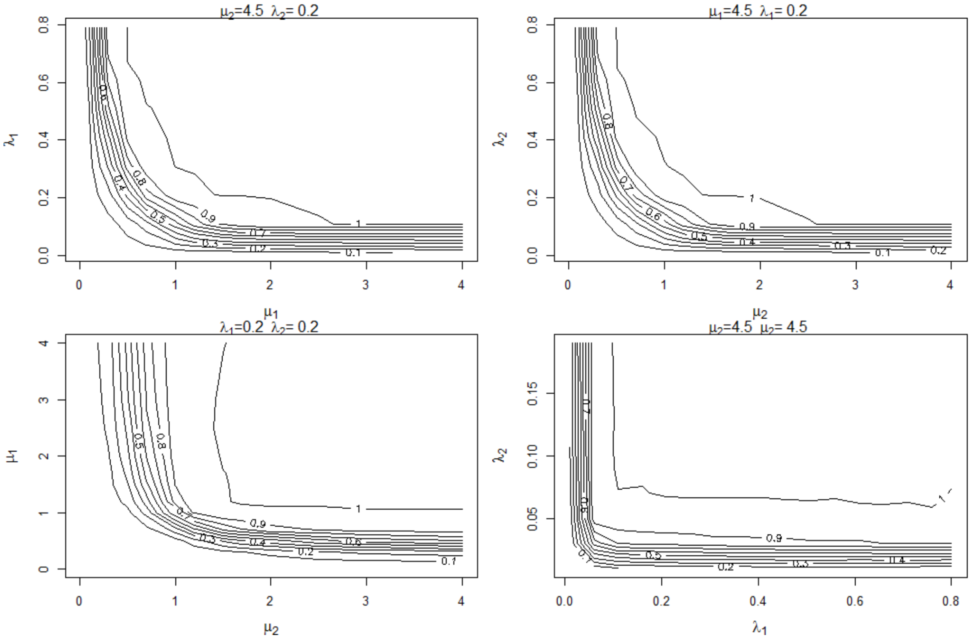

Figure 3.14: Power simulation, $\tau^2 = 0.2$,$\theta = 0.2$

Then, we change $\tau^2 = 0.7$, $\theta = 0.7$, other settings remain the same, perform the power simulation again, the result is shown in Fig 3.15.

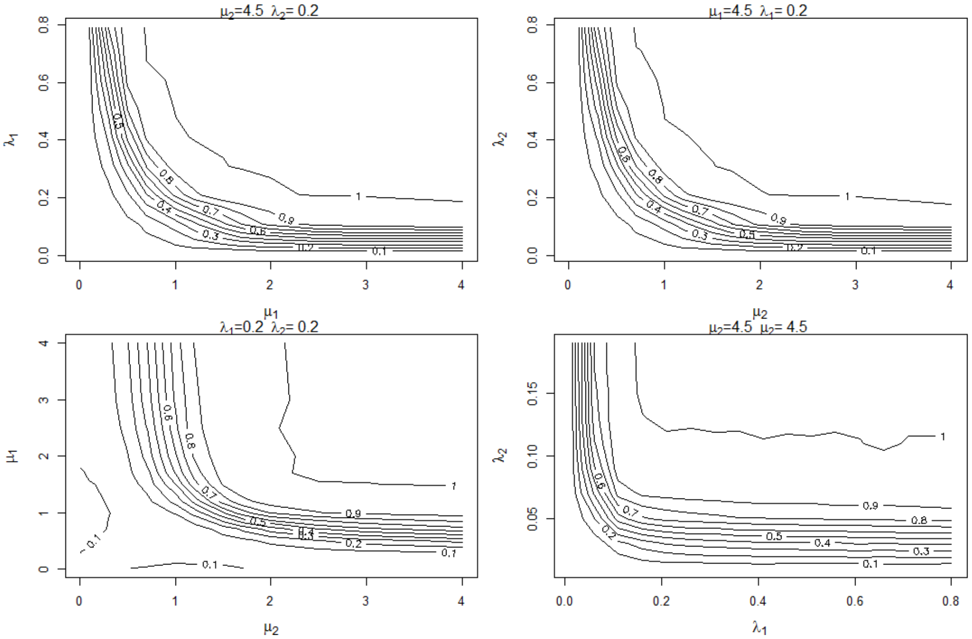

Figure 3.15: Power simulation, $\tau^2 = 0.7$,$\theta = 0.7$

3.7 Application

In application, we use the Down's syndrome data same as shown in Chapter 2(data can be download from http://www.partek.com). Follow the transform in Chapter 2, we can get the histogram of 251 Z-statistics, and the fitted 3 component normal mixture model(red line)$3.58 * 10^{-7}N(0, 1) + 0.4511324N(1.96, 1) + 0.5488673N(-1.04, 1)$, parameters are estimated from EM algorithm assume data are independent.

Next, we need to group the Z-statistics, still according to the Chromosomal locations, we divided the Z-statistics into 10 groups, with group size close to each other. More specifically, the groups are (q22.1.1, q22.1.2, q22.2, q22.3.1, q22.3.2, q22.3.3, q22.3.4, q22.3.5, q23,q21 and other locations), with group size (27, 27, 23, 25, 25, 24, 24, 25, 23, 29). Note that in order to have similar group size, we divide the location q22.1 into 2 parts: q22.1.1, q22.1.2; similarly, we divided location q22.3 into 5

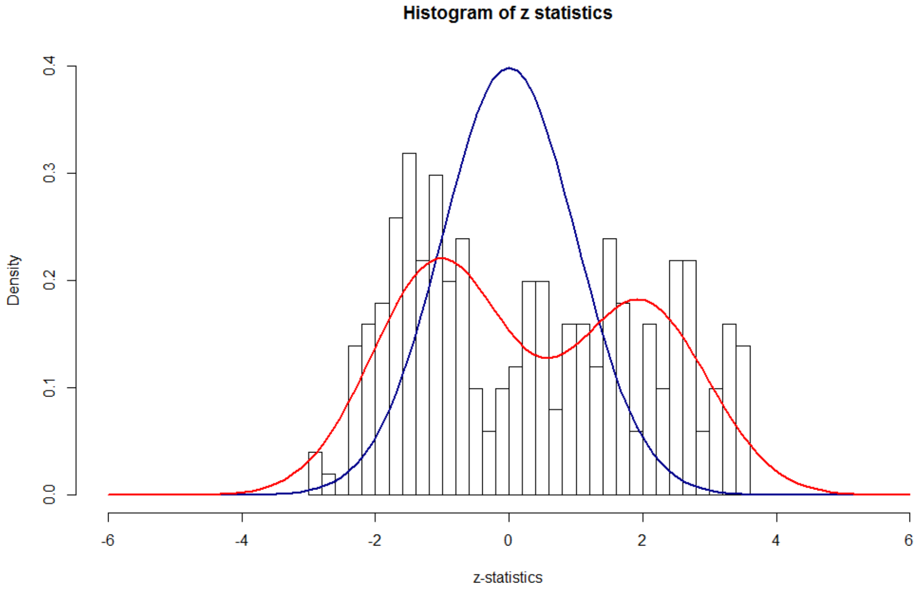

Figure 3.16: Histogram of Z-statistics, blue line is fitted standard normal curve, red line is fitted 3 component normal mixture model

sub-groups.Here we assume that the correlation structure within each group is known as compound symmetric with known θ and τ^2. Figure 3.17 shows the contour plots of P-values. As we can see in the plot, the contour are close to the straight lines, and as we increase θ and τ^2, the p-value also increases. We use a red dashed line to separate the region of accepting and rejecting $H_0 : \mu\lambda = 0$ at $\alpha = 0.05$. Similar to the previous chapter, failing to take into account correlation may massively understate the p-value.

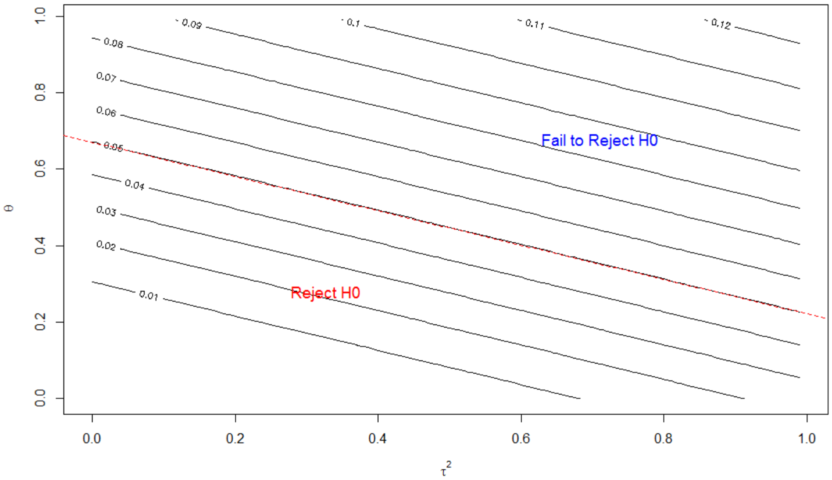

Figure 3.17: Contour plot of P-values

Chapter 4 Singular Bayesian Information Criterion For Hierarchical Normal Mixture Models

4.1 Introduction of Hierarchical Normal Mixture Models

Hierarchical models(also multilevel models or nested models) as described in Raudenbush and Bryk [2002] are particularly used when data are organized at more than one level, and are widely applied in many areas: social science, biology or public health research, etc. For example, in social and behavioral studies, we collect data of risk factors of early drop-out from two levels: the level describing an individual student(grades, gender, hours of course-work), also the level describing schools(such as types of school). Then the model could be built in these two levels, Rumberger [1995] indicates that the level-1 model is a logistic regression model of whether a student have early drop-out depending on the individual student level characteristics, while the coefficients in the level-1 model varied from the school level as a function of school characteristic. Therefore we may apply hierarchical models to simultaneously handle measurements made from different levels.

In this chapter, we focus on hierarchical normal mixture models, constructed as following:

$$\begin{aligned} &\mathbf{Y_i}|X_i = j \sim MVN(\boldsymbol{\mu_j}, \boldsymbol{\Sigma_j}), \\ &\log \tfrac{P(X_i=l|\mathbf{W_i}=\mathbf{w_i})}{P(X_i=m|\mathbf{W_i}=\mathbf{w_i})} = \alpha_l + \beta_l^T \mathbf{w_i}, \end{aligned} \tag{4.1}$$

where $j = 1, 2, ..., m$ is the index for component, $i = 1, 2, ..., n$ is the index of observation. $l = 1, 2...(m-1)$, $\alpha_m = \beta_m = 0$ and $\mathbf{W_i} \overset{iid}{\sim} MVN(\boldsymbol{\nu}, \boldsymbol{\tau})$. Then X is a random variable which determines the mixture component, and the conditional probability of X given $\mathbf{W}$ satisfies a multinomial logistic regression. $\boldsymbol{\mu_j}$ and $\boldsymbol{\Sigma_j}$ are mean and covariance matrix of $\mathbf{Y}$ in the j^{th} component.

This interpretation of model 4.1 is that: $\mathbf{Y}$ is the response variable, $\mathbf{W}$ is the predictor, X is a latent indicator between $\mathbf{W}$ and $\mathbf{Y}$, so that $\mathbf{W}$ affects $\mathbf{Y}$ through X but not directly. Furthermore, we could consider another observed variable $\mathbf{Z}$ as a predictor affecting $\mathbf{Y}$ directly. Then, this model is a hierarchical normal mixture model.The group-based trajectory model(see Nagin and Tremblay [1999],Nagin and Tremblay [2001], Charnigo et al. [2011]) is an example of hierarchical mixture model. In group-based trajectory model, the response variable $\mathbf{Y}$ is a longitudinal sequence of individual measurements over time, while $\mathbf{W}$(risk factor) is a vector of time-stable covariates which influence the probabilities of the individual belonging to a particular group defined by X, but not affecting $\mathbf{Y}$ directly. Also we can have a time-dependent covariates $\mathbf{Z}$ which affect $\mathbf{Y}$ directly(Jones etal. 2001). Neither model is a subset of the other, however they do share a special case. More specifically, model 4.1 is simplified from a general hierarchical mixture by not including $\mathbf{Z}$, while group-based trajectory model is also simplified from general hierarchical mixture models by assuming constraint on the component vector $\boldsymbol{mu_j}$. Then model 4.1 and group-based trajectory model shares a special case by intersecting the two aforementioned simplifications.

4.2 EM Algorithm for Parameter Estimation

For mixture models, we usually apply EM algorithm to estimate the unknown parameters, since EM approach has a more natural interpretation than MLE in the context of incomplete data, and could provide an approximation to MLE without requiring numerical solutions to difficult high-dimensional optimization problems(see Dempster et al. [1977], Redner and Walker [1984]). Here we first apply EM algorithm to estimate the parameters of hierarchical normal mixture model. First, assume the total number of components is known as m. Then define $P(X_i = j|\mathbf{W_i} = \mathbf{w_i})$ as p_{ij}, where $i = 1, 2, ...n$, $j = 1, 2, ..., m$ then we have

$$\begin{cases} \sum_{j=1}^{m} p_{ij} = 1 \\ log(\frac{p_{ij}}{p_{im}}) = \alpha_j + \boldsymbol{\beta}_j^T \mathbf{w_i} \end{cases} \tag{4.2}$$

Then we will have $p_{ij} = \frac{\exp(\alpha_j + \beta_{\mathbf{j}}^{\mathbf{T}} \mathbf{w_i})}{\sum_{l=1}^{m} \exp(\alpha_l + \boldsymbol{\beta}_{\boldsymbol{l}}^{\boldsymbol{T}} \mathbf{w_i})}$. Define $\Theta = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu, \tau, \alpha, \boldsymbol{\beta}\}$. Also, we can write out the likelihood function:

$$L(\Theta|\mathbf{W}, \mathbf{Y}, \mathbf{X}) = f(\mathbf{Y}|\mathbf{X}, \boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{\Sigma}_{\boldsymbol{X}}) f(\mathbf{X}|\mathbf{W}, \alpha, \boldsymbol{\beta}) f(\mathbf{W}|\boldsymbol{\nu}, \boldsymbol{\tau}).$$

We can then form the following two cases to analyze:

Case 1: Y is a scalar for an individual and a vector for full sample

Here we assume that W is also a scalar for an individual. Then the data structure can be written as

$$\begin{aligned} & Y_i | X_i = j \sim N(\mu_j, \sigma_j^2), \\ & \log \tfrac{P(X_i = l | W_i = w_i)}{P(X_i = m | W_i = w_i)} = \alpha_l + \beta_l w_i, \\ & W_i \sim N(\nu, \tau^2) \end{aligned}$$

$$\begin{aligned} L(\boldsymbol{\Theta}|\mathbf{W}, \mathbf{Y}, \mathbf{X}) &= f(\mathbf{Y}|\mathbf{X}, \mu_{\mathbf{X}}, \sigma_{\mathbf{X}}) f(\mathbf{X}|\mathbf{W}) f(\mathbf{W}|\nu, \tau) \\ &= \prod_{i=1}^{n} \sum_{j=1}^{m} \mathbb{I}_{[X_i = j]} p_{ij} \tfrac{1}{2\pi\sigma_j\tau} \exp[-\tfrac{1}{2}(\tfrac{(y_i - \mu_j)^2}{\sigma_j^2} + \tfrac{(w_i - \nu)^2}{\tau^2}).] \end{aligned} \tag{4.3}$$

Also, using Bayes' Theorem, we can get

$$\mathbb{E}(\mathbb{I}_{[X_i = j]}|\mathbf{Y}, \mathbf{W}) = \mathbb{P}(X_i = j|\mathbf{Y}, \mathbf{W}) = \frac{p_{ij} f(y_i|w_i, X_i = j)}{\sum_{l=1}^{m} p_{il} f(y_i|w_i, X_i = l)},$$

we then could define $\mathbb{E}(\mathbb{I}_{[X_i = j]}|\mathbf{Y}, \mathbf{W}) := \Phi ij$, and the approximated complete data log-likelihood function could be written as:

$$\begin{aligned} l(\boldsymbol{\Theta}) &= \sum_{i=1}^{n} \sum_{j=1}^{m} \{\Phi_{ij} [\log p_{ij} - \log \sigma_j - \log(2\pi\tau) - \tfrac{1}{2}(\tfrac{(y_i - \mu_j)^2}{\sigma_j^2} + \tfrac{(w_i - \nu)^2}{\tau^2})]\} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} \{\Phi_{ij} [\alpha_j + \beta_j w_i - \log(\sum_{l=1}^{m} \exp(\alpha_l + \beta_l w_i)) - \log \sigma_j - \tfrac{1}{2} \tfrac{(y_i - \mu_j)^2}{\sigma_j^2} \\ &\quad - \log(2\pi\tau) - \tfrac{1}{2} \tfrac{(w_i - \nu)^2}{\tau^2}]\}. \end{aligned} \tag{4.4}$$

Since for ν and τ^2, we can use MLE to get the estimated $\widehat{\nu} = \frac{\sum_{i=1}^n w_i}{n}$ and $\widehat{\tau^2} = \frac{\sum_{i=1}^n (w_i-\widehat{\nu})^2}{n}$, then the above log-likelihood approximately can be replaced by:

$$l(\mathbf{\Theta}) = \sum_{i=1}^{n}\sum_{j=1}^{m}\{\Phi_{ij}[\alpha_j + \beta_j w_i - \log(\sum_{l=1}^{m}\exp(\alpha_l + \beta_l w_i)) - \log\sigma_j - \frac{1}{2}\frac{(y_i-\mu_j)^2}{\sigma_j^2}]\} + C,$$

where C is constant with respect to Θ.

Then we can define the Q function as:

$$\begin{aligned} Q(\mathbf{\Theta}, \mathbf{\Theta^{(t)}}) &:= \mathbb{E}[l(\mathbf{\Theta})|\mathbf{Y}, \mathbf{W}, \mathbf{\Theta^{(t)}}] \\ &= \textstyle\sum_{i=1}^n\{\sum_{j=1}^{m-1}\Phi_{ij}^{(t)}[\alpha_j + \beta_j w_i - \log(1+\sum_{l=1}^{m-1}\exp(\alpha_l+\beta_l w_i)) - \log\sigma_j - \frac{1}{2}\frac{(y_i-\mu_j)^2}{\sigma_j^2}] \\ &\quad \textstyle +\Phi_{im}^{(t)}[-\log(1+\sum_{l=1}^{m-1}\exp(\alpha_l+\beta_l w_i)) - \log\sigma_m - \frac{1}{2}\frac{(y_i-\mu_m)^2}{\sigma_m^2}]\} + C, \end{aligned} \tag{4.5}$$

where $\mathbf{\Theta^{(t)}}$ is the estimation of $\mathbf{\Theta}$ after t iterations of the EM algorithm, and Q is the evaluation of log-likelihood approximation when Φ_{ij} is evaluated at $\mathbf{\Theta^{(t)}}$.

Next, we take derivative with respect to every element of $\mathbf{\Theta}$ and maximizing the Q function gives the estimators after the (t+1) iterations, the new estimators are:

$$\begin{aligned} \mu_j^{(t+1)} &= \frac{\sum_{i=1}^n \Phi_{ij}^{(t)} y_i}{\sum_{i=1}^n \Phi_{ij}^{(t)}}, \\ \sigma_j^{(t+1)} &= \sqrt{\frac{\sum_{i=1}^n \Phi_{ij}^{(t)}(y_j-\mu^{(t+1)})^2}{\sum_{i=1}^n \Phi_{ij}^{(t)}}}. \end{aligned}$$

For $\alpha_j^{(t+1)}$ and $\beta_j^{(t+1)}$, when $j \neq m$ after differentiation we have the following equations:

$$\begin{aligned} \frac{\partial Q}{\partial \alpha_j} &= \textstyle\sum_{i=1}^n \Phi_{ij}^{(t)} - \sum_{i=1}^n\sum_{l=1}^m \Phi_{il}\frac{\exp(\alpha_j+\beta_j w_i)}{1+\sum_{q=1}^{m-1}\exp(\alpha_q+\beta_q w_i)} \\ \frac{\partial Q}{\partial \beta_j} &= \textstyle\sum_{i=1}^n \Phi_{ij}^{(t)} w_i - \sum_{i=1}^n\sum_{l=1}^m \Phi_{il}\frac{w_i\exp(\alpha_j+\beta_j w_i)}{1+\sum_{q=1}^{m-1}\exp(\alpha_q+\beta_q w_i)} \end{aligned}$$

Note that $\alpha_m^{(t+1)} = \beta_m^{(t+1)} = 0$. We then could use numerical method to find the optimal value of $\alpha_j^{(t+1)}$ and $\beta_j^{(t+1)}$ for $j \neq m$. Here we use 'optim' in R to get maximize Q function with respect to α_j and β_j respectively, and then set the value to $\alpha_j^{(t+1)}$ and $\beta_j^{(t+1)}$.

Case 2: Y is a vector for an individual and a matrix for full sample

For a more general case, suppose $\mathbf{Y}$ is a matrix, then $\widehat{\boldsymbol{\nu}} = \frac{\sum_{i=1}^{n}\mathbf{w_i}}{n}$ and $\widehat{\boldsymbol{\tau}^{\mathbf{2}}} = \frac{\sum_{i=1}^{n}(\mathbf{w_i}-\widehat{\boldsymbol{\nu}})(\mathbf{w_i}-\widehat{\boldsymbol{\nu}})^T}{n}$.

Here we further assume that $\mathbf{Y}$ and $\mathbf{W}$ are both $n \times 2$ matrices, which means that the data is formed as following:

$$\begin{aligned} \mathbf{W_i} &\sim MVN(\boldsymbol{\nu} = (\nu_1, \nu_2)^T, \boldsymbol{\tau}^{\mathbf{2}}) \\ p_{ij} = \mathbb{P}(X_i = j|\mathbf{W_i} = w_i) &= \frac{exp(\alpha_j+\beta_{1j}w_{i1}+\beta_{2j}w_{i2})}{1+\sum_{l=1}^{m-1}\exp(\alpha_l+\beta_{1l}w_{i1}+\beta_{2l}w_{i2})} \\ \mathbf{Y_i}|X_i = j &\sim MVN(\boldsymbol{\mu_j} = (\mu_{1j}, \mu_{2j})^T, \boldsymbol{\Sigma_j}), \end{aligned} \tag{4.6}$$

where $\boldsymbol{\tau}^{\mathbf{2}} = \begin{pmatrix} \tau_{11}^2 & 0 \\ 0 & \tau_{22}^2 \end{pmatrix}$ and $\boldsymbol{\Sigma_j} = \begin{pmatrix} \sigma_{11j}^2 & \sigma_{12j}^2 \\ \sigma_{12j}^2 & \sigma_{22j}^2 \end{pmatrix}$ The Q function can be written as a matrix form:

$$\begin{aligned} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t)}) \; &:= \textstyle\sum_{i=1}^{n}\sum_{j=1}^{m}\{\Phi_{ij}^{(t)}[\log p_{ij} - \frac{1}{2}\log|\Sigma_j| - \frac{1}{2}(Y_i-\mu_j)^T\Sigma_j^{-1}(Y_i-\mu_j)]\} + C \\ &= \textstyle\sum_{i=1}^{n}\{\sum_{j=1}^{m-1}\Phi_{ij}^{(t)}[\alpha_j + \beta_{1j}w_{i1} + \beta 2jw_{i2} - \log(1+\sum_{l=1}^{m-1}\exp(\alpha_l + \beta_{1l}w_{i1} + \beta_{2l}w_{i2})) \\ &\textstyle -\frac{1}{2}\log|\mathbf{\Sigma_j}| - \frac{1}{2}(\mathbf{y_i} - \mu_{\mathbf{j}})^T\mathbf{\Sigma_j}^{-1}(\mathbf{y_i} - \mu_{\mathbf{j}})] \\ &\textstyle +\Phi_{im}^{(t)}[-\log(1+\sum_{l=1}^{m-1}\exp(\alpha_l + \beta_{1l}w_{i1} + \beta_{2l}w_{i2})) \\ &\textstyle -\frac{1}{2}\log|\mathbf{\Sigma_m}| - \frac{1}{2}(\mathbf{y_i} - \mu_{\mathbf{m}})^T\mathbf{\Sigma_m}^{-1}(\mathbf{y_i} - \mu_{\mathbf{m}})]\}, \end{aligned} \tag{4.7}$$

the estimators after the (t+1) iterations are:

$$\begin{aligned} \boldsymbol{\mu_j}^{(t+1)} &= \frac{\sum_{i=1}^{n}\Phi_{ij}^{(t)}\mathbf{y_i}}{\sum_{i=1}^{n}\Phi_{ij}^{(t)}}, \\ \boldsymbol{\Sigma}_j^{(t+1)} &= \frac{\sum_{i=1}^{n}\Phi_{ij}^{(t)}(\mathbf{y_i}-\boldsymbol{\mu_j}^{(t+1)})(\mathbf{y_i}-\boldsymbol{\mu_j}^{(t+1)})^T}{\sum_{i=1}^{n}\Phi_{ij}^{(t)}}. \end{aligned}$$

For $\alpha_j^{(t+1)}$, $\beta_{1j}^{(t+1)}$ and $\beta_{2j}^{(t+1)}$ when $j \neq m$, we still can use numerical method to get the optimal values, here we use 'optim' in R to get maximized Q function with respect to α_j and $\boldsymbol{\beta_j}$ and set the value to $\alpha_j^{(t+1)}$ and $\boldsymbol{\beta_j^{(t+1)}}$. Also we force $\alpha_m^{(t+1)} = \beta_{1m}^{(t+1)} = \beta_{2m}^{(t+1)} = 0$.

4.3 Singular Bayesian Information Criteria

Introduction of Singular Bayesian Information Criteria

As discussed in Chapter 1, neither AIC(Akaike Information Criteria) nor BIC(Bayesian information criterion) is appropriate when dealing with singular model selection problems due to the non-invertibility of Fisher-information matrices(Keribin [2000], Drton [2009]), such as determining the number of components for mixture model with three or more components, determining the rank in reduced-rank regression, etc.

Drton and Plummer [2013] proposed a new information criterion called sBIC(singular Bayesian information criterion), which is a Bayesian information criterion in context of a singular model selection problem. First we give an introduction of sBIC based on Drton and Plummer [2013] and Watanabe [2009b].

Suppose $Y_n = (Y_{n1}, Y_{n2}, ..., Y_{nn})$ be a sample of iid observations, $\{M_i, i \in I\}$ be a finite set of candidate models. For each model M_i, we specify a a prior distribution $P(\pi_i|M_i)$ for the probability distributions $\pi_i \in M_i$. Also parameterize M_i as $M_i = \{\pi_i(\omega_i)\, \omega_i \in \Omega_i\}$, where $\Omega_i \subseteq \mathbb{R}^{d_i}$ is a d_i-dimensional parameter space. Then we could write the marginal likelihood of Y_n as

$$L(M_i) \; = P(Y_n|M_i) = \int_{\Omega_i} P(Y_n|\pi_i(\omega_i), M_i) dP(\omega_i|M_i). \tag{4.8}$$

As mentioned in Drton and Plummer(2013), we can use Laplace approximation $\int e^{-nh(x)}dx \approx e^{-nh(\widehat{x})}(2\pi)^{(d/2)}|\Sigma|^{1/2}n^{-d/2}$ with accuracy $O(n^{-1/2})$ to approximate equation (4.8 at point $\widehat{x}$, where Σ is the inverse of the Hessian of $h(x)$ evaluated at $\widehat{x}$. For equation(4.8), we have

$$\begin{aligned} L(M_i) \; &= \textstyle\int_{\Omega_i} P(Y_n|\pi_i(\omega_i), M_i) dP(\omega_i|M_i) \\ &= \textstyle\int_{\Omega_i} \exp(-n(\frac{1}{n} \log P(Y_n|\pi_i(\omega_i), M_i))) dP(\omega_i|M_i) \\ &\approx P(Y_n|\widehat{\pi}_i, M_i)(2\pi)^{d_i/2}|\Sigma|^{1/2}n^{-d_i/2}, \end{aligned}$$

where $P(Y_n|\widehat{\pi}_i, M_i)$,is the maximum of the likelihood function. Thus, we take logarithm to get

$$\log L(M_i) = \log P(Y_n|\widehat{\pi}_i, M_i) - \frac{d_i}{2}\log(n) + O_p(1)$$

where $O_p(1)$ stands for a remainder that is bounded in probability. Then the Bayesian information criterion for model M_i (Schwarz et al. [1978]) is

$$BIC(M_i) = \log P(Y_n|\widehat{\pi}_i, M_i) - \frac{d_i}{2}\log(n).$$

However, A large-sample quadratic approximation to the log-likelihood function is not possible when the model is singular. Thus Watanabe [2009a] Theorem 6.7 shows that, for singular models, the approximation of log-likelihood has the property that, for Y_n drawn from $\pi_0 \in M_i$:

$$\log L(M_i) = \log P(Y_n|\pi_0, M_i) - \lambda_i(\pi_0)\log(n) + [m_i(\pi_0) - 1]\log\log(n) + O_p(1),$$

where $\lambda_i(\pi_0)$ and $m_i(\pi_0)$ are known as the learning coefficient and its multiplicity respectively. Also as mentioned in Watanabe [2009a], $\lambda_i(\pi_0) \in [0, d_i/2]$ is a rational number and $m_i(\pi_0) \in \{1, 2, ..., d_i\}$ is an integer.

Then Drton and Plummer [2013], shows that if likelihood ratios $P(Y_n|\widehat{\pi}_i, M_i)/P(Y_n|\pi_0, M_i)$ is bounded in probability, we could also write the log-likelihood as:

$$\log L(M_i) = \log P(Y_n|\widehat{\pi}_i, M_i) - \lambda_i(\pi_0)\log(n) + [m_i(\pi_0) - 1]\log\log(n) + O_p(1).$$

Drton and Plummer [2013] also shows that exponential families have the properties that the likelihood ratios bounded in probability. Moreover, Azaïs et al. [2009] shows that for mixture models, likelihood ratios is bounded in probability if we assume compactness on the parameter space.

Then the difficulty is to determine learning coefficients. Since if π_0 is known, the marginal likelihood could be written as:

$$L'_{\pi_0}(M_i) \propto P(Y_n|\widehat{\pi}_i, M_i) n^{-\lambda_i(\pi_0)}(\log n)^{m_i(\pi_0)-1}.$$

Then Drton and Plummer [2013] propose that for π_0 unknown, we could give a probability distribution Q_i to the distributions in model M_i, and approximate the marginal likelihood as:

$$L^{'}_{Q_i}(M_i) \ = \int_{M_i} L^{'}_{\pi_0}(M_i) dQ_i(\pi_0).$$

Then he mentioned that, by the usage of posterior distribution, we could choose Q_i, by conditioning on all sub-models of M_i, as

$$Q_i(\pi_0) \ = P(\pi_0|\{M : M \subseteq M_i\}, Y_n) = \frac{\sum_{j \preceq i} P(\pi_0|M_j, Y_n) P(M_j|Y_n)}{\sum_{j \preceq i} P(M_j|Y_n)}, \tag{4.9}$$

where we define $j \preceq i$ if $M_j \subseteq M_i$.

For example, in normal mixture model, define posterior distribution of 2 component normal mixture as p_2, posterior of normal as p_1, then

$$Q(2comp) = \frac{p_2 * P(2comp|data) + p_1 * P(normal|data)}{P(2comp|data) + P(normal|data)}$$

Since under certain conditions, $\lambda_i(\pi_0)$ and $m_i(\pi_0)$ are almost surely constants, then denote:

$$L^{'}_{ij} \ = P(Y_n|\widehat{\pi}_i, M_i) n^{-\lambda_{ij}} (\log n)^{m_{ij}-1} > 0.$$

Since if we define $L^{'}(M_i) \ = L^{'}_{Q_i}(M_i)$, we could get:

$$L^{'}(M_i) = \frac{\sum_{j \preceq i} L^{'}_{ij} L^{'}(M_j) P(M_j)}{\sum_{j \preceq i} L^{'}(M_j) P(M_j)}, \quad i \in I,$$

and further we can get

$$\sum_{j \preceq i} [L^{'}(M_i) - L^{'}_{ij}] L^{'}(M_j) P(M_j) = 0, \quad i \in I. \tag{4.10}$$

Then, follow the definition 3.1 in Drton(2013),

$$sBIC(M_i) = \log(L^{'}(M_i)),$$

where $\{L^{'}(M_i), i \in I\}$ is the unique solution to equation (4.10) that has all entries positive.

Also, singular BIC can be written as

$$sBIC(M_i) = \log P(Y_n|\widehat{\pi}_i, M_i) - penalty(M_i),$$

where $penalty(M_i) \leq dim(M_i)/2 * \log n$, is milder than ordinary BIC penalty. Hence sBIC will select a greater or equal number of component to BIC.

For our hierarchical normal mixture model (4.1) described in section 4.1 , assume that the parameter space is compact, if the variance is equal and known, the learning coefficients have been determined by Aoyagi [2010]. If variance is unequal and unknown, according to Drton and Plummer [2013], we could apply methods in Watanabe [2009b] Section 7.3.

For case 1: Y is a scalar for an individual and a vector for full sample, suppose l is the total number of normal mixture components in the learning machine, m is the number of components in a true model when $\pi_0 \in M_m \subset M_l$, and $m < l$. Then for the hierarchical normal mixture model, in layer 1, both τ^2 and ν have counts 1; for layer 2, both α and β have counts $(l-1)$; for layer 3 both μ and σ have counts l. Thus the parameter space have dimension $4l$ However, when $\pi_0 \in M_m \subset M_l$, we will leave $l-m$ of α and $l-m$ of β free. Then it leads to the bound of learning coefficient:

$$\lambda_{lm} \leq \frac{1}{2}(4l-2(l-m)) = \frac{1}{2}(2l+2m) = l+m < \frac{dim(M_l)}{2} = \frac{1}{2}*4l = 2l \quad when \quad m < l.$$

For case 2: Since now **Y** and **W** are vectors. Then suppose l is the number of normal mixture components in the learning machine, m is the number of components in a true distribution when $\pi_0 \in M_m \subset M_l$, and $m < l$. The dimension of total parameter space is $8l + 2$(5 from ν and T; $l-1$ from α, β_1 and β_2 respectively; $5l$ from μ and Σ). However, when $\pi_0 \in M_m \subset M_l$, we can leave $l-m$ of α, $l-m$ of β_1 and $l-m$ of β_2 free. Then it leads to the bound of learning coefficient:

$$\lambda_{lm} \leq \frac{1}{2}(3l + 5m + 2) < \frac{dim(M_l)}{2} = 4l + 1 \quad when \quad m < l.$$

Consistency of sBIC for hierarchical normal mixture model

Drton and Plummer [2013] exhibits three assumptions about the likelihood ratios and the learning coefficients and their multiplicities to prove consistency:

- (A1) For any i, j, if M_i and M_j are true models, then the ratio of their likelihoods are bounded in probability as the sample size goes to infinity.

- (A2) For any i, k, if M_i is a true model and M_k is a false model(π_0 not in M_k), then there exists $\delta_{ik} > 0$ such that

  $$P(\frac{L(Y_n, W_n|\widehat{\pi}_k, M_k)}{L(Y_n, W_n|\widehat{\pi}_i, M_i)} \leq e^{-\delta_{ik}n}) \to 1, n \to \infty,$$

  where $\mathbf{Y_n}$ and $\mathbf{W_n}$ stands for n-dimensional vector as mentioned before.

- (A3) For any true models M_i, M_k and their corresponding sub-models $M_j \subseteq M_i$, $M_l \subseteq M_k$, the Bayes complexity is monotonically increasing, i.e $(\lambda_{ij}, m_{ij}) < (\lambda_{kl}, m_{kl})$, if $i \prec k$ and $j \preceq l$.

To prove the consistency of sBIC for hierarchical normal mixture model, we need to prove (A1)-(A3) for our model. First we assume that the parameter space is compact.

Proof:

Proof. For (A1), suppose $\mathcal{G}$ is a set of densities g with respect to Lebesgue measure v, M_a, M_b are two true models with densities $g_a \in \mathcal{G}$,$g_b \in \mathcal{G}$. Let the smallest true model be M_c, $M_c \subseteq M_a$ and $M_c \subseteq M_b$. Let g_c be the unique true density of order c. The density of M_c is $g_c \in \mathcal{G}$. The log-likelihood function of model M_d can be written as $l_n(M_d) = \log L(Y_n, W_n|M_d, \widehat{\pi}_d)$, where W_n is a random vector. Then a log-likelihood ratio test statistic for $H_0 : g_c = f \in \mathcal{G}$ can be written as,

$$LRT_d = \sup_{g \in \mathcal{G}_d}(l_n(g) - l_n(f)),$$

Thus, $\log P(Y_n, W_n|M_c, \pi_c) - \log P(Y_n, W_n|M_a, \widehat{\pi}_a)$ and $\log P(Y_n, W_n|M_c, \pi_c) - \log P(Y_n, W_n|M_b, \widehat{\pi}_b)$ are both negative log-likelihood ratio test statistics. Following by Gassiat [2002], for any $g \in \mathcal{G}_d$, if we define the score function

$$s_g = \frac{\frac{g-f}{f}}{\| \frac{g-f}{f} \|_2} = \frac{\frac{g}{f} - 1}{\| \frac{g}{f} - 1 \|_2},$$

where $\| \cdot \|_2$ is the norm in $L^2(fdv)$, then

$$0 \leq \sup_{g \in \mathcal{G}_d} (l_n(g) - l_n(f)) \leq \frac{1}{2} \sup_{g \in \mathcal{G}_d} \frac{(\sum_{i=1}^n s_g(Y_i, W_i))^2}{\sum_{i=1}^n (s_g)_-^2 (Y_i, W_i)}. \tag{4.11}$$

Note that, the left hand side of the inequality is 1 log-likelihood ratio test statistic. For right hand side, we claim that it is $O_p(1)$. With this claim, we could conclude that

$$\log P(Y_n, W_n | M_b, \widehat{\pi}_b) - \log P(Y_n, W_n | M_c, \pi_c) = O_p(1)$$

and

$$\log P(Y_n, W_n | M_c, \pi_c) - \log P(Y_n, W_n | M_a, \widehat{\pi}_a) = O_p(1).$$

Thus,

$$\begin{aligned} \log P(Y_n, W_n | M_a, \widehat{\pi}_a) - \log P(Y_n, W_n | M_b, \widehat{\pi}_b) &= \log P(Y_n, W_n | M_a, \widehat{\pi}_a) - \log P(Y_n, W_n | M_c, \pi_c) \\ &\quad -(\log P(Y_n, W_n | M_b, \widehat{\pi}_b) - \log P(Y_n, W_n | M_c, \pi_c)) \\ &= O_p(1) \end{aligned} .$$

Exponentiate both sides, we can further conclude that the ratios of their likelihoods are bounded in probability, which shows that assumption (A1) holds.

Next, we argue that our claim is right. Define $\mathcal{S}_d$ to be the set of all score functions corresponding to $\mathcal{G}_d$. Following Gassiat [2002], we assume finite integration of square root entropy, i.e. let $H_\beta(u)$ be the entropy with bracket of $\mathcal{S}$ with respect to $\| \cdot \|_2$, and assume that

$$\int_0^1 \sqrt{H_\beta(u)} du < +\infty.$$

Actually the finite integration of square root entropy for $g_a, g_c \in \mathcal{G}$ are,

$$g(Y, W | M_a, \widehat{\pi}_a) = \sum_{j=1}^{J_a} \frac{\exp(\widehat{\beta}_j w)}{1 + \sum_{l=1}^{J_a - 1} \exp(\widehat{\beta}_l w)} f_{N(\widehat{\mu_j}, \widehat{\sigma_j}^2)}(y),$$

$$g(Y, W | M_c, \pi_c) = \sum_{j=1}^{J_c} \frac{\exp(\beta_j w)}{1 + \sum_{l=1}^{J_c - 1} \exp(\beta_l w)} f_{N(\mu_j, \sigma_j^2)}(y),$$

where $f_{N(\mu_j, \sigma_j^2)}(y)$ is the pdf of $N(\mu_j, \sigma_j^2)$.

According to Van der Vaart [2000], the class of all monotone functions taking values in $[-1,1]$ have finite integration of square root entropy(belongs to P-Donsker class). Thus, $\mathcal{F}_\infty = \{f_\theta = \frac{\exp(\beta_j w)}{1+\exp(\beta_j w)}\}$ is a Donsker class. Then class of normal pdf functions $\mathcal{F}_\in = \{f_\theta, \theta \in \Theta\}$ is also a Donsker class, where Θ correspond to total parameter space of normal probability density. It is a parametric class and satisfy the condition that there exist a measurable function m such that

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq m(x) \parallel \theta_1 - \theta_2 \parallel .$$

Actually, since we suppose our parameter space is compact, then there exist $M > 0$ such that $-M < |\mu_j| < M$ and $\frac{1}{M} < |\sigma_j| < M$. Then by Taylor Expansion we have

$$f_{(\mu_1,\sigma_1)}(x) - f_{(\mu_2,\sigma_2)}(x) = (\mu_1 - \mu_2)\frac{\partial f}{\partial \mu}|_{(0,1)} + (\sigma_1 - \sigma_2)\frac{\partial f}{\partial \sigma}|_{(0,1)} + R_1,$$

where R_1 is a remainder and satisfy $R_1 \leq M_2 * (\parallel \mu_1 - \mu_2 \parallel + \parallel \sigma_1 - \sigma_2 \parallel)$, M_2 is a upper bound of $|\frac{\partial^2 f}{\partial \mu^2}|$, $|\frac{\partial^2 f}{\partial \sigma^2}|$ and $|\frac{\partial^2 f}{\partial \mu \partial \sigma}|$ along a line segment connecting $(0,1)$ and (μ, σ), where $\mu = \sup(|\mu_1|, |\mu_2|)$ and $\sigma = \sup(\sigma_1, \sigma_2)$. Thus,

$$|f_{(\mu_1,\sigma_1)}(x) - f_{(\mu_2,\sigma_2)}(x)| = (\mu_1 - \mu_2)\frac{\partial f}{\partial \mu}|_{(0,1)} + (\sigma_1 - \sigma_2)\frac{\partial f}{\partial \sigma}|_{(0,1)} + R_1 < |\theta_1 - \theta_2| m(x),$$

Then according to Van der Vaart [2000], the class of normal density functions with compact parameter space is a Donsker class. Also since Linear combination of a finite number of functions with bounded coefficients have finite integration of square root entropy, then $\mathcal{G} = \{g(Y, W|M_a, \widehat{\pi}_a) = \sum_{j=1}^{J_a} \frac{\exp(\widehat{\beta_j} w)}{1+\exp(\widehat{\beta_j} w)} f_{N(\widehat{\mu_j}, \widehat{\sigma_j}^2)}(y)\}$ is a Donsker class. Next, we argue that the score functions have finite integration of square root entropy.

For $\mathcal{S}_d = \{s_g = \frac{\frac{g}{f}-1}{\|\frac{g}{f}-1\|_2}\}$, f is a fixed function(here we define it as the density for smallest true model). Then according to the definition in (Van der Vaart [2000]), we have two bracketing functions l and u with finite $L(P)$ -norms. Since $l \leq g \leq u$, $g \in \mathcal{G}$, then for a fixed density function f, $\frac{l}{f} - 1 \leq \frac{g}{f} - 1 \leq \frac{u}{f} - 1$. Then for $\frac{l}{f} - 1$ and $\frac{u}{f} - 1$, we have

$$\int |\frac{u}{f} - 1| dP = \int |\frac{u}{f} - 1| f dv = \int |u - f| dv \leq \int |u| dv + \int |f| dv = \int |u| dv + 1,$$

$$\int |\frac{l}{f} - 1|dP = \int |\frac{l}{f} - 1|fdv = \int |l - f|dv \geq \int |l|dv - \int |f|dv = \int |l|dv - 1.$$

Then we have

$$\int |l|dv - 1 \leq \int |\frac{g}{f} - 1|dv \leq \int |u|dm + 1.$$

Thus, the class of score functions still have finite integration of square root bracketing numbers since l and u have finite $L(P)$ -norms.

So we have proved that the class of score functions has finite integration of square root entropy. Then according to Theorem 1 of Doukhan et al. [1995]

$$\sup_{s\in\mathcal{S}} \frac{1}{n}(\sum_{i=1}^{n} s(Y_i, W_i))^2 = O_p(1).$$

Also, it is shown in Gassiat [2002] that

$$\lim_{n\to\infty} \inf_{s\in\mathcal{S}} \frac{1}{n} \sum_{i=1}^{n} s_-^2(X_i) = \inf_{s\in\mathcal{S}} \parallel s_- \parallel_2^2,$$

and

$$\inf_{s\in\mathcal{S}} \parallel s_- \parallel_2^2 > 0.$$

Thus following Gassiat [2002], take X_i to be the vector of (Y_i, W_i), the right hand side of inequality (4.11) satisfies

$$\sup_{g\in\mathcal{G}} \frac{(\sum_{i=1}^{n} s_g(Y_i, W_i))^2}{\sum_{i=1}^{n} (s_g)_-^2(Y_i, W_i)} = O_p(1)$$

For(A2), suppose M_t is true model, M_f is false model and M_c is the smallest true model. Then we have:

$$\begin{aligned} \log(\tfrac{P(Y_n,W_n|\widehat{\pi_f},M_f)}{P(Y_n,W_n|\pi_c)}) &= \log f_{M_f}(Y_n, W_n|\widehat{\pi_f}) - \log f_{M_c}(Y_n, W_n|\pi_c) \\ & \qquad . \\ &= \textstyle\sum_{i=1}^{n} \log f_{M_f}(Y_i, W_i|\widehat{\pi_f}) - \sum_{i=1}^{n} \log f_{M_c}(Y_i, W_i|\pi_c) \end{aligned}$$

Assume $f(x, \theta)$ is a pdf, the assumptions for uniform law of large numbers are: firstly the parameter space Θ is compact; secondly, $f(x, \theta)$ is continuous at each $\theta \in \Theta$; then, there exists a dominating function $d(x)$, which is independent of parameters, such that $\|f(x,\theta)\| \leq d(x)$ for every x, $x = (y, w)$. According to our case, the first two assumptions are met. For the last one:

$$f_{M_f}(Y, W|\theta) \leq \sum_{j=1}^{m_f} \frac{1}{\sqrt{2\pi}\sigma_j\tau} \exp\left(-\frac{1}{2}\left(\frac{(Y-\mu_j)^2}{\sigma_j^2} + \frac{(w-\nu)^2}{\tau^2}\right)\right),$$

since the parameter space is compact, then there exist U_1, U_2,U_3 and $U_4 > 0$, which satisfy that $\max_{1\leq j\leq m_f} |\mu_j| \leq U_1$, $|\nu| \leq U_2$, $\max_{1\leq j\leq m_f} |\sigma_j| \geq U_3$, $|\tau_j| \geq U_4$ and $\min_{1\leq j\leq m_f} |\sigma_j| \geq U_3$. Then the domination function can be defined as following:

$$d(Y,W) = \begin{cases} U_5U_4/2\pi \exp(-\frac{1}{2}[U_3^2(Y+U_1)^2 + U_4^2(W+U_2)^2]), & < Y < -U_1,\ W < -U_2; \\ U_5U_4/2\pi \exp(-\frac{1}{2}U_3^2(Y+U_1)^2), & Y < -U_1, -U_2 \leq W \leq U_2; \\ U_5U_4/2\pi \exp(-\frac{1}{2}[U_3^2(Y+U_1)^2 + U_4^2(W-U_2)^2]), & Y < -U_1,\ W > U_2; \\ U_5U_4/2\pi \exp(-\frac{1}{2}U_4^2(W+U_3)^2), & W < -U_2, -U_1 \leq Y \leq U_1; \\ U_5U_4/2\pi, & -U_2 \leq W \leq U_2, -U_1 \leq Y \leq U_1; \\ U_5U_4/2\pi \exp(-\frac{1}{2}[U_3^2(Y-U_1)^2 + U_4^2(W+U_2)^2]), & Y > U_1,\ W < -U_2; \\ U_5U_4/2\pi \exp(-\frac{1}{2}U_3^2(Y-U_1)^2), & Y > U_1, -U_2 \leq W \leq U_2; \\ U_5U_4/2\pi \exp(-\frac{1}{2}[U_3^2(Y-U_1)^2 + U_4^2(W-U_2)^2]), & Y > U_1,\ W > U_2. \\ U_5U_4/2\pi \exp(-\frac{1}{2}U_4^2(W-U_2)^2) & W > U_2, -U_1 \leq Y \leq U_1; \end{cases}$$

Then by uniform law of large numbers:

$$\begin{aligned} \tfrac{1}{n}\textstyle\sum_{i=1}^n \log f_{M_f}(Y_i, w_i|\widetilde{\pi_f}) &\overset{a.s}{\rightarrow} \mathbb{E}[\log f_{M_f}(Y_1, w_1|\widetilde{\pi_f})] \\ \tfrac{1}{n}\textstyle\sum_{i=1}^n \log f_{M_c}(Y_i, w_i|\pi_c) &\overset{a.s}{\rightarrow} \mathbb{E}[\log f_{M_c}(Y_1, w_1|\pi_c)], \end{aligned}$$

suppose, for the false model M_f, $\widetilde{\pi_f} := \arg\max_{\pi_f} \mathbb{E}[\log f_{M_f}(Y, W|\widetilde{\pi_f})]$, then, $\widehat{\pi_f} \overset{a.s}{\rightarrow} \widetilde{\pi_f}$ and $\pi_c \overset{a.s}{\rightarrow} \pi_c$ as $n \rightarrow \infty$. Thus, by Slutskys theorem we have

$$\begin{aligned} &\tfrac{1}{n}\textstyle\sum_{i=1}^n \log f_{M_f}(Y_i, w_i|\widehat{\pi_f}) - \tfrac{1}{n}\textstyle\sum_{i=1}^n \log f_{M_c}(Y_i, w_i|\pi_c) \\ &\overset{a.s}{\rightarrow} \mathbb{E}[\log f_{M_f}(Y_1, W_1|\widetilde{\pi_f})] - \mathbb{E}[\log f_{M_c}(Y_1, W_1|\pi_c)] \\ &= \textstyle\int\int \log \tfrac{f_{M_f}(Y_1,W_1|\widetilde{\pi_f})}{f_{M_c}(Y_1,w_1|\pi_c)} f_{M_c}(Y_1, W_1|\pi_c) dY_1 dw_1 \\ &\ \mathit{by\ Jensen's\ inequality} \\ &< \log \textstyle\int\int \tfrac{f_{M_f}(Y_1,W_1|\widetilde{\pi_f})}{f_{M_c}(Y_1,w_1|\pi_c)} f_{M_c}(Y_1, W_1|\pi_c) dY_1 dw_1 \\ &= \log(1) = 0. \end{aligned}$$

Thus,

$$\frac{1}{n}\sum_{i=1}^{n}\log\frac{f_{M_f}(Y_i, w_i|\widehat{\pi_f})}{f_{M_c}(Y_i, w_i|\pi_c)} \overset{a.s}{\to} -\kappa < 0.$$

Then for any $\varepsilon > 0$ there exist N_ε, such that $n > N_\varepsilon$

$$\mathbb{P}(\sum_{i=1}^{n}\log\frac{f_{M_f}(Y_i, w_i|\widehat{\pi_f})}{f_{M_c}(Y_i, w_i|\pi_c)} \geq> -n\delta) \leq \varepsilon$$

Then

$$\mathbb{P}(\frac{P(Y_n, w_n|\widehat{\pi_f}, M_f)}{P(Y_n, w_n|\pi_c, M_c)} \geq e^{-n\delta}) \leq \varepsilon.$$

Also since M_c is the smallest true model, we have $\frac{P(Y_n|\pi_c,M_c)}{P(Y_n,w_n|\widehat{\pi}_t,M_t)} \leq 1$, then we have

$$\mathbb{P}(\frac{P(Y_n, w_n|\widehat{\pi_f}, M_f)}{P(Y_n, w_n|\widehat{\pi}_t, M_t)}) = \frac{P(Y_n, w_n|\widehat{\pi_f}, M_f)}{P(Y_n, w_n|\pi_c, M_c)}) * \frac{P(Y_n, w_n|\pi_c, M_c)}{P(Y_n\delta, w_n|\widehat{\pi}_t, M_t)}) \geq e^{-n\delta}) \leq \varepsilon.$$

Then we have proved that for any i, k, if M_i is true model, M_k is false model(π_0 not in M_k), then there exist $\delta > 0$, such that

$$P(\frac{P(Y_n, w_n|\widehat{\pi}_k, M_k)}{P(Y_n, w_n|\widehat{\pi}_i, M_i)} \leq e^{-\delta n}) \to 1, n \to \infty$$

For (A3), since for our model, we have for Y is a scalar for each individual:$\lambda_{ij} = 0.5(2i + 2j)$ and $m_{ij} = 1$, then if there exist $M_l \subseteq M_k$ and $i \prec k$ and $j \preceq l$, then $0.5(2i + 2j) < \lambda_{ij} < \lambda_{kl} < 0.5(2k + 2l)$ and $m_{ij} = m_{kl} = 1$. Thus we have prove that $(\lambda_{ij}, m_{ij}) \preceq (\lambda_{kl}, m_{kl})$, which means that the bayes complexity are monotonically increasing.

For Y is a vector for each individual $\lambda_{ij} = \frac{1}{2}(3i + 5j + 2)$ and $m_{ij} = 1$. Then assume there exist $M_l \subseteq M_k$ and $i \prec k$ and $j \preceq k$, then $\frac{1}{2}(3i + 5j + 2) = \lambda_{ij} < \lambda_{kl} = \frac{1}{2}(3k + 5l + 2)$, $m_{ij} = m_{kl} = 1$. Thus we have prove that $(\lambda_{ij}, m_{ij}) \preceq (\lambda_{kl}, m_{kl})$, which means that the bayes complexity are monotonically increasing.

Similarly, we could prove the assumption (A1) and (A2) are hold when Y is a vector. □

Since we have proved that our hierarchical normal mixture model satisfies these three assumption, then by Drton and Plummer [2013] theorem 4.1 the sBIC is consistent for hierarchical normal mixture model. Moreover, by Lemma 4.1 in Drton(2013), if M_i is a smallest true model then since assumption (A2) satisfied, we have

$$sBIC(M_i) = \log(L_{ii}^{'}) + o_p(1).$$

Thus, in the simulation and application section of Chapter 5, we will apply this Lemma and approximate $sBIC(M_i) \approx \log(L_{ii}^{'})$ in order to avoid calculating a large number of likelihood functions when n is large(for example n is larger than 500).

Chapter 5 A Flexible Singular Information Criterion For Hierarchical Normal Mixture Models

5.1 Introduction

In Chapter 1 we mentioned that AIC is an inconsistent estimator when n is large, while the BIC will underestimate the number of components when n is small, so Pilla and Charnigo [2007] proposed a new model selection criterion named FLIC(Flexible information criterion).Pilla and Charnigo [2007] shows that FLIC works better than BIC when sample size is small, also working better than AIC for large samples. In addition, the penalty of FLIC are data generated, i.e "it takes into account the structure of the data to determine the strength of the penalty term". In this chapter, we will develop a new flexible information criterion following by Pilla and Charnigo [2007] for a singular hierarchical normal mixture model 4.1; we will use the birth weight data as an example to illustrate our methodology for and compare the new information criterion with AIC, BIC and sBIC. Notice that our work is different from Pilla and Charnigo [2007] from following three aspects: 1) we consider a hierarchical mixture model with varying coefficients; 2) we consider the data with vector response and vector covariates.

5.2 Singular Flexible Information Criterion for Hierarchical Normal Mixture Model

As mentioned in Pilla and Charnigo [2007], the penalty term for FLIC is $2(\log\sqrt{n})^{B(n,\delta)}(3m-1)$, where 3m-1 is the number of free parameters in a m-component mixture model from Pilla and Charnigo [2007], n is the sample size, δ denotes the fraction of within-component variability to the total variability, and $B(n,\gamma) = \frac{\Phi[(\log\sqrt{n})^{\gamma}]-\Phi(1)}{1-\Phi(1)}$ is a bivariate function taking a value between 0 and 1. For multivariate case, according to Dunteman [1984] and Fan, we could define the data y_{ilj} as the observation from ith outcome of lth measurement in jth component of the mixture, where $i = 1, 2, ..., p$, $j = 1, 2, ...J$, $l = 1, 2, ..., n_j$(where $\sum_{j=1}^{J} n_j = N$). Then according to the MANOVA(multivariate analysis of variance) we would have that

$$SStotal_J = \sum_{j=1}^{J} SSwithin_j + SSbetween,$$

where $SStotal_J$, $SSwithin$, $SSbetween$ stand for total variability, within-component variability and between-component variability for J-component mixture model, respectively.(J $\in$ index set I) $SSwithin_j$ is the within variability for jth component. For example, in our model we could assume $Y = (Y_1, Y_2)^T$, $W = (W_1, W_2)^T$, where

$$W \sim MVN(\nu = \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix}, \tau = \begin{pmatrix} \tau_{11}^2 & 0 \\ 0 & \tau_{22}^2 \end{pmatrix}),$$

$$Y|X = j \sim MVN(\mu_j = \begin{pmatrix} \mu_{1j} \\ \mu_{2j} \end{pmatrix}, \Sigma_j = \begin{pmatrix} \sigma_{11j}^2 & \sigma_{12j}^2 \\ \sigma_{12j}^2 & \sigma_{22j}^2 \end{pmatrix}).$$

Then we will have the MANOVA as following:

$$SSwithin_j = \begin{pmatrix} SW_{j11} & SW_{j12} & SW_{j13} & SW_{j14} \\ SW_{j21} & SW_{j22} & SW_{j23} & SW_{j24} \\ SW_{j31} & SW_{j32} & SW_{j33} & SW_{j34} \\ SW_{j41} & SW_{j42} & SW_{j43} & SW_{j44} \end{pmatrix},$$

where

$$\begin{aligned}
SW_{j11} &= \sum_{l=1}^{n_j}(Y_{1lj} - \overline{Y_{1\cdot j}})^2 \approx \sum_{i=1}^{N} \widehat{p_{ij}}\widehat{\sigma_{11j}^2} \\
SW_{j22} &= \sum_{l=1}^{n_j}(Y_{2lj} - \overline{Y_{2\cdot j}})^2 \approx \sum_{i=1}^{N} \widehat{p_{ij}}\widehat{\sigma_{22j}^2} \\
SW_{j33} &= \sum_{l=1}^{n_j}(W_{1lj} - \overline{W_{1\cdot j}})^2 \approx \sum_{i=1}^{N} \widehat{p_{ij}} var(W_{i1}|X_i = j) \\
SW_{j44} &= \sum_{l=1}^{n_j}(W_{2lj} - \overline{W_{2\cdot j}})^2 \approx \sum_{i=1}^{N} \widehat{p_{ij}} var(W_{i2}|X_i = j) \\
SW_{j12} &= \sum_{l=1}^{n_j}(Y_{1lj} - \overline{Y_{1\cdot j}})(Y_{2lj} - \widehat{Y_{2\cdot j}}) \approx \sum_{i=1}^{N} \widehat{p_{ij}}\widehat{\sigma_{12j}^2} \\
SW_{j13} &= \sum_{l=1}^{n_j}(Y_{1lj} - \overline{Y_{1\cdot j}})(W_{1lj} - \overline{W_{1\cdot j}}) \approx \sum_{i=1}^{N} \widehat{p_{ij}}\sigma_{11j}\sqrt{var(W_{i1}|X_i = j)} * 0 = 0 \\
SW_{j14} &= \sum_{l=1}^{n_j}(Y_{1lj} - \overline{Y_{1\cdot j}})(W_{2lj} - \widehat{W_{2\cdot j}}) \approx \sum_{i=1}^{N} \widehat{p_{ij}}\widehat{\sigma_{11j}}\sqrt{var(W_{i2}|X_i = j)} * 0 = 0 \\
SW_{j23} &= \sum_{l=1}^{n_j}(Y_{2lj} - \overline{Y_{2\cdot j}})(W_{1lj} - \overline{W_{1\cdot j}}) \approx \sum_{i=1}^{N} \widehat{p_{ij}}\widehat{\sigma_{22j}}\sqrt{var(W_{i1}|X_i = j)} * 0 = 0 \\
SW_{j24} &= \sum_{l=1}^{n_j}(Y_{2lj} - \overline{Y_{2\cdot j}})(W_{2lj} - \overline{W_{2\cdot j}}) \approx \sum_{i=1}^{N} \widehat{p_{ij}}\widehat{\sigma_{22j}}\sqrt{var(W_{i2}|X_i = j)} * 0 = 0 \\
SW_{j34} &= \sum_{l=1}^{n_j}(W_{1lj} - \overline{W_{1\cdot j}})(W_{2lj} - \overline{W_{2\cdot j}}) \approx \sum_{i=1}^{N} \widehat{p_{ij}}\sqrt{var(W_{i1}|X_i = j)}\sqrt{var(W_2|X = j)} * 0 = 0
\end{aligned}$$

Since we have $p_{ij} = \mathbb{P}(X_i = j|\begin{pmatrix} W_{i1} \\ W_{i2} \end{pmatrix} = \begin{pmatrix} w_{i1} \\ w_{i2} \end{pmatrix}) = \frac{\exp(\alpha_j+\beta_{1j}w_{i1}+\beta_{2j}w_{i2})}{\sum_{q=1}^{J}\exp(\alpha_q+\beta_{1q}w_{i1}+\beta_{2q}w_{i2})}$, thus by Bayes Theorem, we can get

$$\begin{aligned}
f(\begin{pmatrix} w_1 \\ w_2 \end{pmatrix}|X_i = j) &= \frac{\mathbb{P}(X_i=j|W_i=w_i)f(w_i)}{\mathbb{P}(X_i=j)} = \frac{\mathbb{P}(X_i=j|W_i=w_i)f(w_i)}{\int\int f(X_i=j|W_i=w_i)f(w_i)dw_i} \\
&= \frac{\frac{\exp(\alpha_j+\beta_{1j}w_{i1}+\beta_{2j}w_{i2})}{\sum_{q=1}^{J}\exp(\alpha_q+\beta_{1q}w_{i1}+\beta_{2q}w_{i2})}*(2\pi)^{-1}\exp(-\frac{(w_{i1}-\nu_1)^2}{2\tau_{11}^2}-\frac{(w_{i2}-\nu_2)^2}{2\tau_{22}^2})}{\int\int\frac{\exp(\alpha_j+\beta_{1j}w_{i1}+\beta_{2j}w_{i2})}{\sum_{q=1}^{J}\exp(\alpha_q+\beta_{1q}w_{i1}+\beta_{2q}w_{i2})}*(2\pi)^{-1}\exp(-\frac{(w_{i1}-\nu_1)^2}{2\tau_{11}^2}-\frac{(w_{i2}-\nu_2)^2}{2\tau_{22}^2})dw_{i1}dw_{i2}}
\end{aligned}.$$

Also,

$$\begin{aligned}
var(W_{1i}|X_i = j) &= \mathbb{E}(W_{1i}^2|X_i = j) - (\mathbb{E}(W_{1i}|X_i = j))^2 \\
&= \int w_{1i}^2 f(w_{1i}|X_i = j)dw_{1i} - (\int w_{1i}f(w_{1i}|X_i = j)dw_{1i})^2
\end{aligned}.$$

The integral above could be solved by Monte Carlo Integration. In addition, these quantities also depend on the estimation of α_j and β_j. For simulation convenience, we could estimate the conditional variance of W_{ij} as following:

$$\widehat{var(W_{i1}|X_i} = j) = \frac{1}{\sum_{i=1}^{n}\widehat{p_{ij}}}\sum_{i=1}^{n}\widehat{p_{ij}}(w_{i1} - \frac{\sum_{i=1}^{n}\widehat{p_{ij}}w_{i1}}{\sum_{i=1}^{n}\widehat{p_{ij}}})^2.$$

For between component variability, we have

$$SSbetween = \begin{pmatrix} B_{11} & B_{12} & B_{13} & B_{14} \\ B_{21} & B_{22} & B_{23} & B_{24} \\ B_{31} & B_{32} & B_{33} & B_{34} \\ B_{41} & B_{42} & B_{43} & B_{44} \end{pmatrix},$$

where

$$\begin{aligned}
B_{11} &= \textstyle\sum_{j=1}^{J} n_j(\overline{Y_{1\cdot j}} - \overline{Y_{1\cdot\cdot}})^2 \approx \sum_{j=1}^{J}\sum_{i=1}^{N} \widehat{p_{ij}}(\widehat{\mu_{1j}} - \overline{Y_1})^2 \\
B_{22} &= \textstyle\sum_{j=1}^{J} n_j(\overline{Y_{2\cdot j}} - \overline{Y_{2\cdot\cdot}})^2 \approx \sum_{j=1}^{J}\sum_{i=1}^{N} \widehat{p_{ij}}(\widehat{\mu_{2j}} - \overline{Y_2})^2 \\
B_{33} &= \textstyle\sum_{j=1}^{J} n_j(\overline{W_{1\cdot j}} - \overline{W_{1\cdot\cdot}})^2 \approx \sum_{j=1}^{J}\sum_{i=1}^{N} \widehat{p_{ij}}(\widehat{\mathbb{E}(W_{1i}|X_i = j)} - \overline{W_1})^2 \\
B_{44} &= \textstyle\sum_{j=1}^{J} n_j(\overline{W_{2\cdot j}} - \overline{W_{2\cdot\cdot}})^2 \approx \sum_{j=1}^{J}\sum_{i=1}^{N} \widehat{p_{ij}}(\widehat{\mathbb{E}(W_{2i}|X_i = j)} - \overline{W_2})^2 \\
B_{12} &\approx \textstyle\sum_{j=1}^{J}\sum_{i=1}^{N} \widehat{p}_j(\widehat{\mu_{1j}} - \overline{Y_1})(\widehat{\mu_{2j}} - \overline{Y_2}) \\
B_{13} &\approx \textstyle\sum_{j=1}^{J}\sum_{i=1}^{N} \widehat{p_{ij}}(\widehat{\mu_{1j}} - \overline{Y_1})(\widehat{\mathbb{E}(W_{1i}|X_i = j)} - \overline{W_1}) \\
B_{14} &\approx \textstyle\sum_{j=1}^{J}\sum_{i=1}^{N} \widehat{p_{ij}}(\widehat{\mu_{1j}} - \overline{Y_1})(\widehat{\mathbb{E}(W_{2i}|X_i = j)} - \overline{W_2}) \\
B_{23} &\approx \textstyle\sum_{j=1}^{J}\sum_{i=1}^{N} \widehat{p_{ij}}(\widehat{\mu_{2j}} - \overline{Y_2})(\widehat{\mathbb{E}(W_{1i}|X_i = j)} - \overline{W_1}) \\
B_{24} &\approx \textstyle\sum_{j=1}^{J}\sum_{i=1}^{N} \widehat{p_{ij}}(\widehat{\mu_{2j}} - \overline{Y_2})(\widehat{\mathbb{E}(W_{2i}|X_i = j)} - \overline{W_2}) \\
B_{34} &\approx \textstyle\sum_{j=1}^{J}\sum_{i=1}^{N} \widehat{p_{ij}}(\widehat{\mathbb{E}(W_{1i}|X_i = j)} - \overline{W_1})(\widehat{\mathbb{E}(W_{2i}|X_i = j)} - \overline{W_2}).
\end{aligned}$$

Thus, the total variability is

$$\begin{aligned}
SStotal_J &= \textstyle\sum_{j=1}^{J} SSwithin_j + SSbetween \\
&= \textstyle\sum_{j=1}^{J} \begin{pmatrix} SW_{j11} & SW_{j12} & SW_{j13} & SW_{j14} \\ SW_{j21} & SW_{j22} & SW_{j23} & SW_{j24} \\ SW_{j31} & SW_{j32} & SW_{j33} & SW_{j34} \\ SW_{j41} & SW_{j42} & SW_{j43} & SW_{j44} \end{pmatrix} + \begin{pmatrix} B_{11} & B_{12} & B_{13} & B_{14} \\ B_{21} & B_{22} & B_{23} & B_{24} \\ B_{31} & B_{32} & B_{33} & B_{34} \\ B_{41} & B_{42} & B_{43} & B_{44} \end{pmatrix}.
\end{aligned}$$

Then the fraction of within-component variability to total variability for the estimated J-component mixture model is

$$\xi_J = \frac{|SSwithin|}{|SStotal_J|},$$

where $|SSwithin_J|$ and $|SStotal_J|$ are the determinant of matrix $SSwithin_J$ and $SStotal_J$, respectively. Following to Pilla and Charnigo(2007), the penalty statistic can be defined as:

$$\Lambda(Y,W) := \sum_{J\in I} \xi_J \mathbb{P}(Jcomponents),$$

where $\mathbb{P}(Jcomponents)$ is the prior of J-component model, the prior-weighted average fraction of within-component variability to total variability over the $|I|$ estimated mixture models. First we claim that $|SStotal_J| = |SSwithin_J + SSbetween_j| \geq |SSwithin_J| + |SSbetween_J|$, then $\Lambda(Y,W) \in [\mathbb{P}(1component), 1]$. Suppose there exist $1 > \varepsilon > 0$ such that $\mathbb{P}(1component) \geq \varepsilon$, then we have $\Lambda(Y,W) \in [\varepsilon, 1]$. A larger penalty statistic suggests less heterogeneity.

Since according to Marcus and Minc(1964),

$$\begin{aligned} \det(A+B) &= \det(A)\det(I + A^{-1/2}BA^{-1/2}) \\ &\geq \det A[1 + \det(A^{-1/2}BA^{-1/2})] \\ &= \det(A) + \det(A^{-1/2})\det(A^{-1/2}BA^{-1/2})\det(A^{-1/2}) \\ &= \det(A) + \det(B) \end{aligned}.$$

Then the above claim is proved. Furthermore, define the bivariate ratio function:

$$B(n,\kappa) := \frac{\Phi((\log n)^\kappa) - \Phi(1)}{1 - \Phi(1)}$$

for $n > \exp(2)$, $\kappa \in [\varepsilon, 1]$ and $\Phi(\cdot)$ is the cumulative distribution function of standard Gaussian distribution. As mentioned in Pilla and Charnigo(2007), $B(n,\kappa)$ is non-negative and increasing in both n and κ. Also for a fixed $\kappa \in [\varepsilon, 1]$, $\lim_{n\to\infty} B(n,\kappa) = 1$. Then let

$$P_n(Y,W) := (\log n)^{B(n,\Lambda(Y,W))}.$$

We could then define the singular flexible information criterion (sFLIC) as

Definition 1. *The singular flexible information criterion (sFLIC) of model M_j is:*

$$\begin{aligned} sFLIC_j &= \widehat{l(M_j)} + \lambda_j(\pi_0)P_n(Y,W) \\ &= \widehat{l(M_j)} + \lambda_j(\pi_0)(\log n)^{B(n,\Lambda(Y,W))}, \end{aligned}$$

where $\widehat{l(M_j)}$ is the maximum log-likelihood of model M_j, $\lambda_j(\pi_0) \in [0, d_j/2]$ is the learning coefficient(here we use λ_j as a short form of λ_{ij} mentioned in sBIC), and d_j is the dimension of the parameter space Ω_j.

The singular flexible information criterion is a modified flexible information criterion, which has a mild penalty term since $\lambda_j(\pi_0) \in [0, d_j/2]$. Also, following Pilla and Charnigo(2007), since $0 < B(n, \Lambda(Y, W)) \leq 1$, then for large n, the penalty term approximately equal to $\lambda_j \log n$, which equals to the penalty of sBIC, unless $\Lambda(Y, W)$ is very small. Also notice that, if n is small, the penalty term is much less than $\lambda_j \log n$, while for moderate n, it is sensitive to $\Lambda(Y, W)$. Thus for moderate n, if the data indicates a strong heterogeneity then the criterion has a light penalty.

Referring to the birth weight data, we can suppose Y_1 is the weight of new born infant, Y_2 stands for the obstetric gestation, W_1 is the age, and W_2 is the years of smoking.

5.3 Consistency of sFLIC

In this section, we show consistency results for the sFLIC from Definition 5.1. First, since we assume the same model, we have assumptions (A1) to (A3) from Chapter 4. Then for a finite set of models $\{M_i : i \in I\}$ and fixed data-generating distribution $\pi_0 \in \bigcup_{i \in I} M_i$, we have the following theorem:

Theorem 5.3.1. *(Consistency). Let M_i be the model selected by the sFLIC,*

$$i = \arg\max_{j \in I} sFLIC(M_j).$$

Under assumptions (A1)-(A3), the probability that M_i is a true model of minimal Bayes complexity tends to 1 as $n \to \infty$.

Following Drton and Plummer(2013), to show the theorem, it is sufficient to show that 1) the sFLIC of any true model is asymptotically larger that of any false model; 2) the sFLIC of a true model can be asymptotically maximal only if the model minimizes Bayes complexity among all true models. We will prove the two parts in the following lemmas.

Lemma 5.3.2. *If M_i is true a model and M_k is a false model, then under assumption (A2)*

$$\mathbb{P}(sFLIC(M_i) > sFLIC(M_k)) \to 1,$$

as $n \to \infty$.

Proof. Since we have shown in Chapter 4 that assumption (A2) holds for our hierarchical model with varying coefficient:

$$\mathbb{P}(sBIC(M_i) > sBIC(M_k)) \to 1,$$

as $n \to \infty$. Moreover, by Drton(2013) we have $\exp(sBIC(M_i)) = o(\exp(sBIC(M_k)))$. Also we already know that

$$\begin{aligned} sBIC(M_i) &= \widehat{l(M_i)} - \lambda_{ij} \log n \\ sFLIC(M_i) &= \widehat{l(M_i)} - \lambda_{ij} (\log n)^{B(n,\Lambda)}. \end{aligned}$$

We then prove

$$\frac{\exp(sFLIC(M_i))}{\exp(sFLIC(M_k))} = o_p(1).$$

Since

$$\frac{\exp(sFLIC(M_i))}{\exp(sFLIC(M_k))} = \frac{\exp(sFLIC(M_i))}{\exp(sBIC(M_i))} \frac{\exp(sBIC(M_i))}{\exp(sBIC(M_k))} \frac{\exp(sBIC(M_k))}{\exp(sFLIC(M_k))},$$

and note that $\frac{\exp(sFLIC(M_i))}{\exp(sBIC(M_i))} > 0$, $\frac{\exp(sBIC(M_k))}{\exp(sFLIC(M_k))} = o_p(1)$ and $\frac{\exp(sBIC(M_i))}{\exp(sBIC(M_k))} = o_p(1)$. Since $sBIC(M_j) < sFLIC(M_j)$, then we only need to show that for any model M_j, $\frac{\exp(sFLIC(M_j))}{\exp(sBIC(M_j))}$ is bounded($= O_p(1)$).

Since

$$\frac{\exp(sBIC(M_j)}{\exp(sFLIC(M_j))} = \frac{\widehat{L(M_j)} \exp(\lambda_j \log n)}{\widehat{L(M_j)} \exp(\lambda_j (\log n)^{B(n,\Lambda)})},$$

it suffices to show that

$$\frac{\exp(\lambda_j \log n)}{\exp(\lambda_j (\log n)^{B(n,\Lambda)})} = \exp(\lambda_j (\log n - (\log n)^{B(n,\Lambda)})) \to 1$$

in probability as $n \to \infty$.

Since we have that $\Lambda \in [\varepsilon, 1]$, then

$$\log n - (\log n)^{B(n,\Lambda)} \geq 0.$$

We also have that

$$\frac{(\log n)^{B(n,\Lambda)}}{\log n} = (\log n)^{B(n,\Lambda)-1}$$

Since according to Fan(2014),

$$B(n, \Lambda) \leq 1 - \frac{\exp(-(\log n)^{2\Lambda}/2)}{1 - \Phi(1)},$$

details are shown in Fan(2014) page 86,thus $B(n, \Lambda)-1$ converges to 0 faster than $\log n$ diverges to infinity. Actually, $B(n, \Lambda) - 1$ converges to 0 faster than $\log n * \log \log n$ diverges to infinity.

We have

$$\lim_{n\to\infty} (1 - B(n, \Lambda)) \log n \log \log n \leq \lim_{n\to\infty} \frac{\log n \log \log n}{C \exp((\log n)^{2\Lambda}/2)},$$

where $C > 0$ is a constant. Note that, since $\Lambda \in [\varepsilon, 1]$ then $\log n \geq (\log n)^{2\Lambda-1} \geq (\log n)^{2\varepsilon-1} > (\log n)^{-1}$. By L'Hospital's Rule:

$$0 \leq \lim_{n\to\infty} (1 - B(n, \Lambda)) \log n \log \log n \leq \lim_{n\to\infty} \frac{\frac{1}{n} \log \log n + \frac{1}{n}}{C \exp((\log n)^{2\varepsilon}/2) * \varepsilon(\log n)^{2\varepsilon-1} * \frac{1}{n}}.$$

Case 1: if $2\varepsilon - 1 \leq 0$ then as $n \to \infty$:

$$\frac{\log \log n + 1}{C \exp((\log n)^{2\varepsilon}/2) * \varepsilon(\log n)^{2\varepsilon-1}} \to 0;$$

Case 2: if $2\varepsilon - 1 > 0$ then define $\log n := u$,

$$\frac{\frac{1}{n} \log \log n + \frac{1}{n}}{C \exp((\log n)^{2\varepsilon}/2) * \varepsilon(\log n)^{2\varepsilon-1} * \frac{1}{n}} = \frac{(\log u + 1) * u^{1-2\varepsilon}}{C \exp(u^{2\varepsilon}/2) * \varepsilon} \to 0$$

as $u \to \infty$.

Thus, $(B(n, \Lambda) - 1) * \log n \log \log n \to 0$ in probability as $n \to \infty$.

Next,

$$\log n - (\log n)^{B(n,\Lambda)} = \log n(1 - (\log n)^{B(n,\Lambda)-1}).$$

We have

$$(\log n)^{(B(n,\Lambda)-1)} = \exp((B(n,\Lambda)-1)\log\log n).$$

Next we apply Taylor's expansion:

$$\exp((B(n,\Lambda)-1)\log\log n) = 1+(B(n,\Lambda)-1)\log\log n+O_p(((B(n,\Lambda)-1)\log\log n)^2).$$

Thus,

$$\log n(1-(\log n)^{(B(n,\Lambda)-1)}) = (1-B(n,\Lambda))\log n\log\log n+\log n*O_p(((B-1)\log\log n)^2)] \to 0,$$

Thus, $\frac{sFLIC(M_j)}{sBIC(M_j)}$ is bounded, then

$$\frac{\exp(sFLIC(M_i))}{\exp(sFLIC(M_k))} = \frac{\exp(sFLIC(M_i))}{\exp(sBIC(M_i))}\frac{\exp(sBIC(M_i))}{\exp(sBIC(M_k))}\frac{\exp(sBIC(M_k))}{\exp(sFLIC(M_k))} = o_p(1).$$

□

Thus, we have shown that the sFLIC of any true model is asymptotically larger than that of any false model.

Lemma 5.3.3. *Suppose $\pi_0 \in M_k$, but M_k does not minimize the Bayes complexity among all true models, assume there exist a true model M_i which minimize the Bayes complexity such that*

$$\mathbb{P}(sFLIC(M_i) > sFLIC(M_k)) \to 1,$$

as $n \to \infty$.

Proof. As shown by Drton and Plumner(2013) Proposition 4.2, the conclusion in Lemma 5.3.2 holds for any model M_j, as in the proof of Lemma 5.3.2 then imply,

$$\mathbb{P}(sFLIC(M_i) > sFLIC(M_k)) \to 1,$$

as $n \to \infty$. □

5.4 Simulation study

In simulation study, we generate data from the hierarchical normal mixture model as following first generate $\mathbf{W_i}$ from multivariate normal with mean $\boldsymbol{\nu}$ and covariance matrix $\boldsymbol{\tau}$. Then calculate $P(X_i = j|W_i)$ with given parameter a, b_1 and b_2. Then we generate Y_i given $X_i = j$ with mean $\boldsymbol{\mu_j}$ and covariance matrix $\boldsymbol{\Sigma_j}$ from multivariate normal distribution, with $j = 1, 2, ...m$ and number of non-redundant components $m \in \{2, 3, 4, 5\}$. We take sample size to $n = 10000$ then calculate the AIC, BIC and sBIC for each candidate model. For each combination of m and n, we generate 10 datasets with following parameter settings and compare the results.

2 components model: $\boldsymbol{\mu_1} = (-2, 1)^T$, $\boldsymbol{\mu_2} = (2, 3)^T$,

$$\boldsymbol{\Sigma_1} = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}, \boldsymbol{\Sigma_2} = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix},$$

$\mathbf{a} = (0.9, 0)$, $\mathbf{b_1} = (1.1, 0)$ and $\mathbf{b_2} = (2, 0)$.

3 components model: $\boldsymbol{\mu_1} = (-2, 2)^T$, $\boldsymbol{\mu_2} = (2, 4)^T$, $\boldsymbol{\mu_3} = (5, 7)^T$,

$$\boldsymbol{\Sigma_1} = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}, \boldsymbol{\Sigma_2} = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}, \boldsymbol{\Sigma_3} = \begin{pmatrix} 1 & 0.09 \\ 0.09 & 1 \end{pmatrix},$$

$\mathbf{a} = (0.9, 0.8, 0)$, $\mathbf{b_1} = (1.1, 1.2, 0)$ and $\mathbf{b_2} = (2, 1.8, 0)$.

4 components model: $\boldsymbol{\mu_1} = (-3, 3)^T$, $\boldsymbol{\mu_2} = (2, 1)^T$, $\boldsymbol{\mu_3} = (6, 5)^T$, $\boldsymbol{\mu_4} = (9, 9)^T$,

$$\boldsymbol{\Sigma_1} = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}, \boldsymbol{\Sigma_2} = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}, \boldsymbol{\Sigma_3} = \begin{pmatrix} 1 & 0.09 \\ 0.09 & 1 \end{pmatrix}, \boldsymbol{\Sigma_4} = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix},$$

$\mathbf{a} = (0.9, 0.8, 0.8, 0)$, $\mathbf{b_1} = (1.1, 1.2, 1.2, 0)$ and $\mathbf{b_2} = (2, 1.8, 1.9, 0)$.

5 components model: $\boldsymbol{\mu_1} = (-4, -3)^T$, $\boldsymbol{\mu_2} = (-1, 1)^T$, $\boldsymbol{\mu_3} = (2, 5)^T$, $\boldsymbol{\mu_4} = (6, 9)^T$, $\boldsymbol{\mu_5} = (9, 11)^T$,

$$\boldsymbol{\Sigma_1} = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}, \boldsymbol{\Sigma_2} = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}, \boldsymbol{\Sigma_3} = \begin{pmatrix} 1 & 0.09 \\ 0.09 & 1 \end{pmatrix}, \boldsymbol{\Sigma_4} = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix},$$

$$\boldsymbol{\Sigma_5} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix},$$

$\mathbf{a} = (0.9, 0.8, 0.8, 0.9, 0)$, $\mathbf{b_1} = (1.1, 1.2, 1.2, 1.1, 0)$ and $\mathbf{b_2} = (2, 1.8, 1.9, 1.9, 0)$. Also, we have $\boldsymbol{\nu} = (0, 0)$, $\boldsymbol{\tau} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ for all models.

The contour plots of joint density of Y_1 and Y_2 are shown in figure 5.1:

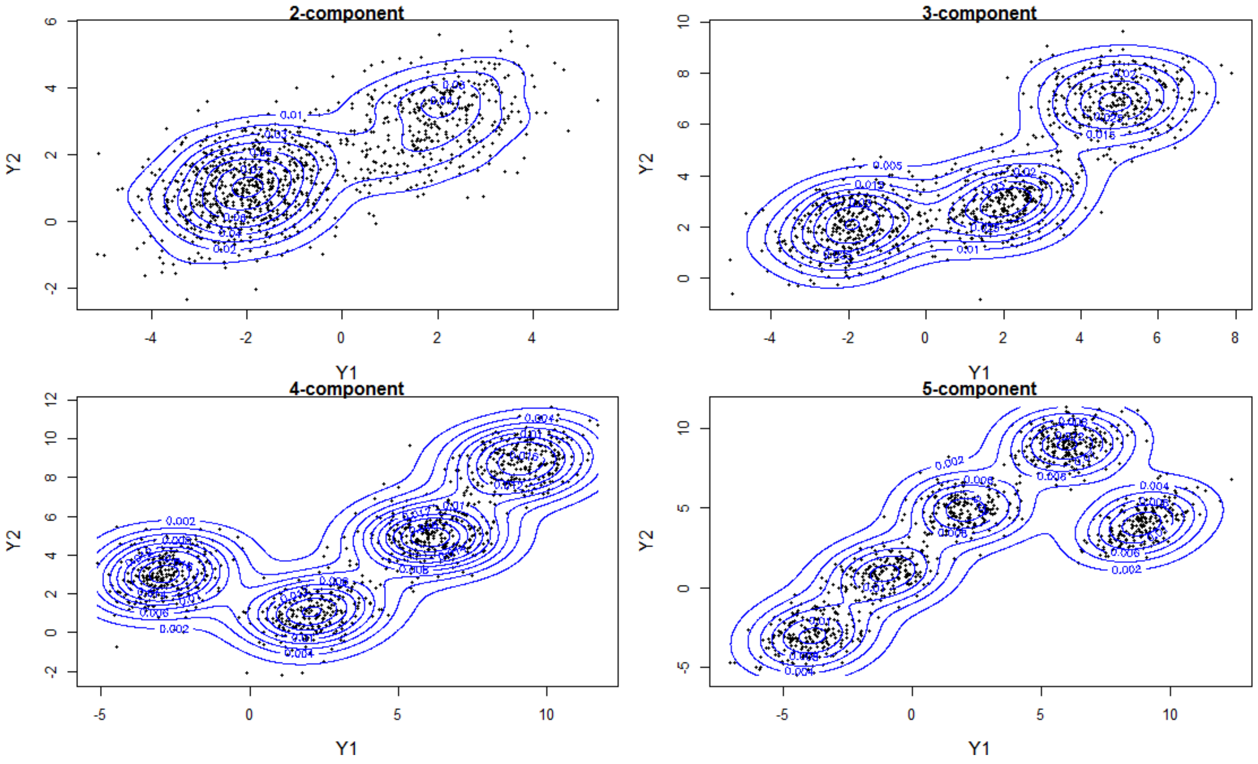

Figure 5.1: Contour plot of fitted joint density for sample data from the indicated distribution

The following table shows the result of model being selected by AIC, BIC, sBIC and sFLIC:

True/Select	AIC				BIC				sBIC				sFLIC			
	2	3	4	5	2	3	4	5	2	3	4	5	2	3	4	5
2 components	8	1	1	0	10	0	0	0	10	0	0	0	9	1	0	0
3 components	0	6	3	1	0	10	0	0	0	10	0	0	0	9	1	0
4 components	0	0	8	2	0	0	10	0	0	0	10	0	0	0	9	1
5 components	0	0	0	10	0	0	0	10	0	0	0	10	0	0	0	10

For large n, BIC and sBIC work well for detecting true number of components; AIC works worst and tend to select a larger number of components, while sFLIC is intermediate between AIC and BIC. More specifically, when true models have two components(M=2), AIC selects 8 out of 10 correctly, sFLIC selects 9 out of 10 correctly, while BIC and sBIC selects 10 correctly. When the true models have 3 components, AIC selects 6 out of 10 correctly, BIC and sBIC have correct classification rates of 100%, while sFLIC is intermediate between AIC and BIC or sBIC, it selects 9 out of 10 correctly. The situation is similar for true models with 4 components, sFLIC chooses 9 out of 10 correctly, which is between AIC(8) and BIC or sBIC(10). For 5 component models, all of these criteria work well. Since our candidate sets only contains up to 5 components model, then it is not possible to choose more components.

5.5 Real Data Applications

In this section, we apply our model 4.1 to NCHS' Vital Statistics Natality Birth Data from the National Vital Statistics System of the National Center for Health Statistics in year 2014. The data is publicly available online at

$$http://www.nber.org/data/vital-statistics-natality-data.html.$$

The data are based on information abstracted from birth certificates filed in vital statistics offices of each State and the District of Columbia.

We wish to analyze the relationship between gestational age and birth weight adjusted for other variables like mother's age and father's age. According to Charnigo et al. [2010], since records with birthweight less than 500 grams or gestational age less than 22 weeks were not consistently documented, then we select data with known birthweight between 500 and 5500 grams and gestational age larger than or equal to 22 weeks. Referring to our model, we treat mother's age(in years) and father's age(in years) as $\mathbf{W}$, gestational age(in weeks) and birthweight(in grams) as $\mathbf{Y}$. We randomly draw 10 samples of sizes $n = 500, 1000, 2500, 5000, 10000$ respectively and fit models with number of non-redundant components $m \in \{2, 3, 4, 5\}$. For each m and n combination, we use EM algorithm to estimate parameters and use AIC, BIC, sBIC and sFLIC to infer the true number of components. Table 5.1 summarizes the EM algorithm parameter estimates and Table 5.2-Table 5.5 shows the model selection results by information criterion average over 10 samples.

Table 5.1: Preferences of Model Selection Criteria in Real Data

Sample size	AIC select				BIC select				sBIC select				sFLIC select			
	2	3	4	5	2	3	4	5	2	3	4	5	2	3	4	5
n=500	1	6	2	1	10	0	0	0	9	1	0	0	8	2	0	0
n=1000	0	2	1	7	7	3	0	0	4	6	0	0	4	6	0	0
n=2500	0	0	5	5	0	0	8	2	0	0	8	2	0	0	8	2
n=5000	0	1	6	3	0	1	8	1	0	1	8	1	0	1	8	1
n=10000	0	0	2	8	0	0	5	5	0	0	5	5	0	0	5	5

Table 5.1 shows the results of 10 random samples. For small sample size(n=500 or n=1000), BIC, sBIC and sFLIC tend to select 2 or 3 component models. More specifically, when n=500, BIC selects all models as 2-component, sFLIC selects 2 out of 10 as 3-component models others as 2-component; AIC selects all models but tends to prefer to 3-component. When n=1000, AIC shows preference to 5 component models, BIC still favors 2-component models, but sBIC and sFLIC choose more 3-component than 2-component models. However, due to the numerical optimization and the inherent multi-modality of the likelihood function, the results may not be stable for small samples. As sample size increases to 2500, all criteria seem to choose more components, none of them selects 2 or 3 component models. AIC shows equal preference to 4 and 5 component models, while BIC, sBIC and sFLIC have similar results and choose more 4-component models. For n=5000, still BIC, sBIC and sFLIC agree with each other and tend to select more 4-component models. AIC as well selects 6 out of 10 as 4-component models, but still chooses 3 as 5-component models, which agrees with the theory that AIC is not stable when sample size is large as mentioned in Pilla and Charnigo [2007]. When sample size increases to n=10000, BIC, sBIC and sFLIC shows equal preference for 4 and 5-component models, while AIC chooses 8 out of 10 as 5-component models, and 2 as 4-component models.

The scatter plots of real data and contour plots of the fitted densities for gestation and birthweight, marginalized over mother and father's age, are shown in Figure 5.2. The parameters for fitted density are from table 5.2 to table 5.5 using $n = 1000$, which is plotted against 1000 new observations not used to fit the model; the use of new data could provide insight into the generalizability of the model to new data.

For 2-component model, we see that the first component accounts for most of the low birthweight and premature babies, and has large variation in both Y_1 and Y_2 to capture the diversity of these cases. The second component accounts for the majority of the data, which captures most normal to high birthweight cases whose gestational ages are of full term or nearly so.

For the 3-component model, similar to the 2-component model, there is still one component with large variation accounting for cases with low birthweights and short gestational ages. One component accounts for the majority of the data. Besides that there is an additional component accounting for many cases with high gestational ages.

Fitted density of 4-component model seems more reasonable since the first component tends to capture more low birthweight and premature cases, while other three components together account for almost all normal to high birthweight cases whose gestational ages are of full term or nearly so.

The 5-component model seems a little over-fitted, since compared to 4-component model, it has an additional component to capture relatively low birthweight but normal gestational age cases, which are infrequent. This also shows the reason that BIC, sBIC and sFLIC tend to select 4-component model as the best fitted model.

Also, since the estimated b_1 and b_2 in the tables are very close to zero, we can conclude that mother's age and father's age have very little influence on component membership for the birthweight and gestational age. However, the model does not explore whether parents' age may influence birthweight and gestational ages within a component.

Table 5.2: Estimating Parameters in a 2-Component Mixture Model

Parameters	n=1000	n=2500	n=5000	n=10000
μ_1	$(36.8, 2753)^T$	$(36.4, 2862)^T$	$(37.3, 2848)^T$	$(37.4, 2853)^T$
μ_2	$(39.1, 3516)^T$	$(38.9, 3456)^T$	$(39.1, 3521)^T$	$(39.1, 3516)^T$
Σ_1	$\begin{pmatrix} 17.5 & 2395 \\ 2395 & 552236 \end{pmatrix}$	$\begin{pmatrix} 15.4 & 1971 \\ 1971 & 510546 \end{pmatrix}$	$\begin{pmatrix} 16.5 & 2247 \\ 2247 & 590887 \end{pmatrix}$	$\begin{pmatrix} 17.6 & 2217 \\ 2217 & 568091 \end{pmatrix}$
Σ_2	$\begin{pmatrix} 1.49 & 156.9 \\ 156.9 & 190990 \end{pmatrix}$	$\begin{pmatrix} 1.14 & 147.3 \\ 147.3 & 191553 \end{pmatrix}$	$\begin{pmatrix} 1.63 & 154.3 \\ 154.3 & 205321 \end{pmatrix}$	$\begin{pmatrix} 1.44 & 149.6 \\ 149.6 & 192252 \end{pmatrix}$
a	$(1.75, 0)$	$(2.44, 0)$	$(1.49, 0)$	$(1.24, 0)$
b_1	$(0.004, 0)$	$(0.002, 0)$	$(0.014, 0)$	$(0.010, 0)$
b_2	$(-0.005, 0)$	$(-0.003, 0)$	$(-0.006, 0)$	$(-0.007, 0)$
ν	$(28.79, 32.79)^T$	$(28.89, 32.50)^T$	$(28.57, 32.75)^T$	$(28.37, 32.89)^T$
τ	$\begin{pmatrix} 861.2 & 944.2 \\ 944.2 & 1436 \end{pmatrix}$	$\begin{pmatrix} 866.6 & 938.8 \\ 938.8 & 1387 \end{pmatrix}$	$\begin{pmatrix} 850.1 & 934.3 \\ 934.3 & 1442 \end{pmatrix}$	$\begin{pmatrix} 857.8 & 943.0 \\ 943.0 & 1443 \end{pmatrix}$

Table 5.3: Estimating Parameters in a 3-Component Mixture Model

Parameters	n=1000	n=2500	n=5000	n=10000
μ_1	$(36.7, 2772)^T$	$(36.9, 2817)^T$	$(37.2, 2808)^T$	$(36.9, 2791)^T$
μ_2	$(38.6, 3518)^T$	$(38.5, 3471)^T$	$(38.5, 3505)^T$	$(38.7, 3548)^T$
μ_3	$(39.1, 3072)^T$	$(39.5, 2941)^T$	$(39.3, 2986)^T$	$(39.2, 3035)^T$
Σ_1	$\begin{pmatrix} 21.7 & 2558 \\ 2558 & 803403 \end{pmatrix}$	$\begin{pmatrix} 21.4 & 2627 \\ 2627 & 851339 \end{pmatrix}$	$\begin{pmatrix} 23.2 & 3139 \\ 3139 & 836011 \end{pmatrix}$	$\begin{pmatrix} 19.9 & 2279 \\ 2279 & 724088 \end{pmatrix}$
Σ_2	$\begin{pmatrix} 1.28 & 105.9 \\ 105.9 & 182895 \end{pmatrix}$	$\begin{pmatrix} 1.33 & 109.4 \\ 109.4 & 185928 \end{pmatrix}$	$\begin{pmatrix} 1.27 & 115.1 \\ 115.1 & 183038 \end{pmatrix}$	$\begin{pmatrix} 1.02 & 97.8 \\ 97.8 & 174403 \end{pmatrix}$
Σ_3	$\begin{pmatrix} 3.13 & 432.9 \\ 432.9 & 151404 \end{pmatrix}$	$\begin{pmatrix} 2.77 & 582.2 \\ 582.2 & 192953 \end{pmatrix}$	$\begin{pmatrix} 2.61 & 546.6 \\ 546.6 & 199232 \end{pmatrix}$	$\begin{pmatrix} 3.26 & 433.1 \\ 433.1 & 141631 \end{pmatrix}$
a	$(-1.16, 1.02, 0)$	$(-1.21, 1.08, 0)$	$(-1.16, 0.93)$	$(-0.72, 1.39, 0)^T$
b_1	$(0.086, 0.28, 0)$	$(0.06, 0.05, 0)$	$(0.057, 0.050, 0)$	$(0.047, 0.059, 0)$
b_2	$(0.027, 0.039, 0)$	$(-0.015, 0.007, 0)$	$(0.011, 0.1, 0)$	$(0.013, 0.022, 0)$
ν	$(28.79, 32.79)^T$	$(28.89, 32.50)^T$	$(28.57, 32.75)^T$	$(28.37, 32.89)^T$
τ	$\begin{pmatrix} 861.2 & 944.2 \\ 944.2 & 1436 \end{pmatrix}$	$\begin{pmatrix} 866.6 & 938.8 \\ 938.8 & 1387 \end{pmatrix}$	$\begin{pmatrix} 850.1 & 934.3 \\ 934.3 & 1442 \end{pmatrix}$	$\begin{pmatrix} 857.8 & 943.0 \\ 943.0 & 1443 \end{pmatrix}$

Table 5.4: Estimating Parameters in a 4-Component Mixture Model

Parameters	n=1000	n=2500	n=5000	n=10000
μ_1	$(36.1, 2258)^T$	$(36.2, 2139)^T$	$(35.9, 2246)^T$	$(36.1, 2198)^T$
μ_2	$(38.1, 3160)^T$	$(37.9, 3283)^T$	$(37.6, 3182)^T$	$(38.4, 3200)^T$
μ_3	$(39.7, 3551)^T$	$(39.7, 3561)^T$	$(39.8, 3574)^T$	$(39.9, 3612)^T$
μ_4	$(38.8, 3281)^T$	$(38.9, 3031)^T$	$(38.8, 3051)^T$	$(38.9, 3176)^T$
Σ_1	$\begin{pmatrix} 20.1 & 2833 \\ 2833 & 459540 \end{pmatrix}$	$\begin{pmatrix} 20.7 & 2566 \\ 2566 & 495962 \end{pmatrix}$	$\begin{pmatrix} 18.6 & 2900 \\ 2900 & 458287 \end{pmatrix}$	$\begin{pmatrix} 20.6 & 2863 \\ 2863 & 458287 \end{pmatrix}$
Σ_2	$\begin{pmatrix} 14.5 & 996 \\ 996 & 479716 \end{pmatrix}$	$\begin{pmatrix} 11.61 & 704 \\ 704 & 414000 \end{pmatrix}$	$\begin{pmatrix} 10.60 & 696 \\ 696 & 477921 \end{pmatrix}$	$\begin{pmatrix} 11.3 & 675 \\ 675 & 392403 \end{pmatrix}$
Σ_3	$\begin{pmatrix} 2.76 & 103.3 \\ 103.3 & 204631 \end{pmatrix}$	$\begin{pmatrix} 1.22 & 139.9 \\ 139.9 & 207002 \end{pmatrix}$	$\begin{pmatrix} 1.38 & 142.9 \\ 142.9 & 193885 \end{pmatrix}$	$\begin{pmatrix} 2.97 & 93.2 \\ 93.2 & 170252 \end{pmatrix}$
Σ_4	$\begin{pmatrix} 2.55 & 227.1 \\ 227.1 & 122116 \end{pmatrix}$	$\begin{pmatrix} 3.97 & 247.3 \\ 247.3 & 145140 \end{pmatrix}$	$\begin{pmatrix} 3.08 & 306.0 \\ 306.0 & 138422 \end{pmatrix}$	$\begin{pmatrix} 2.33 & 239.5 \\ 239.5 & 119216 \end{pmatrix}$
a	$(-1.44, 1.056, 2.83, 0)$	$(-1.32, -0.62, 1.49, 0)$	$(-1.16, -0.22, 1.83, 0)$	$(-1.82, -0.62, 0.25, 0)$
b_1	$(-0.01, 0.05, 0.04, 0)$	$(0.06, 0.11, 0.08, 0)$	$(0.016, 0.2, 0.16, 0)$	$(0.05, 0.02, 0.06, 0)$
b_2	$(0.02, 0.22, 0.21, 0)$	$(0.04, 0.07, 0.05, 0)$	$(0.04, 0.05, 0.04, 0)$	$(0.01, 0.03, 0.03, 0)^T$
ν	$(28.79, 32.79)^T$	$(28.89, 32.50)^T$	$(28.57, 32.75)^T$	$(28.37, 32.89)^T$
τ	$\begin{pmatrix} 861.2 & 944.2 \\ 944.2 & 1436 \end{pmatrix}$	$\begin{pmatrix} 866.6 & 938.8 \\ 938.8 & 1387 \end{pmatrix}$	$\begin{pmatrix} 850.1 & 934.3 \\ 934.3 & 1442 \end{pmatrix}$	$\begin{pmatrix} 857.8 & 943.0 \\ 943.0 & 1443 \end{pmatrix}$

Table 5.5: Estimating Parameters in a 5-Component Mixture Model

Parameters	n=1000	n=2500	n=5000	n=10000
μ_1	$(35.3, 2214)^T$	$(35.8, 2196)^T$	$(35.7, 2179)^T$	$(35.5, 2232)^T$
μ_2	$(37.8, 3189)^T$	$(37.8, 3173)^T$	$(37.3, 3192)^T$	$(37.8, 3206)^T$
μ_3	$(38.9, 3319)^T$	$(39.2, 3324)^T$	$(39.1, 3279)^T$	$(38.9, 3317)^T$
μ_4	$(40.1, 3670)^T$	$(39.9, 3526)^T$	$(39.9, 3579)^T$	$(39.9, 3654)^T$
μ_5	$(38.8, 3173)^T$	$(38.8, 3146)^T$	$(38.7, 3078)^T$	$(38.8, 3130)^T$
Σ_1	$\begin{pmatrix} 20.2 & 2441 \\ 2441 & 520451 \end{pmatrix}$	$\begin{pmatrix} 20.3 & 2709 \\ 2709 & 536479 \end{pmatrix}$	$\begin{pmatrix} 20.1 & 2022 \\ 2022 & 534640 \end{pmatrix}$	$\begin{pmatrix} 20.8 & 2614 \\ 2614 & 519112 \end{pmatrix}$
Σ_2	$\begin{pmatrix} 9.03 & 517.4 \\ 517.4 & 323629 \end{pmatrix}$	$\begin{pmatrix} 8.16 & 408.3 \\ 408.3 & 29480 \end{pmatrix}$	$\begin{pmatrix} 8.63 & 530.3 \\ 530.3 & 335853 \end{pmatrix}$	$\begin{pmatrix} 9.06 & 529.9 \\ 529.9 & 322149 \end{pmatrix}$
Σ_3	$\begin{pmatrix} 0.91 & 67.4 \\ 67.4 & 105828 \end{pmatrix}$	$\begin{pmatrix} 1.20 & 74.05 \\ 74.05 & 121725 \end{pmatrix}$	$\begin{pmatrix} 1.23 & 91.2 \\ 91.2 & 109495 \end{pmatrix}$	$\begin{pmatrix} 0.68 & 71.49 \\ 71.49 & 123333 \end{pmatrix}$
Σ_4	$\begin{pmatrix} 1.49 & 94.1 \\ 94.1 & 190121 \end{pmatrix}$	$\begin{pmatrix} 1.13 & 87.79 \\ 87.79 & 180379 \end{pmatrix}$	$\begin{pmatrix} 1.51 & 102.1 \\ 102.1 & 210575 \end{pmatrix}$	$\begin{pmatrix} 1.99 & 92.44 \\ 92.44 & 190431 \end{pmatrix}$
Σ_5	$\begin{pmatrix} 5.52 & 304.2 \\ 304.2 & 189227 \end{pmatrix}$	$\begin{pmatrix} 5.27 & 290.7 \\ 290.7 & 217762 \end{pmatrix}$	$\begin{pmatrix} 6.49 & 239.4 \\ 239.4 & 171510 \end{pmatrix}$	$\begin{pmatrix} 6.92 & 350.0 \\ 350.0 & 182401 \end{pmatrix}$
a	$(0.71, 1.77, 3, 2.93, 0)$	$(1.21, 0.29, 1.57, 2.99, 0)$	$(0.91, 2.47, 2.98, 2.98, 0)$	$(1.79, 2.26, 3, 2.3, 0)$
b_1	$(0.09, 0.45, 0.22, 0.37, 0)$	$(0.38, 0.5, 0.37, 0.38, 0)$	$(0.28, 0.35, 0.33, 0.37, 0)$	$(0.26, 0.30, 0.24, 0.27, 0)$
b_2	$(0.05, 0.22, 0.12, 0.21, 0)$	$(0.10, 0.49, 0.21, 0.2, 0)$	$(0.09, 0.17, 0.14, 0.17, 0)$	$(0.08, 0.09, 0.17, 0.10, 0)$
ν	$(28.79, 32.79)^T$	$(28.89, 32.50)^T$	$(28.57, 32.75)^T$	$(28.37, 32.89)^T$
τ	$\begin{pmatrix} 861.2 & 944.2 \\ 944.2 & 1436 \end{pmatrix}$	$\begin{pmatrix} 866.6 & 938.8 \\ 938.8 & 1387 \end{pmatrix}$	$\begin{pmatrix} 850.1 & 934.3 \\ 934.3 & 1442 \end{pmatrix}$	$\begin{pmatrix} 857.8 & 943.0 \\ 943.0 & 1443 \end{pmatrix}$

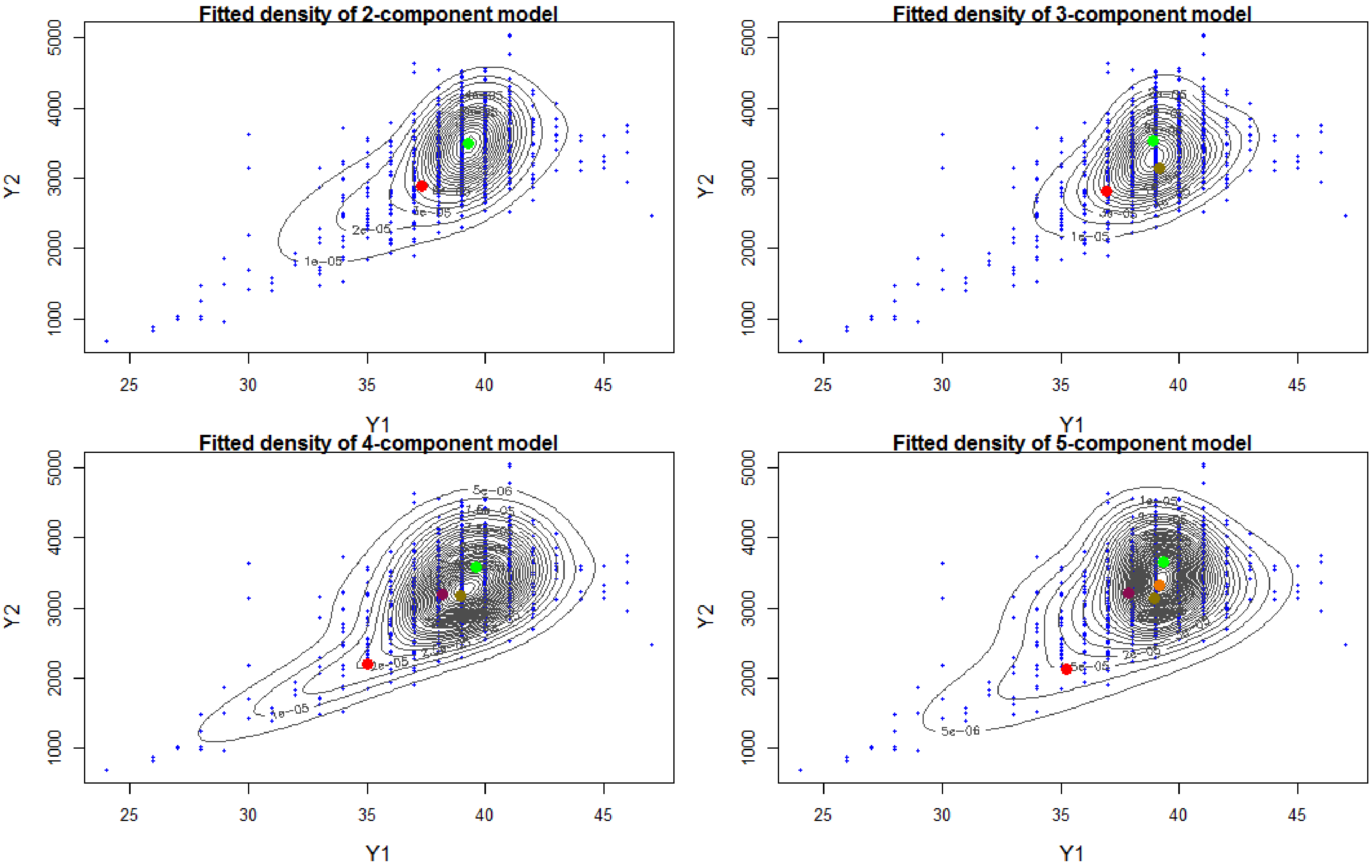

Figure 5.2: Contour plots of fitted joint density(n=1000), with component means in colored dots

5.6 Summary and Discussion

In this chapter, we develop a new data dependent information criterion sFLIC ,inspired in part by Pilla and Charnigo [2007] 's FLIC, for bivariate hierarchical mixture models 4.1 with varying weights. Our work is different from Pilla and Charnigo [2007] in the following three aspects: 1)our hierarchical mixture model has varying weights; 2) our method is derived recognizing the singular structure of the models; 3) we consider data with a vector response and a vector covariate.

In section 5.3, we proved asymptotic properties for our sFLIC criterion which accommodates singularity of our hierarchical model (4.1). In section 5.4 simulation study, we showed that sFLIC works as well as sBIC for large samples. In section 5.5, we applied sFLIC to birthweight data from the National Center for Health Statistics in year 2014. We saw that sFLIC shows concordant results with BIC and sBIC, which tend to choose 4-component models as best for birthweight and gestational age. Surprisingly, this result agrees with Charnigo et al. [2010], who considered only birthweight as a variable and has no covariate.

But there are some limitations to this study. First non-uniqueness of local maxima of the likelihood (Karlis and Xekalaki [2003]) and numerical optimization imply the results may not be stable, especially for small samples; estimates may be local maxima instead of global maxima. Second,since for birthweight and gestational age data, we only include those cases with birthweight between 500 and 5500 grams and gestational age larger than or equal to 22 weeks, the model is technically no longer a normal mixture, but rather a truncated normal mixture. As such, we anticipate that model selection criterion will try to capture empirical distribution which trend to select more component for large n, meaning that the theoretical consistency may not be observed in practice.

Bibliography

Kamel Jedidi, Harsharanjeet S Jagpal, and Wayne S DeSarbo. Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science*, 16(1):39–59, 1997.

Christopher G Lamoureux and William D Lastrapes. Endogenous trading volume and momentum in stock-return volatility. *Journal of Business & Economic Statistics*, 12(2):253–260, 1994.

PM Atkinson, MEJ Cutler, and H Lewis. Mapping sub-pixel proportional land cover with avhrr imagery. *International Journal of Remote Sensing*, 18(4):917–935, 1997.

Richard Charnigo, Lorie W Chesnut, Tony LoBianco, and Russell S Kirby. Thinking outside the curve, part ii: modeling fetal-infant mortality. *BMC pregnancy and childbirth*, 10(1):44, 2010.

Allen J Wilcox and Ian T Russell. Birthweight and perinatal mortality: Iii. towards a new method of analysis. *International Journal of Epidemiology*, 15(2):188–196, 1986.

Timothy B Gage and Gene Therriault. Variability of birth-weight distributions by sex and ethnicity: Analysis using mixture models. *Human biology*, pages 517–534, 1998.

Jurg Ott. *Analysis of human genetic linkage.* JHU Press, 1999.

Richard Charnigo and Ramani S Pilla. Semiparametric mixtures of generalized exponential families. *Scandinavian Journal of Statistics*, 34(3):535–551, 2007.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26(2):195–239, 1984.

Hanfeng Chen, Jiahua Chen, and John D Kalbfleisch. A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(1):19–29, 2001.

PM Hartigan. Algorithm as 217: Computation of the dip statistic to test for unimodality. *Applied Statistics*, pages 320–325, 1985.

Geoffrey J McLachlan and Kaye E Basford. Mixture models: Inference and applications to clustering. *Applied Statistics*, 1988.

Bruce G Lindsay. Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–163. JSTOR, 1995.

Didier Dacunha-Castelle, Elisabeth Gassiat, et al. Testing the order of a model using locally conic parametrization: population mixtures and stationary arma processes. *The Annals of Statistics*, 27(4):1178–1209, 1999.

Hongying Dai and Richard Charnigo. Omnibus testing and gene filtration in microarray data analysis. *Journal of Applied Statistics*, 35(1):31–47, 2008.

Richard Charnigo and Jiayang Sun. Testing homogeneity in a mixture distribution via the l 2 distance between competing models. *Journal of the American Statistical Association*, 99(466):488–498, 2004.

Richard Charnigo and Jiayang Sun. Asymptotic relationships between the d-test and likelihood ratio-type tests for homogeneity. *Statistica Sinica*, 20(2):497, 2010.

Richard Charnigo and Jiayang Sun. Testing homogeneity in discrete mixtures. *Journal of Statistical Planning and Inference*, 138(5):1368–1388, 2008.

Jiahua Chen and Pengfei Li. Hypothesis test for normal mixture models: The em approach. *The Annals of Statistics*, pages 2523–2542, 2009.

Mário AT Figueiredo and Jose Leitã. Simulated tearing: An algorithm for discontinuity-preserving visual surface reconstruction. In *Computer Vision and Pattern Recognition, 1993. Proceedings CVPR'93., 1993 IEEE Computer Society Conference on*, pages 28–33. IEEE, 1993.

Htrotugu Akaike. Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, 60(2):255–265, 1973.

Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

Parhasarathi Lahiri. Model selection. IMS, 2001.

Christine Keribin. Consistent estimation of the order of mixture models. *The Indian Journal of Statistics, Series A*, pages 49–66, 2000.

RS Pilla and R Charnigo. Consistent estimation and model selection in semiparametric mixtures. 2006.

Mathias Drton and Martyn Plummer. A bayesian information criterion for singular models. *arXiv preprint arXiv:1309.0911*, 2013.

Sumio Watanabe. *Algebraic geometry and statistical learning theory*, volume 25. Cambridge University Press, 2009a.

Qian Fan. Normal mixture and contaminated model with nuisance parameter and applications. 2014.

David B Allison, Gary L Gadbury, Moonseong Heo, José R Fernández, Cheol-Koo Lee, Tomas A Prolla, and Richard Weindruch. A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis*, 39(1):1–20, 2002.

Hongying Dai and Richard Charnigo. Contaminated normal modeling with application to microarray data analysis. *Canadian Journal of Statistics*, 38(3):315–332, 2010.

Linyong Mao, Chris Mackenzie, Jung H Roh, Jesus M Eraso, Samuel Kaplan, and Haluk Resat. Combining microarray and genomic data to predict dna binding motifs. *Microbiology*, 151(10):3197–3213, 2005.

Richard Charnigo, Qian Fan, Douglas Bittel, and Hongying Dai. Testing unilateral versus bilateral normal contamination. *Statistics & Probability Letters*, 83(1):163–167, 2013.

Jelle J Goeman and Peter Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, 2007.

Thomas Shelburne Ferguson. *A course in large sample theory*, volume 49. Chapman & Hall London, 1996.

Georges Aad, B Abbott, J Abdallah, O Abdinov, R Aben, M Abolins, OS AbouZeid, H Abramowicz, H Abreu, R Abreu, et al. Combined measurement of the higgs boson mass in pp collisions at s= 7 and 8 tev with the atlas and cms experiments. *Physical Review Letters*, 114(19), 2015.

Stephen W Raudenbush and Anthony S Bryk. *Hierarchical linear models: Applications and data analysis methods*, volume 1. Sage, 2002.

Russell W Rumberger. Dropping out of middle school: A multilevel analysis of students and schools. *American educational Research journal*, 32(3):583–625, 1995.

Daniel Nagin and Richard E Tremblay. Trajectories of boys' physical aggression, opposition, and hyperactivity on the path to physically violent and nonviolent juvenile delinquency. *Child development*, 70(5):1181–1196, 1999.

Daniel S Nagin and Richard E Tremblay. Analyzing developmental trajectories of distinct but related behaviors: a group-based method. *Psychological methods*, 6 (1):18, 2001.

Richard Charnigo, Richard Kryscio, Michael T Bardo, Donald Lynam, and Rick S Zimmerman. Joint modeling of longitudinal data in multiple behavioral change. *Evaluation & the health professions*, 34(2):181–200, 2011.

Mathias Drton. Likelihood ratio tests and singularities. *The Annals of Statistics*, pages 979–1012, 2009.

Sumio Watanabe. A limit theorem in singular regression problem. *arXiv preprint arXiv:0901.2376*, 2009b.

Jean-Marc Azaïs, Elisabeth Gassiat, and Cécile Mercadier. The likelihood ratio test for general mixture models with or without structural parameter. *ESAIM: Probability and Statistics*, 13:301–327, 2009.

Miki Aoyagi. A bayesian learning coefficient of generalization error and vandermonde matrix-type singularities. *Communications in StatisticsTheory and Methods*, 39 (15):2667–2687, 2010.

Elisabeth Gassiat. Likelihood ratio inequalities with applications to various mixtures. In *Annales de l'IHP Probabilités et statistiques*, volume 38, pages 897–906, 2002.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Paul Doukhan, Pascal Massart, and Emmanuel Rio. Invariance principles for absolutely regular empirical processes. In *Annales de l'IHP Probabilités et statistiques*, volume 31, pages 393–427, 1995.

RS Pilla and R Charnigo. Consistent estimation and model selection in semiparametric mixtures. 2007.

George Henry Dunteman. *Introduction to multivariate analysis*. Sage Publications, Inc, 1984.

Dimitris Karlis and Evdokia Xekalaki. Choosing initial values for the em algorithm for finite mixtures. *Computational Statistics & Data Analysis*, 41(3):577–590, 2003.

Vita

Education

- Master of Science in Statistics, University of Kentucky, 2011-2013
- Bachelor of Science in Mathematics, Shandong University, 2007-2011

Employment

- Statistics Consultant, University of Kentucky, 2013-2015
- Research and Development Scientist Summer Intern, J.M Sumkers, 2015
- Research Assistant, University of Kentucky, 2014-2015
- Teaching Assistant, University of Kentucky, 2011-2013