Stony Brook University

Restricted Mixture Linear Regression Models: Estimation, Power and Sample Size Calculations

A Dissertation Presented

by

Zhongming Yang

to

The Graduate School

in Partial fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

May 2007

Stony Brook University

The Graduate School

Zhongming Yang

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree,

hereby recommend acceptance of this dissertation.

Stephen J. Finch, Professor,

Department of Applied Mathematics & Statistics

Nancy R. Mendell, Professor,

Department of Applied Mathematics & Statistics

John Chen, Associate Professor,

Department of Preventive Medicine

Derek Gordon, Associate Professor,

Department of Genetics, Rutgers, The State University of New Jersey

This dissertation is accepted by the Graduate School

Lawrence Martin

Dean of the Graduate School

Abstract of the Dissertation

Restricted Mixture Linear Regression Models:
Estimation, Power and Sample Size Calculations

by

Zhongming Yang

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

2007

This dissertation focuses on restricted mixture linear regression models (RMLRM). RMLRM are mixture models that have some regression parameters set to be equal across the mixture components, while other regression parameters (unrestricted) may be unequal across components. We use the Expectation-Maximization (EM) algorithm to calculate the maximum likelihood estimates (MLE) for the regression parameters and mixing proportions. We provide the standard errors for the MLE. We further provide two EM initialization procedures for two specific RMLRM: the mixture intercept model (MIM), where only intercepts may differ across components, and the mixture slope model (MSM), where only slopes may differ across components. We also propose two approximate formulas to calculate the power and sample size for two-component normal mixture model (NMM), MIM and MSM.

Through simulation studies, we (1) investigate the null LRTS distributions of the test

for two-component mixture using NMM, MIM and MSM models; (2) verify that RMLRM techniques are more powerful to detect some specific mixtures compared to the unrestricted mixture linear regression model; (3) verify that our power and sample size formulas are accurate under a broad range of sample sizes.

We also apply our RMLRM in two case studies. These case studies document us that RMLRM is an useful tool to detect different mixture mechanisms.

Contents

List of Figures

List of Tables

Acknowledgements

I am deeply grateful to my advisor, Professor Stephen J. Finch, for his great encouragement and guidance in these past two and half years. The fruitful and inspiring research work I have done with Professor Finch helped me truly open a new frontier for my future career development. I also thank my committee members, Professor Nancy R. Mendell, Professor John Chen and Professor Derek Gordon. Their advise and help were indispensable for my research projects and this dissertation.

I thank our department for providing an excellent academic environment. In particular, I wish to thank Professor James Glimm, Dr. Yan Yu and all other Seawulf Cluster members for their support on my computations. I thank Christine Rota, Pam Wolfskill and Nancy Policastro for all of their help on adminstration issues. I am also very appreciated for the help and friendship from all the graduate students in our department.

I thank all of my colleagues in the Department of Preventive Medicine. They stimulated my passion for statistical research. I especially want to thank Professor M. Cristina Leske. Her dedication to public health research encouraged me very much.

I thank my wife for her sacrifice and her support on my academic pursuit.

Chapter 1

Introduction

1.1 Motivation

Finite mixture models are a class of statistical models that have been used in a wide range of applications. One of the most important application areas is statistical genetics, in which mixture distributions are used to model various qualitative and quantitative traits controlled by underlying genetic factors [36]. When it is assumed that we know the underlying genetic factors, variance components, logistic regression and other traditional statistical procedures can be used to study the effects of underlying genetic factors on some specific qualitative or quantitative traits. For example, with multiple regression models, Caspi *et al.* [5] found that a functional polymorphism in the gene encoding the neurotransmitter-metabolizing enzyme monoamine oxidase A (MAMO) moderated the antisocial behavior associated with childhood maltreatment. In reality, the process to obtain the underlying genetic mechanism is very costly and time consuming. A statistically oriented way to initiate this research is: first to perform a mixture analysis on a set of data to check whether the trait can be described by a mixture distribution. Then use segregation analysis to determine if there

may be a suitable genetic model that is consistent with the mixture distribution detected by the previous step. Finally, find the approximate location of the gene through linkage analysis [36]. In this process, obtaining a valid genetic model depends on the accurate estimation of mixture distribution parameters, especially, the mixture proportions. Therefore, a sound mixture analysis strategy can assist many genetic studies.

To illustrate some specific mixture analysis tasks, consider a selection of variables describing the attributes of Pima Indians Diabetes Database [30]. The purpose of the complete Pima data set is to study the risk factors for the incidence of diabetes in women of Pima Indian heritage. In this selected data set we only include the following attributes:

- Number of times pregnant;
- Plasma glucose concentration in an oral glucose tolerance test;
- Diastolic blood pressure ($mm\ Hg$);
- Triceps skin fold thickness (mm);
- 2-Hour serum insulin ($mu\ U/ml$);
- Body mass index ($weight\ in\ kg/(height\ in\ m)^2$);
- Age (years).

We have excluded the variable *diabetes pedigree function*. Our main interest in this data set is glucose concentration. From data screening we know that there is apparently some linear relationship between glucose concentration and serum insulin, and we want to know how serum insulin influences the glucose concentration. In order to study whether there is some genetic model that can be used to interpret this data set, we test the following hypotheses:

1. Do the data come from a two-component mixture, such that under each one there is a linear relationship among glucose concentration, serum insulin and other covariates?

2. Do the two linear regression lines have the same intercept? In other words, after controlling for other covariates, does glucose concentration have the same group mean at serum insulin baseline (value 0), but different increase rates proportional to serum insulin in the two components?

3. Do the two linear regression lines have the same slope for serum insulin? In other words, after controlling for other covariates, does glucose concentration have the same increase rates proportional to serum insulin, but with a different group mean at serum insulin baseline (value 0) in the two components?

To answer the first question, we can use the mixture linear regression models that were first studied by Quandt and Ramsey [34]. There are no reported methods to handle the second and third questions. Therefore, in this dissertation we carry out extensive studies to develop methods to handle parameter estimation, statistical inference, sample size, and power calculations for the second and third questions.

1.2 Restricted Mixture Linear Regression Models

In this dissertation, we concentrate on a subclass of the following *mixture linear regression models (MLRM)*:

$$y \sim \begin{cases} \boldsymbol{x}^T\boldsymbol{\beta}_1 + \varepsilon_j & \text{with probability } p_1, \\ \dots & \\ \boldsymbol{x}^T\boldsymbol{\beta}_g + \varepsilon_g & \text{with probability } p_g, \end{cases} \tag{1.1}$$

$$\varepsilon_j \overset{iid}{\sim} N(0, \sigma_j^2),$$

$$0 \leq p_j \leq 1, \qquad \sum_{j=1}^{g} p_j = 1,$$

where $\overset{iid}{\sim}$ is the notation for 'are independent and identically distributed as'. In our subclass model, some elements of $\boldsymbol{\beta}_j$ may have the same values across all mixture components. Because of this restriction on some $\boldsymbol{\beta}_j$ elements, we call these models *restricted mixture linear regression models (RMLRM)*. We consider in detail two specific models:

Mixture Intercept Model (MIM) : two model components with different intercepts and error terms but the same slope:

$$y \sim \begin{cases} \alpha + \boldsymbol{x}^T\boldsymbol{\beta} + \varepsilon_1 & \text{with probability } p, \\ (\alpha + \delta) + \boldsymbol{x}^T\boldsymbol{\beta} + \varepsilon_2 & \text{with probability } 1-p. \end{cases} \tag{1.2}$$

Mixture Slope Model (MSM) : two model components with different slopes and error terms but the same intercept:

$$y \sim \begin{cases} \alpha + \gamma x_t + \boldsymbol{x}_c^T\boldsymbol{\beta} + \varepsilon_1 & \text{with probability } p, \\ \alpha + (\gamma + \delta) x_t + \boldsymbol{x}_c^T\boldsymbol{\beta} + \varepsilon_2 & \text{with probability } 1-p. \end{cases} \tag{1.3}$$

In the mixture slope model, we assume x_t is the treatment variable whose effect we wish to study for and $\boldsymbol{x}_c$ are covariates that we need control for.

For clear presentation, in this dissertation, we arrange the design matrix so that the differences in mixture parameters $\boldsymbol{\beta}$s only exist in the first d elements out of total m elements, and let

$$\boldsymbol{\beta}_j = \boldsymbol{\beta}_1 + \begin{bmatrix} \boldsymbol{\delta}_j \\ \mathbf{0} \end{bmatrix} \qquad (j = 2, \ldots, g). \tag{1.4}$$

Chapter 2

Mathematic Background and Literature Review

In this chapter, I present the mathematical background and literature review of this dissertation.

2.1 Modes of Convergence

There are many ways for a sequence of random variables, X_1, X_2, ..., to converge to another random variable X. In this dissertation, we use the following two modes of convergence(modified from [9]):

***Convergence in probability*:** X_n converges in probability to X, $X_n \xrightarrow{P} X$, if for any $\varepsilon > 0$

$$\Pr(\mid X_n - X \mid > \varepsilon) \to 0 \qquad as \qquad n \to \infty;$$

***Convergence in distribution*:** X_n converges in distribution to X, $X_n \xrightarrow{D} X$, if

$$\Pr(X_n \leq x) \to \Pr(X \leq x) \qquad as \qquad n \to \infty$$

at every x for which the distribution function $\Pr(X \leq x)$ is continuous.

2.2 Maximum Likelihood Estimation

Suppose $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is a data vector of n independent and identically-distributed (IID) random variables with *probability density function (PDF)* $f(y; \boldsymbol{\theta})$ ($\boldsymbol{\theta}$ is a m dimension parameter vector). Then the joint density function is

$$f(\boldsymbol{y}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta}). \tag{2.1}$$

If we switch the roles of $\boldsymbol{y}$ and $\boldsymbol{\theta}$ and take $\boldsymbol{y}$ as fixed, then the joint probability density is a function of an unknown parameter vector $\boldsymbol{\theta}$ and is defined to be the *likelihood function*

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = f(\boldsymbol{y}; \boldsymbol{\theta}). \tag{2.2}$$

The *log likelihood function* is defined to be

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} \log f(y_i; \boldsymbol{\theta}). \tag{2.3}$$

The *maximum likelihood estimate* (MLE) of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, is a vector of $\boldsymbol{\theta}$ values that maximizes the likelihood (2.2) (equivalently the log likelihood (2.3)). Typically, the MLE is a solution of the *likelihood equation*:

$$\frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta}} = \mathbf{0}, \tag{2.4}$$

where $\boldsymbol{u}(\boldsymbol{\theta};\boldsymbol{y}) = \partial\ell(\boldsymbol{\theta};\boldsymbol{y})/\partial\boldsymbol{\theta}$ is called the *score function*. For the MLE, we have following optimality theorem (modified from [9]).

Theorem 2.2.1. *Suppose the probability density function $f(y;\boldsymbol{\theta})$ satisfies the* regularity conditions*:*

1. *the true value $\boldsymbol{\theta}^0$ of $\boldsymbol{\theta}$ is interior to the parameter space $\boldsymbol{\Theta}$, which has finite dimension and is compact;*

2. *the densities defined by any two different values of $\boldsymbol{\theta}$ are different;*

3. *there is a neighborhood $\mathcal{N}$ of $\boldsymbol{\theta}^0$ within which the first three derivatives of the log likelihood with respect to $\boldsymbol{\theta}$ exist almost surely, and for $r,s,t = 1,\ldots,m$, $n^{-1}\mathbf{E}|\partial^3\ell(\boldsymbol{\theta};y)/\partial\theta_r\partial\theta_s\partial\theta_t|$ is uniformly bounded for $\boldsymbol{\theta}\in\mathcal{N}$; and*

4. *within $\mathcal{N}$, the* Fisher information matrix *$I(\boldsymbol{\theta})$ is finite and positive definite, and its elements satisfy*

$$i(\boldsymbol{\theta})_{rs} = \mathbf{E}\left\{\frac{\partial\ell(\boldsymbol{\theta};y)}{\partial\theta_r}\frac{\partial\ell(\boldsymbol{\theta};y)}{\partial\theta_s}\right\} = \mathbf{E}\left\{-\frac{\partial^2\ell(\boldsymbol{\theta};y)}{\partial\theta_r\partial\theta_s}\right\}, \quad r,s = 1,\ldots,m. \tag{2.5}$$

Then:

1. *$\hat{\boldsymbol{\theta}}$ is a* consistent estimator *of $\boldsymbol{\theta}^0$, that is, for every $\epsilon > 0$ and every $\boldsymbol{\theta}^0$,*

$$lim_{n\to\infty}P_{\boldsymbol{\theta}^0}(|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0| \geq \epsilon) = 0. \tag{2.6}$$

2. *$\hat{\boldsymbol{\theta}}$ is* asymptotically efficient*, which means*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \xrightarrow{D} N(0, I^{-1}(\boldsymbol{\theta}^0)). \tag{2.7}$$

and $I(\boldsymbol{\theta}^0)$ achieves the Cramér-Rao Bound*.* ■

The definition of Fisher information matrix (2.5) specifies that it is the covariance

matrix of the score function. The *observed information matrix* is defined to be

$$J(\hat{\boldsymbol{\theta}}) = -\frac{\partial^2 \ell(\hat{\boldsymbol{\theta}}; \boldsymbol{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \tag{2.8}$$

and we have the approximation

$$\hat{\boldsymbol{\theta}} \dot{\sim} N(\boldsymbol{\theta}^0, J^{-1}(\hat{\boldsymbol{\theta}})), \tag{2.9}$$

where $\dot{\sim}$ is the notation for 'is approximately distributed as'. Equation (2.9) is used to construct confidence regions for $\boldsymbol{\theta}$ and to carry out *Wald tests* for components of $\boldsymbol{\theta}^0$.

2.3 Likelihood Ratio Test Statistic (LRTS)

Suppose we want to test the following hypothesis:

$$\mathbf{H}_0 : \theta_1 = \theta_1^0, \ldots, \theta_d = \theta_d^0 \tag{2.10}$$

against the alternative hypothesis $\mathbf{H}_1$ in which $\theta_1, \ldots, \theta_d$ are any possible values. Then the *likelihood ratio test statistic (LRTS)* for this test is defined as

$$\lambda_n(\boldsymbol{\theta}^0; \boldsymbol{y}) = 2\{\ell(\hat{\boldsymbol{\theta}}; \boldsymbol{y}) - \ell(\boldsymbol{\theta}^0; \boldsymbol{y})\}. \tag{2.11}$$

For the LRTS, we have the following theorem (modified from [45])

Theorem 2.3.1. *Suppose the model satisfies the same regularity conditions listed in theorem 2.2.1, and* $\mathbf{H}_0$ *is true. Then as the sample size* $n \to \infty$*,*

$$\lambda_n \xrightarrow{D} \chi^2_d. \tag{2.12}$$

2.4 Expectation Maximization (EM) Algorithm

The *Expectation Maximization (EM)* Algorithm [10] is a popular algorithm to solve maximum likelihood estimation equations for various mixture models. Suppose there are n observations $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ from a mixture distribution with g components. In order to use EM algorithm, we introduce a zero-one membership vector of length g for every observation y_i to indicate which component the observation y_i comes from. That is, $z_{ij} = 1(z_{il} = 0, j \neq l)$ means y_i comes from component j, and $\boldsymbol{y}$ is called the *observed-data* or *incomplete-data* and $\{\boldsymbol{y}, Z\}$ is called the *complete-data*. The *incomplete-data likelihood function* and *incomplete-data log likelihood function* are

$$L(\boldsymbol{\psi}; \boldsymbol{y}) = \prod_{i=1}^{n} \sum_{j=1}^{g} p_j f_j(y_i; \boldsymbol{\theta}_j) \tag{2.13}$$

$$\ell(\boldsymbol{\psi}; \boldsymbol{y}) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{g} p_j f_j(y_i; \boldsymbol{\theta}_j) \right); \tag{2.14}$$

and the *complete-data likelihood function* and *complete-data log likelihood function* are

$$L_c(\boldsymbol{\psi}; \boldsymbol{y}, Z) = \prod_{i=1}^{n} \prod_{j=1}^{g} (p_j f_j(y_i; \boldsymbol{\theta}_j))^{z_{ij}} \tag{2.15}$$

$$\ell_c(\boldsymbol{\psi}; \boldsymbol{y}, Z) = \sum_{i=1}^{n} \sum_{j=1}^{g} z_{ij} \left(\log p_j + \log f_j(y_i; \boldsymbol{\theta}_j) \right), \tag{2.16}$$

where $\boldsymbol{\psi}$ is the nonredundant parameter vector that defines all p and $\boldsymbol{\theta}$.

2.4.1 EM Procedure [26]

In order to handle the unknown Z, the EM algorithm iteratively goes through the following two steps until there is apparent convergence of estimates.

E-Step

Given observed data $\boldsymbol{y}$ and current estimate $\boldsymbol{\psi}^{(r)}$, calculate the *conditional expectation of the complete-data log likelihood*

$$\begin{aligned}\mathbf{E}[\ell_c(\boldsymbol{\psi};\boldsymbol{y},Z)|\boldsymbol{y},\boldsymbol{\psi}^{(r)}] &= \mathbf{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}\left(\log p_j + \log f_j(y_i;\boldsymbol{\theta}_j)\right)\Big|\boldsymbol{y},\boldsymbol{\psi}^{(r)}\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{g}\left(\log p_j + \log f_j(y_i;\boldsymbol{\theta}_j)\right)\mathbf{E}[z_{ij}|\boldsymbol{y},\boldsymbol{\psi}^{(r)}] \qquad (2.17)\end{aligned}$$

Since z_{ij} is a $0-1$ membership function,

$$\begin{aligned}\mathbf{E}[z_{ij}|\boldsymbol{y},\boldsymbol{\psi}^{(r)}] &= \Pr\{z_{ij}=1|\boldsymbol{y},\boldsymbol{\psi}^{(r)}\} \\ &= \frac{p_j^{(r)} f_j(y_i;\boldsymbol{\theta}_j^{(r)})}{\sum_{l=1}^{g} p_l^{(r)} f_l(y_i;\boldsymbol{\theta}_l^{(r)})}. \qquad (2.18)\end{aligned}$$

Let $\tau_j(y_i;\boldsymbol{\psi}^{(r)}) = \mathbf{E}[z_{ij}|\boldsymbol{y},\boldsymbol{\psi}^{(r)}]$, which is the posterior probability that y_i comes from the jth component of the mixture (conditioned on current parameter estimation $\boldsymbol{\psi}^{(r)}$). Substitute $\tau_j(y_i;\boldsymbol{\psi}^{(r)})$ into (2.17) and define $\mathbf{E}[\ell_c(\boldsymbol{\psi};\boldsymbol{y},Z)|\boldsymbol{y},\boldsymbol{\psi}^{(r)}]$ to be $\boldsymbol{Q}(\boldsymbol{\psi};\boldsymbol{\psi}^{(r)})$. Then we have

$$\boldsymbol{Q}(\boldsymbol{\psi};\boldsymbol{\psi}^{(r)}) = \sum_{i=1}^{n}\sum_{j=1}^{g}\tau_j(y_i;\boldsymbol{\psi}^{(r)})\left(\log p_j + \log f_j(y_i;\boldsymbol{\theta}_j)\right) \qquad (2.19)$$

M-Step

Given the conditional expectation of the complete-data log likelihood, $\boldsymbol{Q}(\boldsymbol{\psi};\boldsymbol{\psi}^{(r)})$, the purpose of the M-step is to maximize $\boldsymbol{Q}(\boldsymbol{\psi};\boldsymbol{\psi}^{(r)})$ respect to $\boldsymbol{\psi}$.

The likelihood equation with respect to the mixing proportion vector $\boldsymbol{p} = (p_1,\ldots,p_g)^T$

is

$$\frac{\partial \boldsymbol{Q}(\boldsymbol{\psi}; \boldsymbol{\psi}^{(r)})}{\partial \boldsymbol{p}} = \frac{\partial \left(\sum_{i=1}^{n} \sum_{j=1}^{g} \tau_j(y_i; \boldsymbol{\psi}^{(r)}) \log p_j \right)}{\partial \boldsymbol{p}} = \mathbf{0} \tag{2.20}$$

The solution of (2.20) is

$$p_j^{(r+1)} = \sum_{i=1}^{n} \tau_j(y_i; \boldsymbol{\psi}^{(r)})/n \qquad (j = 1, \ldots, g) \tag{2.21}$$

The likelihood equation with respect to $\boldsymbol{\theta}_j$ is

$$\frac{\partial \boldsymbol{Q}(\boldsymbol{\psi}; \boldsymbol{\psi}^{(r)})}{\partial \boldsymbol{\theta}_j} = \frac{\partial \left(\sum_{i=1}^{n} \tau_j(y_i; \boldsymbol{\psi}^{(r)}) \log f_j(y_i; \boldsymbol{\theta}_j) \right)}{\partial \boldsymbol{\theta}_j} = \mathbf{0} \tag{2.22}$$

which should be solved using the specific component PDF to obtain updates of $\boldsymbol{\theta}^{(r+1)}$.

2.4.2 Empirical Information Matrix

Since the information matrix is the covariance of the score function, the *empirical information matrix* (see [27]) is

$$J_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\psi}; \boldsymbol{y}) = \sum_{i=1}^{n} \frac{\partial \ell(\boldsymbol{\psi}; y_i)}{\partial \boldsymbol{\psi}} \left[\frac{\partial \ell(\boldsymbol{\psi}; y_i)}{\partial \boldsymbol{\psi}} \right]^T - \frac{1}{n} \left\{ \sum_{i=1}^{n} \frac{\partial \ell(\boldsymbol{\psi}; y_i)}{\partial \boldsymbol{\psi}} \right\} \left\{ \sum_{i=1}^{n} \frac{\partial \ell(\boldsymbol{\psi}; y_i)}{\partial \boldsymbol{\psi}} \right\}^T \tag{2.23}$$

Suppose the EM procedure converges to a local maximum $\hat{\boldsymbol{\psi}}$. Then

$$\left\{ \sum_{i=1}^{n} \frac{\partial \ell(\boldsymbol{\psi}; y_i)}{\partial \boldsymbol{\psi}} \right\} \Bigg|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} = \mathbf{0}. \tag{2.24}$$

Under very general regularity conditions that allow interchanging derivative and integral [21],

$$\left\{ \frac{\partial \ell(\boldsymbol{\psi}; \boldsymbol{y})}{\partial \boldsymbol{\psi}} \right\} \Bigg|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} = \left\{ \mathbf{E} \left[\frac{\partial \ell_c(\boldsymbol{\psi}; \boldsymbol{y}, Z)}{\partial \boldsymbol{\psi}} \Big| \boldsymbol{y}, \boldsymbol{\psi} \right] \right\} \Bigg|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}. \tag{2.25}$$

This means that the score function of the incomplete-data log likelihood equals the expectation of the score function of the complete-data log likelihood function. Combining (2.23), (2.24) and (2.25),

$$\begin{aligned} J_{\boldsymbol{\psi}\boldsymbol{\psi}}(\hat{\boldsymbol{\psi}}; \boldsymbol{y}) &= \left\{ \sum_{i=1}^{n} \left\{ \frac{\partial \ell(\boldsymbol{\psi}; \boldsymbol{y})}{\partial \boldsymbol{\psi}} \right\} \left\{ \frac{\partial \ell(\boldsymbol{\psi}; \boldsymbol{y})}{\partial \boldsymbol{\psi}} \right\}^T \right\} \Bigg|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} \\ &= \left\{ \sum_{i=1}^{n} \mathbf{E} \left[\frac{\partial \ell_c(\boldsymbol{\psi}; y_i, \boldsymbol{z}_i)}{\partial \boldsymbol{\psi}} \Big| \boldsymbol{y}, \boldsymbol{\psi} \right] \mathbf{E} \left[\frac{\partial \ell_c(\boldsymbol{\psi}; y_i, \boldsymbol{z}_i)}{\partial \boldsymbol{\psi}} \Big| \boldsymbol{y}, \boldsymbol{\psi} \right]^T \right\} \Bigg|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}. \end{aligned} \tag{2.26}$$

2.4.3 EM Convergence

As to the convergence of the EM algorithm, the following theorem (from [4]) guarantees that every EM step increases the incomplete-data likelihood. Therefore, the EM procedure will eventually reach a local maximum.

Theorem 2.4.1 (**Monotonic EM sequence**). *The sequence* $\boldsymbol{\psi}^{(r)}$ *defined by the EM Procedure satisfies*

$$L\left(\boldsymbol{\psi}^{(r+1)}; \boldsymbol{y}\right) \geq L\left(\boldsymbol{\psi}^{(r)}; \boldsymbol{y}\right) \tag{2.27}$$

with equality holding if and only if successive iterations yield the same value of the maximized expected complete-data log likelihood; that is,

$$\mathbf{E}[\ell_c(\boldsymbol{\psi}^{(r+1)}; \boldsymbol{y}, Z) | \boldsymbol{y}, \boldsymbol{\psi}^{(r)}] = \mathbf{E}[\ell_c(\boldsymbol{\psi}^{(r)}; \boldsymbol{y}, Z) | \boldsymbol{y}, \boldsymbol{\psi}^{(r)}] \tag{2.28}$$

■

2.4.4 Implementation

In order to maximize the log likelihood function, the EM algorithm introduces n vectors of missing value indicators Z. For every observation, there are at least two missing indicators. That is, since we work on a parameter space with much a higher dimension than the original one, one has a chance of converging to a local maximum. The monotonic increase property of the EM procedure can require a large number of iterations until convergence. Given that, there are two key implementation issues in mixture modeling: how to make sure we find a global maximum and how to find a good initialization scheme to set the starting points.

The best way to handle the global maximum issue is to run the EM procedure multiple times with different starting points. For instance, in calculating the MLEs assuming a two-component normal mixture when the data were in fact sampled from an $N(0,1)$ distribution, Finch *et al.* [13] proposed the following procedure for a two component normal mixture.

Procedure 2.4.2.

Step 1: *Obtain multiple random starting values p_0 for mixing proportion p by sampling from a uniform $U(0,1)$ distribution;*

Step 2: *For every p_0, calculate the starting values for the remaining three parameters according to:*

$$\mu_{10} = \sum_{i=1}^{m} x_{(i)}/m$$

$$\mu_{20} = \sum_{i=m+1}^{n} x_{(i)}/(n-m)$$

$$\sigma_0^2 = \left[\sum_{i=1}^{m}(x_{(i)} - \mu_{10})^2 + \sum_{i=m+1}^{n}(x_{(i)} - \mu_{20})^2\right] \Big/ (n-2)$$

where $x_{(i)}$ denotes the ith sample order statistic and m is the integer part of np_0;

Step 3: *Run a numeric optimization algorithm on each starting point;*

Step 4: *Pick the best solution and estimate the probability that it is the real global maximum from the convergence pattern of all solutions.* ■

As to the initialization issue, we need problem specific strategies. For example, the step 2 in the previous procedure and *k-means clustering* [24] in multivariate cases are effective choices for the mixture normal distribution.

2.5 Introduction to Mixture Models and Literature Review

Mixture modeling is a flexible strategy to handle population heterogeneity. Let the random variable y have the *probability density function (PDF)*

$$f(y;\boldsymbol{\psi}) = \sum_{j=1}^{g} p_j f_j(y;\boldsymbol{\theta}_j), \tag{2.29}$$

$$0 \leq p_j \leq 1 \qquad (j = 1, \ldots, g),$$

$$\sum_{j=1}^{g} p_j = 1.$$

Each $f_j(y;\boldsymbol{\theta}_j)$ is the PDF for a homogeneous subpopulation with parameter vector $\boldsymbol{\theta}_j$, and $\boldsymbol{\psi}$ is the nonredundant parameter vector that defines all p and $\boldsymbol{\theta}$.

In this mixture model, there are two levels of randomness:

Membership Level : The component membership of each observation follows a g-class multinomial distribution with probabilities $(p_1, \ldots, p_g)$;

Component Level : Conditional on the component membership, each observation has a specific PDF $f_j(y; \boldsymbol{\theta}_j)$.

Pearson [33] and Cohen [8] estimated the parameters for the mixtures of two normal distributions with the *method of moments*. Besides the inefficiency problem, another drawback of the method of moments is that we need to solve a series of algebra equations obtained by equating sample moments to corresponding population (theoretic) moments. These algebraic equations will change dramatically with the number of mixing components and other assumptions. Rao [35] used *maximum likelihood estimation (MLE)* to solve the same problem but with a restriction of equal variance for the two components. Assuming equal variance, Tan and Chang [38] obtained *asymptotic covariance matrices* for the point estimations through the method of moments and MLE respectively, and concluded that maximum likelihood is a much more efficient method than the method of moments. In their landmark paper [10] on the *expectation-maximization (EM) algorithm*, Dempster *et al.* provided the EM algorithm for finding MLEs for any mixture model described by (2.29), and pointed out that Ceppellini *et al.* [6] had already used EM algorithm in mixture modeling for genetic data. Everitt [11] compared six algorithms for finding the MLE using simulations on three data sets, and concluded that the best two methods were Newton's method using exact values of the gradient and Hessian matrices and the EM algorithm.

The preceding papers on mixture model estimation always assumed that the number of mixing components is known. In reality, determining the number of mixing components through hypothesis testing is an extremely difficult problem. Ghosh and Sen [15] pointed out the violation of the regularity conditions for the classical asymptotic theory of the likelihood ratio test statistic (LRTS). For the case of testing homogeneity against two-component normal mixture with equal and known variance, Ghosh and Sen [15] also showed that, under certain separation and boundness conditions on two normal means, the null LRTS is

asymptotically distributed as a function of a complicated Gaussian process. For the same problem, Hartigan [17] conjectured that, without the separation and bounded conditions, the null LRTS is asymptotically diverges to infinite at the rate $\frac{1}{2}\log(\log(n))$. Liu and Shao [18] proved that, for the simple homogeneity test

$$\begin{aligned} \mathbf{H_0} &: N(0,1) \\ \mathbf{H_A} &: pN(t,1) + (1-p)N(0,1) \\ & \quad p \in (0,1], t \in \Re \backslash \{0\}, \end{aligned} \tag{2.30}$$

the null LRTS λ_n follows

$$\lim_{n\to\infty} \mathbf{P}\{\lambda_n - \log(\log(n)) + \log(2\pi^2) \leq x\} = e^{-e^{-x/2}}, \quad x > 0. \tag{2.31}$$

The divergence of the asymptotic distribution (2.31) is at a very slow rate which could not be detected, even with sample size of 5000 in authors' simulation. For seven different cases of two-component normal mixture models, Garel [14] provided the null LRTSs which also were represented by the functions of a Gaussian process. Garel's simulation results are different from others. For example, his power to detect the normal mixture with Mahalanobis distance ($|\mu_1 - \mu_2|/\sigma$) of 1.5, mixing proportion of 0.7 and sample size of 100 is 99.8%. Under almost the same conditions (except the Mahalanobis distance is 2.0), Mendell *et al.* [29] report power of 26.0%. Based on Kullback-Leibler information, Lo *et al.* [20] proposed the following asymptotic distribution of the null LRTS λ_n

$$\lim_{n\to\infty} \mathbf{P}\{\lambda_n \leq x\} = M_{m_1+m_2}(x;\boldsymbol{\iota}), \quad x > 0, \tag{2.32}$$

to test the number of components in a normal mixture. In (2.32), m_1 and m_2 are the numbers of parameter for the two hypothetical normal mixture models, $M_{m_1+m_2}(\cdot)$ is the

weighted sum of χ^2_1 distributions as defined by Vuong [42], and $\boldsymbol{\iota}$ is the vector of $m_1 + m_2$ eigenvalues of a complicated matrix made by eight second order derivative matrices of the two hypothetic normal mixture PDFs. Lo [19] also further extended this result to unequal variance cases. Hall and Stewart [16] reported a theoretical power study based on the Liu and Shao null LRTS result (2.31) for the hypothesis test problem of (2.30). They did not provide any simulation results.

Besides these theoretic results on the null LRTS distribution and power calculations for detecting mixtures, there are many papers using simulation and bootstrap techniques. McLachlan [25] used a parametric bootstrap procedure to assess the null distribution of LRTS for detecting homogeneity versus a mixture of two normal densities. Feng and McCulloch [12] provided some justification for the parametric bootstrapped LRTS by pointing out that the MLE is consistent with the set identifying the true density function. Mendell *et al.* reported simulation studies on the null LRTS distribution [40], sample size and power calculations [29], [28] for detecting two-component normal mixture models. Maclean *et al.* [23] suggested using a Box-Cox transformation to remove skewness from the data and extended the existing normal mixture approaches to other mixture problems. Ning and Finch [31], [32] obtained the null LRTS distribution, sample size and power results through simulations for those Box-Cox transformed mixture analyses.

There are some studies on mixture regression models. Quandt and Ramsey [34] referred to the mixture linear regression model (MLRM) as a switching regression model, and used *the moment generating function (MGF)* to estimate parameters. They also proved that the MGF solution had the properties of *consistency* and *asymptotic normality*. In one of their case studies, Quandt and Ramsey used a mixture linear regression model with an identical coefficient (out of total three coefficients) across two mixing components. Turner [41] obtained the MLE and its standard error formula for MLRM by using the EM algorithm. He

used this approach to estimate the propagation rate of a viral infection in potato plants. Wang *et al.* [44] used the following mixture Poisson regression model

$$\begin{aligned} f(y|\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{p}) &= \sum_{j=1}^{g} p_j f_j(y|\mu_j) \\ f_j(y|\mu_j) &= \frac{1}{y}\mu_j^y e^{-\mu_j} \\ \mu_j &= c e^{\boldsymbol{\alpha}_j^T \boldsymbol{x}} \end{aligned} \tag{2.33}$$

to handle overdispersion and covariate-dependent event rates. In this paper, Wang *et al.* also discussed some specific identical coefficient constraints in their model, but did not provide any further information. In order to handle the sources of extra-binomial variation, Wang and Puterman [43] proposed a class of mixture logistic regression models that allow both the mixing components and mixing proportions to be dependent on covariates.

Chapter 3

EM Algorithm for Restricted Mixture Linear Regression Models

In this chapter, the EM algorithm is used to calculate the MLE of the mixing proportion $\boldsymbol{p}$, regression parameters $\left(\boldsymbol{\beta}_1^T, \boldsymbol{\delta}_2^T, \ldots, \boldsymbol{\delta}_g^T\right)^T$, variance components $(\sigma_1^2, \ldots, \sigma_g^2)$ and their standard errors for the following restricted mixture linear regression models (RMLRM)

$$y = \begin{cases} \boldsymbol{x}^T\boldsymbol{\beta}_1 + \varepsilon_1 & \text{with probability } p_1, \\ \ldots & \\ \boldsymbol{x}^T\boldsymbol{\beta}_g + \varepsilon_g & \text{with probability } p_g, \end{cases} \tag{3.1}$$

$$\varepsilon_j \overset{iid}{\sim} N(0, \sigma_j^2),$$

$$0 \le p_j \le 1, \qquad \sum_{j=1}^{g} p_j = 1,$$

$$\boldsymbol{\beta}_j = \boldsymbol{\beta}_1 + \begin{bmatrix} \boldsymbol{\delta}_j \\ \mathbf{0} \end{bmatrix}, \qquad (j = 2, \ldots, g).$$

Specifically, when we calculate the MLE for all parameters, we assume the EM algorithm has already been iterated r times, and the results are the one-step update of the MLE for the $(r+1)$st iteration. In fact, the standard errors of the parameters can only be calculated at the convergence point of the EM algorithm. When we calculate the standard errors, we in fact assume the EM algorithm has already converged, and the final iteration number is c.

Because of the mathematical nature of this chapter, it might be better for the reader first to review Section 2.4.1 on the EM algorithm and preview Procedure 3.3.1 to get a general idea of what is going to be done in this chapter, and then to read it through section by section.

According to (2.16), the complete-data log likelihood function for mixture regression is

$$\ell_c(\boldsymbol{\psi}; \boldsymbol{y}, Z) = \sum_{i=1}^{n} \sum_{j=1}^{g} z_{ij} \big(\log p_j + \log f_j(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}_j) \big) \tag{3.2}$$

where

$$f_j(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}_j) = \frac{1}{\sigma_j} \phi(\frac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j}{\sigma_j}), \tag{3.3}$$

with ϕ the density of the standard Gaussian distribution, and $\boldsymbol{\theta}_j = \big(\boldsymbol{\beta}_j^T, \sigma_j\big)^T$.

Suppose we have the current estimate $\boldsymbol{\psi}^{(r)}$. The E-step of the $(r+1)$th iteration for the RMLRM follows the standard procedure listed in 2.4.1. Namely,

$$\boldsymbol{Q}(\boldsymbol{\psi}; \boldsymbol{\psi}^{(r)}) = \sum_{i=1}^{n} \sum_{j=1}^{g} \tau_j(y_i; \boldsymbol{\psi}^{(r)}) \big(\log p_j + \log f_j(y_i; \boldsymbol{\theta}_j) \big), \tag{3.4}$$

and

$$\tau_j(y_i; \boldsymbol{\psi}^{(r)}) = \frac{p_j^{(r)} f_j(y_i; \boldsymbol{\theta}_j^{(r)})}{\sum_{l=1}^{g} p_l^{(r)} f_l(y_i; \boldsymbol{\theta}_l^{(r)})}. \tag{3.5}$$

Equations (3.4) and (3.5) are straightforward, and even apply with mixture components

$f_j(y; \boldsymbol{\theta}_j)$ have different PDFs.

In the rest of this chapter, I carry out

- the M-step calculations to obtain the one-step update for the parameter estimation;
- the calculations for the standard error for every parameter at a convergence point of the EM algorithm.

The calculations are separated into equal and unequal variance cases respectively.

3.1 Equal Variance Case

In this section, we assume all the mixture components have the same error term variances ($\sigma_1^2 = \cdots = \sigma_g^2$).

3.1.1 Mixing Proportion ($\boldsymbol{p}$) Estimation

The update for the mixing proportion $\boldsymbol{p}$ follows the standard formula (2.21):

$$p_j^{(r+1)} = \sum_{i=1}^{n} \tau_j(y_i; \boldsymbol{\psi}^{(r)})/n \qquad (j = 1, \ldots, g).$$

3.1.2 Regression Parameter ($\boldsymbol{\beta}$) Estimation

Let $\boldsymbol{\xi}$ be the parameter vector that defines all $\boldsymbol{\beta}_j$. Then according to section 2.4.1, the likelihood equation for $\boldsymbol{\xi}$ is

$$\frac{\partial \boldsymbol{Q}(\boldsymbol{\psi}; \boldsymbol{\psi}^{(r)})}{\partial \boldsymbol{\xi}} = \frac{\partial \sum_{i=1}^{n} \sum_{j=1}^{g} \tau_j(y_i; \boldsymbol{\psi}^{(r)}) \log f_j(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}_j)}{\partial \boldsymbol{\xi}} = \mathbf{0} \tag{3.6}$$

in which

$$\log f_j(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}_j) = -\frac{\log(2\pi\sigma^2)}{2} - \frac{(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j)^2}{2\sigma^2}. \tag{3.7}$$

From (3.6),

$$\frac{\partial \sum_{i=1}^{n} \sum_{j=1}^{g} \tau_j(y_i; \boldsymbol{\psi}^{(r)})(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j)^2/(2\sigma^2)}{\partial \boldsymbol{\xi}} = \mathbf{0}. \tag{3.8}$$

Define $\boldsymbol{Y}$, $\boldsymbol{X}$ and $n \times n$ diagonal matrix $\boldsymbol{W}_j^{(r)}$ as

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix},$$

$$\boldsymbol{W}_j^{(r)} = \begin{pmatrix} \frac{\tau_j(y_1; \boldsymbol{\psi}^{(r)})}{2\sigma^2} & 0 & \dots & 0 \\ 0 & \frac{\tau_j(y_2; \boldsymbol{\psi}^{(r)})}{2\sigma^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{\tau_j(y_n; \boldsymbol{\psi}^{(r)})}{2\sigma^2} \end{pmatrix}. \tag{3.9}$$

Equation (3.8) can be written as

$$\frac{\partial\left[\sum_{j=1}^{g}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_j)^T\boldsymbol{W}_j^{(r)}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_j)\right]}{\partial\boldsymbol{\xi}}=\mathbf{0}. \tag{3.10}$$

Suppose there are differences in mixture parameters $\boldsymbol{\beta}_j$ only in the first d of a total m elements, and assume

$$\boldsymbol{\beta}_j=\boldsymbol{\beta}_1+\begin{bmatrix}\boldsymbol{\delta}_j\\ \mathbf{0}\end{bmatrix}, \qquad (j=2,\ldots,g). \tag{3.11}$$

Define $\boldsymbol{X}_d$ as

$$\boldsymbol{X}_d=\boldsymbol{X}\begin{pmatrix}\boldsymbol{I}_{d\times d}\\ \mathbf{0}_{(m-d)\times d}\end{pmatrix}_{m\times d}, \tag{3.12}$$

and write $\boldsymbol{\xi}$ as

$$\boldsymbol{\xi}=\left(\boldsymbol{\beta}_1^T,\boldsymbol{\delta}_2^T,\ldots,\boldsymbol{\delta}_g^T\right)^T. \tag{3.13}$$

Then from (3.10) we have following likelihood equations

$$\frac{\partial\left[(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_1)^T\boldsymbol{W}_1^{(r)}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_1)+\sum_{j=2}^{g}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_j-\boldsymbol{X}_d\boldsymbol{\delta}_j)^T\boldsymbol{W}_j^{(r)}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_j-\boldsymbol{X}_d\boldsymbol{\delta}_j)\right]}{\partial\boldsymbol{\xi}}=\mathbf{0}$$

$$\Rightarrow\quad \frac{\partial\left[(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_1)^T\left(\sum_{j=1}^{g}\boldsymbol{W}_j^{(r)}\right)(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_1)+\sum_{j=2}^{g}\boldsymbol{\delta}_j^T\boldsymbol{X}_d^T\boldsymbol{W}_j^{(r)}\boldsymbol{X}_d\boldsymbol{\delta}_j-2\sum_{j=2}^{g}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_1)^T\boldsymbol{W}_j^{(r)}\boldsymbol{X}_d\boldsymbol{\delta}_j\right]}{\partial\boldsymbol{\xi}}=\mathbf{0}$$

$$\Rightarrow\quad \frac{\partial\left[\left(\boldsymbol{\beta}_1^T\ \boldsymbol{\delta}_2^T\ \ldots\ \boldsymbol{\delta}_g^T\right)\boldsymbol{A}(r)\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\delta}_2\\ \vdots\\ \boldsymbol{\delta}_g\end{pmatrix}-2\left(\boldsymbol{\beta}_1^T\ \boldsymbol{\delta}_2^T\ \ldots\ \boldsymbol{\delta}_g^T\right)\boldsymbol{B}(r)\right]}{\partial\boldsymbol{\xi}}=\mathbf{0}$$

$$\Rightarrow\quad \frac{\partial\left(\boldsymbol{\xi}^T\boldsymbol{A}(r)\boldsymbol{\xi}-2\boldsymbol{\xi}^T\boldsymbol{B}(r)\right)}{\partial\boldsymbol{\xi}}=\mathbf{0} \tag{3.14}$$

where

$$\boldsymbol{A}(r) = \begin{pmatrix} \boldsymbol{X}^T(\sum_{j=1}^{g} \boldsymbol{W}_j^{(r)})\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{W}_2^{(r)}\boldsymbol{X}_d & \dots & \boldsymbol{X}^T\boldsymbol{W}_g^{(r)}\boldsymbol{X}_d \\ \boldsymbol{X}_d^T\boldsymbol{W}_2^{(r)}\boldsymbol{X} & \boldsymbol{X}_d^T\boldsymbol{W}_2^{(r)}\boldsymbol{X}_d & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{X}_d^T\boldsymbol{W}_g^{(r)}\boldsymbol{X} & 0 & \dots & \boldsymbol{X}_d^T\boldsymbol{W}_g^{(r)}\boldsymbol{X}_d \end{pmatrix},$$

$$\boldsymbol{B}(r) = \begin{pmatrix} \boldsymbol{X}^T(\sum_{j=1}^{g} \boldsymbol{W}_j^{(r)})\boldsymbol{Y} \\ \boldsymbol{X}_d^T\boldsymbol{W}_2^{(r)}\boldsymbol{Y} \\ \vdots \\ \boldsymbol{X}_d^T\boldsymbol{W}_g^{(r)}\boldsymbol{Y} \end{pmatrix}.$$

Therefore, we have the update for ξ as

$$\begin{pmatrix} \boldsymbol{\beta}_1^{(r+1)} \\ \boldsymbol{\delta}_2^{(r+1)} \\ \vdots \\ \boldsymbol{\delta}_g^{(r+1)} \end{pmatrix} = \boldsymbol{A}^{-1}(r)\boldsymbol{B}(r). \tag{3.15}$$

3.1.3 Variance (σ^2) Estimation

The likelihood equation for σ^2 is

$$\begin{aligned}
& \frac{\partial \boldsymbol{Q}(\boldsymbol{\psi}; \boldsymbol{\psi}^{(r)})}{\partial \sigma^2} = \mathbf{0} \\
\Rightarrow \quad & \frac{\partial \sum_{i=1}^{n} \sum_{j=1}^{g} \tau_j(y_i; \boldsymbol{\psi}^{(r)}) \log f_j(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}_j)}{\partial \sigma^2} = \mathbf{0} \\
\Rightarrow \quad & \sum_{i=1}^{n} \sum_{j=1}^{g} \tau_j(y_i; \boldsymbol{\psi}^{(r)}) \frac{\partial \log f_j(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}_j)}{\partial \sigma^2} = \mathbf{0} \\
\Rightarrow \quad & \sum_{i=1}^{n} \sum_{j=1}^{g} \tau_j(y_i; \boldsymbol{\psi}^{(r)}) \left[-\frac{1}{\sigma^2} + \frac{(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j)^2}{\sigma^4} \right] = \mathbf{0} \\
\Rightarrow \quad & \sum_{i=1}^{n} \sum_{j=1}^{g} \tau_j(y_i; \boldsymbol{\psi}^{(r)}) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{g} \tau_j(y_i; \boldsymbol{\psi}^{(r)})(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j)^2}{\sigma^2}.
\end{aligned}$$

Replacing $\boldsymbol{\beta}_j$ with $\boldsymbol{\beta}_j^{(r+1)}$ obtained from (3.15), we have the update

$$\begin{aligned}
(\sigma^2)^{(r+1)} &= \frac{\sum_{i=1}^{n} \sum_{j=1}^{g} \tau_j(y_i; \boldsymbol{\psi}^{(r)})(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j^{(r+1)})^2}{\sum_{i=1}^{n} \sum_{j=1}^{g} \tau_j(y_i; \boldsymbol{\psi}^{(r)})} \\
&= \frac{\sum_{j=1}^{g} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_j^{(r+1)})^T \boldsymbol{W}_j^{(r)} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_j^{(r+1)})}{\sum_{j=1}^{g} \mathbf{Tr}(\boldsymbol{W}_j^{(r)})}. \qquad (3.16)
\end{aligned}$$

3.1.4 Empirical Information Matrix

Following the definition of the empirical information matrix in section 2.4.2, suppose $\hat{\boldsymbol{\psi}}$ is a convergence point of an EM calculation and c is the iteration number for $\hat{\boldsymbol{\psi}}$. We have

$$\begin{aligned}
\left\{\frac{\partial \ell(\boldsymbol{\psi}; y_i)}{\partial \boldsymbol{\psi}}\right\}\Bigg|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} &= \left\{\mathbf{E}\Big[\frac{\partial \ell_c(\boldsymbol{\psi}; y_i, \boldsymbol{z}_i)}{\partial \boldsymbol{\psi}}\Big|\boldsymbol{y}, \hat{\boldsymbol{\psi}}\Big]\right\}\Bigg|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} \\
&= \left\{\mathbf{E}\left[\frac{\partial \sum_{j=1}^{g} z_{ij}\big(\log p_j + \log f_j(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}_j)\big)}{\partial \boldsymbol{\psi}}\Bigg|\boldsymbol{y}, \hat{\boldsymbol{\psi}}\right]\right\}\Bigg|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} \\
&= \left\{\frac{\partial \mathbf{E}\Big[\sum_{j=1}^{g} z_{ij}\big(\log p_j + \log f_j(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}_j)\big)\Big|\boldsymbol{y}, \hat{\boldsymbol{\psi}}\Big]}{\partial \boldsymbol{\psi}}\right\}\Bigg|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} \\
&= \left\{\frac{\partial \sum_{j=1}^{g}\big(\log p_j + \log f_j(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}_j)\big)\mathbf{E}\Big[z_{ij}\Big|\boldsymbol{y}, \hat{\boldsymbol{\psi}}\Big]}{\partial \boldsymbol{\psi}}\right\}\Bigg|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} \\
&= \left\{\frac{\sum_{j=1}^{g} \tau_j(y_i; \hat{\boldsymbol{\psi}})\partial\big(\log p_j + \log f_j(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}_j)\big)}{\partial \boldsymbol{\psi}}\right\}\Bigg|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}. \qquad (3.17)
\end{aligned}$$

Empirical Information Matrix for Mixing Proportion $\boldsymbol{p}$

To calculate $J_{\boldsymbol{pp}}$, the $(g-1)\times(g-1)$ empirical information matrix of mixing proportion $\boldsymbol{p}$, we first have the score function

$$\frac{\partial \ell(\boldsymbol{\psi}; y_i)}{\partial p_j}\Bigg|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} = \frac{\tau_j(y_i; \hat{\boldsymbol{\psi}})}{\hat{p}_j} - \frac{\tau_g(y_i; \hat{\boldsymbol{\psi}})}{\hat{p}_g}, \qquad (j = 1, \ldots, g-1). \qquad (3.18)$$

Then according to (2.26), we have

$$
\begin{aligned}
& J_{\boldsymbol{pp}}(\hat{\boldsymbol{\psi}}; \boldsymbol{y}) \\
= & \left\{ \sum_{i=1}^{n} \left\{ \frac{\partial \ell(\boldsymbol{\psi}; \boldsymbol{y})}{\partial \boldsymbol{p}} \right\} \left\{ \frac{\partial \ell(\boldsymbol{\psi}; \boldsymbol{y})}{\partial \boldsymbol{p}} \right\}^{T} \right\} \Bigg|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} \\
= & \begin{pmatrix} \sum_{i=1}^{n} \left(\frac{\tau_1(i)}{\hat{p}_1} - \frac{\tau_g(i)}{\hat{p}_g} \right)^2 & \cdots & \sum_{i=1}^{n} \left(\frac{\tau_1(i)}{\hat{p}_1} - \frac{\tau_g(i)}{\hat{p}_g} \right) \left(\frac{\tau_{g-1}(i)}{\hat{p}_{g-1}} - \frac{\tau_g(i)}{\hat{p}_g} \right) \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} \left(\frac{\tau_{g-1}(i)}{\hat{p}_{g-1}} - \frac{\tau_g(i)}{\hat{p}_g} \right) \left(\frac{\tau_1(i)}{\hat{p}_1} - \frac{\tau_g(i)}{\hat{p}_g} \right) & \cdots & \sum_{i=1}^{n} \left(\frac{\tau_{g-1}(i)}{\hat{p}_{g-1}} - \frac{\tau_g(i)}{\hat{p}_g} \right)^2 \end{pmatrix} \\
= & \begin{pmatrix} \sum_{i=1}^{n} \left(\frac{f_1(y_i) - f_g(y_i)}{f(y_i)} \right)^2 & \cdots & \sum_{i=1}^{n} \left(\frac{f_1(y_i) - f_g(y_i)}{f(y_i)} \right) \left(\frac{f_{g-1}(y_i) - f_g(y_i)}{f(y_i)} \right) \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} \left(\frac{f_{g-1}(y_i) - f_g(y_i)}{f(y_i)} \right) \left(\frac{f_1(y_i) - f_g(y_i)}{f(y_i)} \right) & \cdots & \sum_{i=1}^{n} \left(\frac{f_{g-1}(y_i) - f_g(y_i)}{f(y_i)} \right)^2 \end{pmatrix} \\
= & \sum_{i=1}^{n} \frac{1}{f(y_i)^2} \begin{pmatrix} (f_1(y_i) - f_g(y_i))^2 & \cdots & (f_1(y_i) - f_g(y_i))(f_{g-1}(y_i) - f_g(y_i)) \\ \vdots & \ddots & \vdots \\ (f_{g-1}(y_i) - f_g(y_i))(f_1(y_i) - f_g(y_i)) & \cdots & (f_{g-1}(y_i) - f_g(y_i))^2 \end{pmatrix}, \quad (3.19)
\end{aligned}
$$

where $\tau_j(i)$ represents $\tau_j(y_i; \hat{\boldsymbol{\psi}})$.

Empirical Information Matrix for Regression Parameters

To calculate the empirical information matrix for regression parameters $\boldsymbol{\xi} = \left(\boldsymbol{\beta}_1^T, \boldsymbol{\delta}_2^T, \ldots, \boldsymbol{\delta}_g^T\right)^T$, we first have the score function

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{\psi}; y_i)}{\partial \boldsymbol{\xi}} &= -\frac{\partial \sum_{j=1}^{g} \tau_j(y_i; \hat{\boldsymbol{\psi}})(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j)^2/(2\sigma^2)}{\partial \boldsymbol{\xi}} \\
&= -\frac{\partial \sum_{j=1}^{g} w_{ji}^{(c)}(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j)^2}{\partial \boldsymbol{\xi}} \\
&= -\frac{\partial \Big[w_{1i}^{(c)}(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_1)^2 + \sum_{j=2}^{g} w_{ji}^{(c)}(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j)^2 \Big]}{\partial \boldsymbol{\xi}} \\
&= -\frac{\partial \Big[\big(\sum_{j=1}^{g} w_{ji}^{(c)}\big)(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_1)^2 + \sum_{j=2}^{g} \boldsymbol{\delta}_j^T \boldsymbol{x}_{di} w_{ji}^{(c)} \boldsymbol{x}_{di}^T \boldsymbol{\delta}_j - 2 \sum_{j=2}^{g} w_{ji}^{(c)}(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_1) \boldsymbol{x}_{di} \boldsymbol{\delta}_j \Big]}{\partial \boldsymbol{\xi}} \\
&= -\frac{\partial \Big[\boldsymbol{\xi}^T A_i(c) \boldsymbol{\xi} - 2 \boldsymbol{\xi}^T B_i(c) \Big]}{\partial \boldsymbol{\xi}}, \qquad (3.20)
\end{aligned}$$

where

$$\boldsymbol{A}_i(c) = \begin{pmatrix} \boldsymbol{x}_i \big(\sum_{j=1}^{g} w_{ji}^{(c)}\big) \boldsymbol{x}_i^T & \boldsymbol{x}_i w_{2i}^{(c)} \boldsymbol{x}_{di}^T & \ldots & \boldsymbol{x}_i w_{gi}^{(c)} \boldsymbol{x}_{di}^T \\ \boldsymbol{x}_{di} w_{2i}^{(c)} \boldsymbol{x}_i^T & \boldsymbol{x}_{di} w_{2i}^{(c)} \boldsymbol{x}_{di}^T & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_{di} w_{gi}^{(c)} \boldsymbol{x}_i^T & 0 & \ldots & \boldsymbol{x}_{di} w_{gi}^{(c)} \boldsymbol{x}_{di}^T \end{pmatrix}, \qquad \boldsymbol{B}_i(c) = \begin{pmatrix} \boldsymbol{x}_i \big(\sum_{j=1}^{g} w_{ji}^{(c)}\big) y_i \\ \boldsymbol{x}_{di} w_{2i}^{(c)} y_i \\ \vdots \\ \boldsymbol{x}_{di} w_{gi}^{(c)} y_i \end{pmatrix},$$

$$\boldsymbol{x}_{di} = \boldsymbol{x}_i \begin{pmatrix} \boldsymbol{I}_{d \times d} \\ \boldsymbol{0}_{(m-d) \times d} \end{pmatrix}_{m \times d}.$$

and $w_{ji}^{(c)}$ is $W_j^{(c)}(i,i)$. Therefore,

$$\begin{aligned}\left.\frac{\partial \ell(\boldsymbol{\psi}; y_i)}{\partial \boldsymbol{\xi}}\right|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} &= \left.-\frac{\partial\Big[\boldsymbol{\xi}^T A_i(c)\boldsymbol{\xi} - 2\boldsymbol{\xi}^T B_i(c)\Big]}{\partial \boldsymbol{\xi}}\right|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} \\ &= -2A_i(c)\boldsymbol{\xi}^{(c)} + 2B_i(c). \end{aligned} \tag{3.21}$$

For the regression parameter vector $\boldsymbol{\beta}_1$, we have the score function

$$\begin{aligned}\left.\frac{\partial \ell(\boldsymbol{\psi}; y_i)}{\partial \boldsymbol{\beta}_1}\right|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} &= -2\left[\boldsymbol{x}_i\Big(\sum_{j=1}^{g} w_{ji}^{(c)}\Big)\boldsymbol{x}_i^T\boldsymbol{\beta}_1^{(c)} + \boldsymbol{x}_i w_{2i}^{(c)}\boldsymbol{x}_{di}^T\boldsymbol{\delta}_2^{(c)} + \cdots + \boldsymbol{x}_i w_{gi}^{(c)}\boldsymbol{x}_{di}^T\boldsymbol{\delta}_g^{(c)}\right] + 2\boldsymbol{x}_i\Big(\sum_{j=1}^{g} w_{ji}^{(c)}\Big)y_i \\ &= -\frac{1}{(\sigma^2)^{(c)}}\sum_{j=1}^{g}\tau_j(i)\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\beta}_j^{(c)} + \frac{1}{(\sigma^2)^{(c)}}\sum_{j=1}^{g}\tau_j(i)\boldsymbol{x}_i y_i \\ &= \frac{1}{(\sigma^2)^{(c)}}\sum_{j=1}^{g}\tau_j(i)\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_j^{(c)}) \\ &= \frac{1}{(\sigma^2)^{(c)}}\sum_{j=1}^{g}\tau_j(i)\boldsymbol{x}_i e_{ij} \end{aligned} \tag{3.22}$$

where $e_{ij} = y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_j^{(c)}$. Therefore, the empirical information matrix for $\boldsymbol{\beta}_1$ is

$$\begin{aligned} J_{\boldsymbol{\beta}_1\boldsymbol{\beta}_1} &= \frac{1}{(\sigma^4)^{(c)}}\sum_{i=1}^{n}\left\{\left[\sum_{j=1}^{g}\tau_j(i)e_{ij}\boldsymbol{x}_i\right]\left[\sum_{j=1}^{g}\tau_j(i)e_{ij}\boldsymbol{x}_i\right]^T\right\} \\ &= \frac{1}{(\sigma^4)^{(c)}}\sum_{i=1}^{n}\left\{\left[\sum_{j=1}^{g}\tau_j(i)e_{ij}\right]^2\boldsymbol{x}_i\boldsymbol{x}_i^T\right\} \\ &= \frac{1}{(\sigma^4)^{(c)}}\boldsymbol{X}^T\begin{pmatrix} \Big(\sum_{j=1}^{g}\tau_j(1)e_{1j}\Big)^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Big(\sum_{j=1}^{g}\tau_j(n)e_{nj}\Big)^2 \end{pmatrix}\boldsymbol{X} \end{aligned} \tag{3.23}$$

For the regression parameter vector $\boldsymbol{\delta}_k$, we first have the score function

$$\begin{aligned}\left.\frac{\partial \ell(\boldsymbol{\psi}; y_i)}{\partial \boldsymbol{\delta}_k}\right|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} &= -2\left[\boldsymbol{x}_{di} w_{ki}^{(c)} \boldsymbol{x}_i^T \boldsymbol{\theta}_1^{(c)} + \boldsymbol{x}_{di} w_{ki}^{(c)} \boldsymbol{x}_{di}^T \boldsymbol{\delta}_k^{(c)}\right] + 2\boldsymbol{x}_{di} w_{ki}^{(c)} y_i \\ &= 2\boldsymbol{x}_{di} w_{ki}^{(c)} e_{ik} \\ &= \frac{1}{(\sigma^2)^{(c)}} \boldsymbol{x}_{di} \tau_k(i) e_{ik}. \end{aligned} \tag{3.24}$$

Therefore, the empirical information matrix for $\boldsymbol{\delta}_k$ is

$$\begin{aligned} J_{\boldsymbol{\delta}_k \boldsymbol{\delta}_k} &= \frac{1}{(\sigma^4)^{(c)}} \sum_{i=1}^{n} \left\{ \tau_k^2(i) e_{ik}^2 \boldsymbol{x}_{di} \boldsymbol{x}_{di}^T \right\} \\ &= \frac{1}{(\sigma^4)^{(c)}} \boldsymbol{X}_d^T \begin{pmatrix} \tau_k^2(1) e_{1k}^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \tau_k^2(n) e_{nk}^2 \end{pmatrix} \boldsymbol{X}_d. \end{aligned} \tag{3.25}$$

Empirical Information Matrix for Variance

The score function of σ^2 is

$$\begin{aligned}\left.\frac{\partial \ell(\boldsymbol{\psi}; y_i)}{\partial \sigma^2}\right|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} &= \sum_{j=1}^{g} \tau_j(i) \left.\frac{\partial\big[-\log(2\pi\sigma^2)/2 - (y_i - \boldsymbol{x}_i^T \boldsymbol{\theta}_j)^2/2\sigma^2 \big]}{\partial \sigma^2}\right|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} \\ &= \sum_{j=1}^{g} \tau_j(i) \Big[-\frac{1}{2(\sigma^2)^{(c)}} + \frac{e_{ij}^2}{2(\sigma^4)^{(c)}} \Big] \end{aligned} \tag{3.26}$$

Therefore, the empirical information matrix for σ^2 is

$$J_{\sigma^2\sigma^2} = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{g} \tau_j(i) \Big[-\frac{1}{2(\sigma^2)^{(c)}} + \frac{e_{ij}^2}{2(\sigma^4)^{(c)}} \Big] \right\}^2. \tag{3.27}$$

3.2 Unequal Variance Case

In this section, we assume that the mixture components might have different error term variances $(\sigma_1^2, \ldots, \sigma_K^2)$ and that the EM algorithm has been iterated r times.

3.2.1 Mixing Proportion ($\boldsymbol{p}$) Estimation

The update for mixing proportion $\boldsymbol{p}$ follows the standard formula (2.21):

$$p_j^{(r+1)} = \sum_{i=1}^{n} \tau_j(y_i; \boldsymbol{\psi}^{(r)})/n \qquad (j = 1, \ldots, g).$$

3.2.2 Solution for $\boldsymbol{\beta}_j$ and σ_j^2

Define $\boldsymbol{\eta}$ as the parameter vector that defines all $\boldsymbol{\beta}_j$ and σ_j^2. According to section 2.4.1, the likelihood equation for $\boldsymbol{\eta}$ is

$$\frac{\partial \sum_{i=1}^{n} \sum_{j=1}^{g} \tau_j(y_i; \boldsymbol{\psi}^{(r)}) \log f_j(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}_j)}{\partial \boldsymbol{\eta}} = \mathbf{0} \tag{3.28}$$

in which

$$\log f_j(y_i, \boldsymbol{x}_i; \boldsymbol{\beta}_j) = -\frac{\log(2\pi\sigma_j^2)}{2} - \frac{(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j)^2}{2\sigma_j^2} \tag{3.29}$$

Since equation (3.28) is a nonlinear likelihood equation, we use coordinate descent methods [22] by going through the following two steps iteratively until apparent convergence:

Step 1: Solve for $\boldsymbol{\beta}_j$

Suppose we have current estimations of all $\hat{\sigma}_j^2$ and define $\boldsymbol{\xi}$ as a parameter vector that defines all $\boldsymbol{\beta}_j$. Then the likelihood equation for $\boldsymbol{\xi}$ becomes

$$\frac{\partial \sum_{i=1}^{n}\sum_{j=1}^{g} \tau_j(y_i;\boldsymbol{\psi}^{(r)}) \log f_j(y_i,\boldsymbol{x}_i;\boldsymbol{\theta}_j)}{\partial \boldsymbol{\xi}} = \mathbf{0}$$

$$\Rightarrow \quad \frac{\partial \sum_{i=1}^{n}\sum_{j=1}^{g} \tau_j(y_i;\boldsymbol{\psi}^{(r)}) \frac{(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}_j)^2}{2\hat{\sigma}_j^2}}{\partial \boldsymbol{\xi}} = \mathbf{0}$$

Change $\boldsymbol{W}_j^{(r)}$ in (3.9) to

$$\hat{\boldsymbol{V}}_j^{(r)} = \begin{pmatrix} \frac{\tau_j(y_1;\boldsymbol{\psi}^{(r)})}{2\hat{\sigma}_j^2} & 0 & \dots & 0 \\ 0 & \frac{\tau_j(y_2;\boldsymbol{\psi}^{(r)})}{2\hat{\sigma}_j^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{\tau_j(y_n;\boldsymbol{\psi}^{(r)})}{2\hat{\sigma}_j^2} \end{pmatrix} \tag{3.30}$$

and change $\boldsymbol{A}(r)$ and $\boldsymbol{B}(r)$ into $\hat{\boldsymbol{C}}(r)$ and $\hat{\boldsymbol{D}}(r)$ as follows,

$$\hat{\boldsymbol{C}}(r) = \begin{pmatrix} \boldsymbol{X}^T(\sum_{j=1}^{g}\hat{\boldsymbol{V}}_j^{(r)})\boldsymbol{X} & \boldsymbol{X}^T\hat{\boldsymbol{V}}_2^{(r)}\boldsymbol{X}_d & \dots & \boldsymbol{X}^T\hat{\boldsymbol{V}}_g^{(r)}\boldsymbol{X}_d \\ \boldsymbol{X}_d^T\hat{\boldsymbol{V}}_2^{(r)}\boldsymbol{X} & \boldsymbol{X}_d^T\hat{\boldsymbol{V}}_2^{(r)}\boldsymbol{X}_d & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{X}_d^T\hat{\boldsymbol{V}}_g^{(r)}\boldsymbol{X} & 0 & \dots & \boldsymbol{X}_d^T\hat{\boldsymbol{V}}_g^{(r)}\boldsymbol{X}_d \end{pmatrix},$$

$$\hat{\boldsymbol{D}}(r) = \begin{pmatrix} \boldsymbol{X}^T(\sum_{j=1}^{g} \hat{\boldsymbol{V}}_j^{(r)})\boldsymbol{Y} \\ \boldsymbol{X}_d^T \hat{\boldsymbol{V}}_2^{(r)} \boldsymbol{Y} \\ \vdots \\ \boldsymbol{X}_d^T \hat{\boldsymbol{V}}_g^{(r)} \boldsymbol{Y} \end{pmatrix}.$$

Then we can follow the same path as in section 3.1.2 to obtain the estimate

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\delta}}_2 \\ \vdots \\ \hat{\boldsymbol{\delta}}_g \end{pmatrix} = \hat{\boldsymbol{C}}(r)^{-1} \hat{\boldsymbol{D}}(r). \tag{3.31}$$

Step 2: Solve for σ_j^2

Suppose we have current estimates for all $\hat{\boldsymbol{\beta}}_j$. From (3.28) we have the likelihood equation for σ_j^2

$$\begin{aligned} & \frac{\partial \sum_{i=1}^{n} \sum_{j=1}^{g} \tau_j(y_i; \boldsymbol{\psi}^{(r)}) \log f_j(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}_j)}{\partial \sigma_j^2} = \mathbf{0} \\ \Rightarrow \quad & \frac{\partial \sum_{i=1}^{n} \tau_j(y_i; \boldsymbol{\psi}^{(r)}) \log f_j(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}_j)}{\partial \sigma_j^2} = \mathbf{0} \\ \Rightarrow \quad & \sum_{i=1}^{n} \tau_j(y_i; \boldsymbol{\psi}^{(r)}) \frac{\partial \log f_j(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}_j)}{\partial \sigma_j^2} = \mathbf{0} \\ \Rightarrow \quad & \sum_{i=1}^{n} \tau_j(y_i; \boldsymbol{\psi}^{(r)}) \left[-\frac{1}{\sigma_j^2} + \frac{(y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_j)^2}{\sigma_j^4} \right] = \mathbf{0} \end{aligned}$$

Therefore

$$\begin{aligned}\hat{\sigma}_j^2 &= \frac{\sum_{i=1}^{n} \tau_j(y_i; \boldsymbol{\psi}^{(r)})(y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_j)^2}{\sum_{i=1}^{n} \tau_j(y_i; \boldsymbol{\psi}^{(r)})} \\ &= \frac{(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_j)^T \hat{\boldsymbol{V}}_j^{(r)} (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_j)}{\mathbf{Tr}(\hat{\boldsymbol{V}}_j^{(r)})}. \end{aligned} \tag{3.32}$$

Once there is convergence for $\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\delta}}_2$, ..., $\hat{\boldsymbol{\delta}}_g$ and $\hat{\sigma}_j^2$, we take them as the solution for $(r+1)$st iteration.

3.2.3 Empirical Information Matrix

Suppose $\hat{\boldsymbol{\psi}}$ is the point of convergence of EM calculation, and c is its iteration number.

Empirical Information Matrix for Mixing Proportion $\boldsymbol{p}$

For $J_{\boldsymbol{pp}}$, we have the same solution as (3.19).

Empirical Information Matrix for Regression Parameters

To calculate the empirical information matrix for regression parameters $\boldsymbol{\xi} = \left(\boldsymbol{\beta}_1^T, \boldsymbol{\delta}_2^T, \ldots, \boldsymbol{\delta}_g^T\right)^T$, we first have the score function

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{\psi}; y_i)}{\partial \boldsymbol{\xi}} &= -\frac{\partial \sum_{j=1}^{g} \tau_j(y_i; \hat{\boldsymbol{\psi}})(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j)^2/(2\sigma_j^2)}{\partial \boldsymbol{\xi}} \\
&= -\frac{\partial \sum_{j=1}^{g} v_{ji}^{(c)}(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j)^2}{\partial \boldsymbol{\xi}} \\
&= -\frac{\partial \Big[v_{1i}^{(c)}(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_1)^2 + \sum_{j=2}^{g} v_{ji}^{(c)}(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j)^2 \Big]}{\partial \boldsymbol{\xi}} \\
&= -\frac{\partial \Big[\big(\sum_{j=1}^{g} v_{ji}^{(c)} \big)(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_1)^2 + \sum_{j=2}^{g} \boldsymbol{\delta}_j^T \boldsymbol{x}_{di} v_{ji}^{(c)} \boldsymbol{x}_{di}^T \boldsymbol{\delta}_j - 2 \sum_{j=2}^{g} v_{ji}^{(c)}(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_1) \boldsymbol{x}_{di} \boldsymbol{\delta}_j \Big]}{\partial \boldsymbol{\xi}} \\
&= -\frac{\partial \Big[\boldsymbol{\xi}^T C_i(c) \boldsymbol{\xi} - 2 \boldsymbol{\xi}^T D_i(c) \Big]}{\partial \boldsymbol{\xi}} \qquad (3.33)
\end{aligned}$$

where

$$\boldsymbol{C}_i(c) = \begin{pmatrix} \boldsymbol{x}_i \big(\sum_{j=1}^{g} v_{ji}^{(c)} \big) \boldsymbol{x}_i^T & \boldsymbol{x}_i v_{2i}^{(c)} \boldsymbol{x}_{di}^T & \ldots & \boldsymbol{x}_i v_{gi}^{(c)} \boldsymbol{x}_{di}^T \\ \boldsymbol{x}_{di} v_{2i}^{(c)} \boldsymbol{x}_i^T & \boldsymbol{x}_{di} v_{2i}^{(c)} \boldsymbol{x}_{di}^T & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_{di} v_{gi}^{(c)} \boldsymbol{x}_i^T & 0 & \ldots & \boldsymbol{x}_{di} v_{gi}^{(c)} \boldsymbol{x}_{di}^T \end{pmatrix}, \qquad \boldsymbol{D}_i = \begin{pmatrix} \boldsymbol{x}_i \big(\sum_{j=1}^{g} v_{ji}^{(c)} \big) y_i \\ \boldsymbol{x}_{di} v_{2i}^{(c)} y_i \\ \vdots \\ \boldsymbol{x}_{di} v_{gi}^{(c)} y_i \end{pmatrix},$$

and $v_{ji}^{(c)}$ is $\hat{V}_j^{(c)}(i,i)$. Therefore

$$\begin{aligned}
\left. \frac{\partial \ell(\boldsymbol{\psi}; y_i)}{\partial \boldsymbol{\xi}} \right|_{\boldsymbol{\psi} = \hat{\boldsymbol{\psi}}} &= \left. -\frac{\partial \Big[\boldsymbol{\xi}^T C_i(c) \boldsymbol{\xi} - 2 \boldsymbol{\xi}^T D_i(c) \Big]}{\partial \boldsymbol{\xi}} \right|_{\boldsymbol{\psi} = \hat{\boldsymbol{\psi}}} \\
&= -2 C_i(c) \boldsymbol{\xi}^{(c)} + 2 D_i(c). \qquad (3.34)
\end{aligned}$$

For the regression parameter vector $\boldsymbol{\beta}_1$, we have the score function

$$\begin{aligned}\left.\frac{\partial \ell(\boldsymbol{\psi}; y_i)}{\partial \boldsymbol{\beta}_1}\right|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} &= -2\left[\boldsymbol{x}_i\Big(\sum_{j=1}^{g} v_{ji}^{(c)}\Big)\boldsymbol{x}_i^T\boldsymbol{\beta}_1^{(c)} + \boldsymbol{x}_i v_{2i}^{(c)}\boldsymbol{x}_{di}^T\boldsymbol{\delta}_2^{(c)} + \cdots + \boldsymbol{x}_i v_{gi}^{(c)}\boldsymbol{x}_{di}^T\boldsymbol{\delta}_g^{(c)}\right] + 2\boldsymbol{x}_i\Big(\sum_{j=1}^{g} v_{ji}^{(c)}\Big)y_i \\ &= \sum_{j=1}^{g}\frac{\tau_j(i)}{(\sigma_j^2)^{(c)}}\boldsymbol{x}_i e_{ij}, \end{aligned} \tag{3.35}$$

where $e_{ij} = y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_j^{(c)}$. Therefore, the empirical information matrix for $\boldsymbol{\beta}_1$ is

$$\begin{aligned} J_{\boldsymbol{\beta}_1\boldsymbol{\beta}_1} &= \sum_{i=1}^{n}\left\{\left[\sum_{j=1}^{g}\frac{\tau_j(i)}{(\sigma_j^2)^{(c)}}e_{ij}\boldsymbol{x}_i\right]\left[\sum_{j=1}^{g}\frac{\tau_j(i)}{(\sigma_j^2)^{(c)}}e_{ij}\boldsymbol{x}_i\right]^T\right\} \\ &= \sum_{i=1}^{n}\left\{\left[\sum_{j=1}^{g}\frac{\tau_j(i)}{(\sigma_j^2)^{(c)}}e_{ij}\right]^2\boldsymbol{x}_i\boldsymbol{x}_i^T\right\} \\ &= \boldsymbol{X}^T\begin{pmatrix} \Big(\sum_{j=1}^{g}\frac{\tau_j(1)}{(\sigma_j^2)^{(c)}}e_{1j}\Big)^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Big(\sum_{j=1}^{g}\frac{\tau_j(n)}{(\sigma_j^2)^{(c)}}e_{nj}\Big)^2 \end{pmatrix}\boldsymbol{X} \end{aligned} \tag{3.36}$$

For regression parameter vector $\boldsymbol{\delta}_k$, we first have the score function

$$\begin{aligned}\left.\frac{\partial \ell(\boldsymbol{\psi}; y_i)}{\partial \boldsymbol{\delta}_k}\right|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} &= -2\left[\boldsymbol{x}_{di}v_{ki}^{(c)}\boldsymbol{x}_i^T\boldsymbol{\beta}_1^{(c)} + \boldsymbol{x}_{di}v_{ki}^{(c)}\boldsymbol{x}_{di}^T\boldsymbol{\delta}_k^{(c)}\right] + 2\boldsymbol{x}_{di}v_{ki}^{(c)}y_i \\ &= 2\boldsymbol{x}_{di}v_{ki}^{(c)}e_{ik} \\ &= \frac{1}{(\sigma_k^2)^{(c)}}\boldsymbol{x}_{di}\tau_k(i)e_{ik}. \end{aligned} \tag{3.37}$$

Therefore, the empirical information matrix for $\boldsymbol{\delta}_k$ is

$$\begin{aligned} J_{\boldsymbol{\delta}_k\boldsymbol{\delta}_k} &= \frac{1}{(\sigma_k^4)^{(c)}}\sum_{i=1}^{n}\{\tau_k^2(i)e_{ik}^2\boldsymbol{x}_{di}\boldsymbol{x}_{di}^T\} \\ &= \frac{1}{(\sigma_k^4)^{(c)}}\boldsymbol{X}_d^T\begin{pmatrix} \tau_k^2(1)e_{1k}^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \tau_k^2(n)e_{nk}^2 \end{pmatrix}\boldsymbol{X}_d. \end{aligned} \tag{3.38}$$

Empirical Information Matrix for Variance

For the score function of σ_j^2, we have

$$\begin{aligned} \left.\frac{\partial\ell(\boldsymbol{\psi};y_i)}{\partial\sigma_j^2}\right|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} &= \tau_j(i)\left.\frac{\partial\big[-\log(2\pi\sigma_j^2)/2-(y_i-\boldsymbol{x}_i^T\boldsymbol{\theta}_j)^2/2\sigma_j^2\big]}{\partial\sigma_j^2}\right|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} \\ &= \tau_j(i)\Big[-\frac{1}{2(\sigma_j^2)^{(c)}}+\frac{e_{ij}^2}{2(\sigma_j^4)^{(c)}}\Big]. \end{aligned} \tag{3.39}$$

Therefore, the empirical information matrix for σ_j^2 is

$$J_{\sigma_j^2\sigma_j^2} = \sum_{i=1}^{n}\left\{\tau_j(i)\Big[-\frac{1}{2(\sigma_j^2)^{(c)}}+\frac{e_{ij}^2}{2(\sigma_j^4)^{(c)}}\Big]\right\}^2. \tag{3.40}$$

3.3 Summary

In summary, without dealing with global maximum and initialization issues, we have the following EM procedure to calculate the MLE and their standard errors for RMLRM (3.1):

Procedure 3.3.1.

given $(\boldsymbol{p}^{(0)},\boldsymbol{\beta}^{(0)},\boldsymbol{\sigma}^{(0)})$

repeat

Calculate posterior probability matrix $\boldsymbol{\tau}$ *(2.18);*

if *Equal Variance* ***then***

Calculate regression parameters $\boldsymbol{\beta}^{(r)}$ *by (3.9), (3.1.2) and (3.15);*

Calculate residual variance $(\sigma^2)^{(r)}$ *by (3.16);*

else

repeat

Calculate regression parameters $\boldsymbol{\beta}^{(r)}$ *by (3.30), (3.2.2) and (3.31);*

Calculate residual variance $(\sigma^2)^{(r)}$ *by (3.32);*

until $(\boldsymbol{\beta}^{(r)}, (\boldsymbol{\sigma}^2)^{(r)})$ *converges*

endif

until $\boldsymbol{Q}(\boldsymbol{\psi}; \boldsymbol{\psi}^{(r)})$ *converges*

Calculate Empirical Information Matrix $J_{\boldsymbol{pp}}$ *by (3.19);*

if *Equal Variance* ***then***

Calculate Empirical Information Matrix $J_{\boldsymbol{\beta\beta}}$ *by (3.23) and (3.25);*

Calculate Empirical Information Matrix $J_{\sigma^2\sigma^2}$ *by (3.27);*

else

Calculate Empirical Information Matrix $J_{\boldsymbol{\beta\beta}}$ *by (3.36) and (3.38);*

Calculate Empirical Information Matrix $J_{\boldsymbol{\sigma^2\sigma^2}}$ *by (3.40);*

endif

return $(\boldsymbol{p}^{(*)}, \boldsymbol{\beta}^{(*)}, (\boldsymbol{\sigma}^2)^{(*)}, J_{\boldsymbol{pp}}, J_{\boldsymbol{\beta\beta}}, J_{\sigma^2\sigma^2})$

Chapter 4

Some Mixture Linear Regression Model (MLRM) Related Theoretic Results and Their Application

In this chapter, we first develop some theoretical results about mixture linear regression models and then give the EM initialization methods for two specific mixture linear regression models: namely, the *mixture intercept model (MIM)* and the *mixture slope model (MSM)*. Besides the direct applications on EM initialization, the theoretic results developed in this chapter also provide a theoretical foundation for power and sample size calculations for MIM and MSM and mixture intercept detection using residual analysis.

4.1 Linear Regression with Mixture Error Terms

Suppose we have the following linear model (a sample from it is shown in figure (4.1)):

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + d_i + \varepsilon_i \tag{4.1}$$

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

$$d_i = \begin{cases} \Delta_1 & \text{with probability } p_1, \\ \dots & \dots \\ \Delta_g & \text{with probability } p_g, \end{cases}$$

$$\mathbf{E}(d_i) = \Delta_1 p_1 + \ldots + \Delta_g p_g = 0$$

where d_i and ε_i are independent of each other. If we add d_i into the intercept, the resulting model is a g-component MIM.

For *ordinary least square (OLS)* regression on this model, the following two theorems hold:

Theorem 4.1.1. *If the design matrix X is full rank (Rank$(X) = m$), then the OLS estimate $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$ for model (4.1) is an unbiased consistent estimator of $\boldsymbol{\beta}$ with variance:*

$$\varrho^2 (X^T X)^{-1}. \tag{4.2}$$

where

$$\varrho^2 = \sigma^2 + \Delta_1^2 p_1 + \ldots + \Delta_g^2 p_g. \tag{4.3}$$

PROOF:

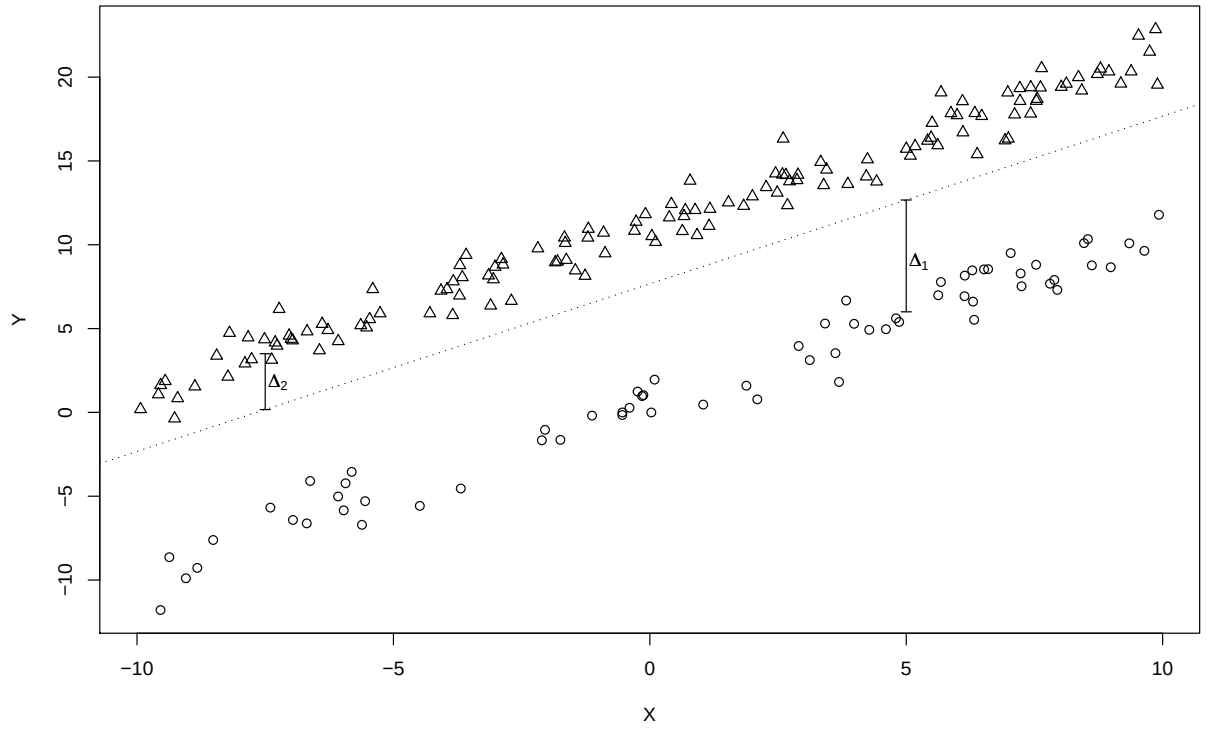

Figure 4.1: Simulated Data Scatterplot of Mixture Intercept Model

Using $\boldsymbol{y}$, $\boldsymbol{d}$ and $\boldsymbol{\varepsilon}$ to represent the observation and two error term vectors, we have

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (X^T X)^{-1} X^T \boldsymbol{y} \\ &= (X^T X)^{-1} X^T X \boldsymbol{\beta} + (X^T X)^{-1} X^T \boldsymbol{d} + (X^T X)^{-1} X^T \boldsymbol{\varepsilon} \\ &= \boldsymbol{\beta} + (X^T X)^{-1} X^T \boldsymbol{d} + (X^T X)^{-1} X^T \boldsymbol{\varepsilon}.\end{aligned}$$

Therefore

$$\begin{aligned}\mathbf{E}(\hat{\boldsymbol{\beta}}) &= \boldsymbol{\beta} + (X^T X)^{-1} X^T \mathbf{E}(\boldsymbol{d}) + (X^T X)^{-1} X^T \mathbf{E}(\boldsymbol{\varepsilon}) \\ &= \boldsymbol{\beta}.\end{aligned}$$

For the random variables d_i, we have

$$\mathbf{E}[d_i^2] = \Delta_1^2 p_1 + \ldots + \Delta_g^2 p_g.$$

For $i \neq j$

$$d_i d_j = \begin{cases} \Delta_1\Delta_2 & \text{with probability } 2p_1p_2, \\ \ldots & \ldots \\ \Delta_{g-1}\Delta_g & \text{with probability } 2p_{g-1}p_g, \\ \Delta_1^2 & \text{with probability } p_1^2, \\ \ldots & \ldots \\ \Delta_g^2 & \text{with probability } p_g^2, \end{cases}$$

and

$$\begin{aligned} \mathbf{E}[d_i d_j] &= \Delta_1^2 p_1^2 + \ldots + \Delta_g^2 p_g^2 + 2\Delta_1\Delta_2 p_1 p_2 + \ldots + 2\Delta_{g-1}\Delta_g p_{g-1} p_g \\ &= (\Delta_1 p_1 + \ldots + \Delta_g p_g)^2 = 0. \end{aligned}$$

Therefore

$$\mathbf{E}[\boldsymbol{d}\boldsymbol{d}^T] = \big(\Delta_1^2 p_1 + \ldots + \Delta_g^2 p_g\big) I. \tag{4.4}$$

Then

$$\begin{aligned} \mathbf{Var}(\hat{\boldsymbol{\beta}}) &= \mathbf{E}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T] \\ &= \mathbf{E}\Big\{ \big[(X^TX)^{-1}X^T\boldsymbol{d} + (X^TX)^{-1}X^T\boldsymbol{\varepsilon}\big] \big[\boldsymbol{d}^T X(X^TX)^{-1} + \boldsymbol{\varepsilon}^T X(X^TX)^{-1}\big] \Big\} \\ &= \mathbf{E}\big[(X^TX)^{-1}X^T\boldsymbol{d}\boldsymbol{d}^T X(X^TX)^{-1}\big] + \mathbf{E}\big[(X^TX)^{-1}X^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T X(X^TX)^{-1}\big] \\ &= \big(\Delta_1^2 p_1 + \ldots + \Delta_g^2 p_g\big)(X^TX)^{-1} + \sigma^2(X^TX)^{-1} \\ &= \varrho^2(X^TX)^{-1}. \end{aligned}$$

Since

$$lim_{n\to\infty}(X^TX)^{-1} = 0,$$

$\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$. ■

Theorem 4.1.2. *If the design matrix X is full rank, then the estimate*

$$\hat{\sigma}^2 = \frac{(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})}{n} \tag{4.5}$$

for model (4.1) has expectation:

$$\frac{(n-m)\varrho^2}{n}. \tag{4.6}$$

and

$$\hat{\sigma}^2 \xrightarrow{P} \varrho^2. \tag{4.7}$$

PROOF:

(1)

$$\begin{aligned} \boldsymbol{y} - X\hat{\boldsymbol{\beta}} &= X\boldsymbol{\beta} + \boldsymbol{d} + \boldsymbol{\varepsilon} - X(X^TX)^{-1}X^T\boldsymbol{y} \\ &= \boldsymbol{d} + \boldsymbol{\varepsilon} - H\boldsymbol{d} - H\boldsymbol{\varepsilon}, \end{aligned}$$

where

$$H = X(X^TX)^{-1}X^T. \tag{4.8}$$

Since

$$\begin{aligned} \hat{\sigma}^2 &= \frac{(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})}{n} \\ &= \frac{\boldsymbol{d}^T(I-H)\boldsymbol{d} + \boldsymbol{\varepsilon}^T(I-H)\boldsymbol{\varepsilon} + 2\boldsymbol{d}^T(I-H)\boldsymbol{\varepsilon}}{n}, \end{aligned}$$

$$
\begin{aligned}
\mathbf{E}(\hat{\sigma}^2) &= \mathbf{E}\Big\{ \frac{\boldsymbol{d}^T(I-H)\boldsymbol{d} + \boldsymbol{\varepsilon}^T(I-H)\boldsymbol{\varepsilon} + 2\boldsymbol{d}^T(I-H)\boldsymbol{\varepsilon}}{n} \Big\} \\
&= \mathbf{E}\Big\{ \frac{\mathrm{Tr}((I-H)\boldsymbol{d}\boldsymbol{d}^T) + \mathrm{Tr}((I-H)\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T)}{n} \Big\} \\
&= \frac{(\sigma^2 + \Delta_1^2 p_1 + \ldots + \Delta_g^2 p_g)\mathrm{Tr}(I-H)}{n} \\
&= \frac{(n-m)\varrho^2}{n}
\end{aligned}
$$

where m is the dimension of $\boldsymbol{\beta}$ and $\mathrm{Tr}(I-H)$ is the trace of the square matrix $I-H$.

(2)

According to *Lemma 3* of [2],

$$
\mathbf{Var}(\hat{\sigma}^2) = \frac{2\varrho^4\{\mathrm{Tr}((I-H)^2) + \frac{1}{2}\gamma_2 \boldsymbol{q}^T\boldsymbol{q}\}}{n^2}
$$

where $\boldsymbol{q}$ is the column vector of diagonal elements of $I-H$, and γ_2 is the standard measure of kurtosis for $\boldsymbol{y}$. From *Proposition 13.1.4* of [7], we have

$$
0 \leq q_i \leq 1.
$$

For $\mathrm{Tr}((I-H)^2)$, we have

$$
\begin{aligned}
\mathrm{Tr}((I-H)^2) &= \mathrm{Tr}(I-H) \\
&= \mathrm{Tr}(I) - \mathrm{Tr}(H) \\
&= n - \mathrm{Tr}(X(X^TX)^{-1}X^T) \\
&= n - \mathrm{Tr}(X^TX(X^TX)^{-1}) \\
&= n - \mathrm{rank}(X).
\end{aligned}
$$

For γ_2, we have

$$\begin{aligned}
\gamma_2 &= \frac{\mathbf{E}[(y_i - \mathbf{E}(y_i))^4]}{\varrho^4} - 3 \\
&= \frac{\mathbf{E}[d_i^4 + \varepsilon_i^4 + 4d_i^3\varepsilon_i + 4d_i\varepsilon_i^3 + 6d_i^2\varepsilon_i^2]}{\varrho^4} - 3 \\
&= \frac{\mathbf{E}[d_i^4 + \varepsilon_i^4 + 6d_i^2\varepsilon_i^2]}{\varrho^4} - 3 \\
&= \frac{\Delta_1^4 p_1 + \ldots + \Delta_g^4 p_g + 3\sigma^4 + 6\sigma^2(\Delta_1^2 p_1 + \ldots + \Delta_g^2 p_g)}{\varrho^4} - 3 \\
&= o(n).
\end{aligned}$$

Therefore,

$$lim_{n\to\infty}\mathbf{Var}(\hat{\sigma}^2) = 0,$$

and

$$\hat{\sigma}^2 \xrightarrow{P} \varrho^2.$$

■

4.2 OLS for Mixture Models

We first study the unrestricted mixture linear regression model:

$$y = \begin{cases} \boldsymbol{x}^T\boldsymbol{\beta}_1 + \varepsilon_j & \text{with probability } p_1, \\ \ldots & \\ \boldsymbol{x}^T\boldsymbol{\beta}_g + \varepsilon_g & \text{with probability } p_g, \end{cases} \tag{4.9}$$

$$\varepsilon_j \overset{iid}{\sim} N(0, \sigma_j^2),$$

$$0 \le p_j \le 1, \qquad \sum_{j=1}^{g} p_j = 1.$$

When we use $\mathbf{E}(\boldsymbol{y}) = X\boldsymbol{\beta}$ to model the data from model (4.9) with OLS, we have the following theorem:

Theorem 4.2.1. *If the design matrix X is full rank, then the expectation of OLS estimate $\hat{\boldsymbol{\beta}}$ for model (4.9) will be:*

$$\mathbf{E}(\hat{\boldsymbol{\beta}}) = p_1\boldsymbol{\beta_1} + \cdots + p_g\boldsymbol{\beta_g}. \tag{4.10}$$

PROOF:

Since the OLS estimate $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y},$$

$$\begin{aligned} \mathbf{E}(\hat{\boldsymbol{\beta}}) &= (X^T X)^{-1} X^T \mathbf{E}(\boldsymbol{y}) \\ &= (X^T X)^{-1} X^T \left(p_1 X^T \boldsymbol{\beta_1} + \ldots + p_g X^T \boldsymbol{\beta_g}\right) \\ &= p_1\boldsymbol{\beta_1} + \cdots + p_g\boldsymbol{\beta_g}. \end{aligned}$$

■

Consider the restricted mixture linear regression model:

$$y = \begin{cases} \boldsymbol{x}^T\boldsymbol{\beta}_1 + \varepsilon_1 & \text{with probability } p_1, \\ \ldots & \\ \boldsymbol{x}^T\boldsymbol{\beta}_g + \varepsilon_g & \text{with probability } p_g, \end{cases} \tag{4.11}$$

$$\varepsilon_j \overset{iid}{\sim} N(0, \sigma_j^2)$$

$$0 \le p_j \le 1, \qquad \sum_{j=1}^{g} p_j = 1,$$

$$\boldsymbol{\beta}_j = \boldsymbol{\beta}_1 + \begin{bmatrix} \boldsymbol{\delta}_j \\ \mathbf{0} \end{bmatrix} \qquad (j = 2, \ldots, g).$$

We call variables (dimension of $d \times 1$)with nonzero $\boldsymbol{\delta}_j$ unrestricted, and the other variables restricted. That is, the regression coefficients of $\boldsymbol{x}_{d+1}$, ..., $\boldsymbol{x}_m$ are the same for each component.

When we apply theorem 4.2.1 to the restricted mixture linear regression model (4.11), we have the following corollary:

Corollary 4.2.2. *If the design matrix X is full rank, then the expectation of the OLS estimator $\hat{\boldsymbol{\beta}}$ applied to the data from model (4.11) is:*

$$\mathbf{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta_1} + p_2 \begin{bmatrix} \boldsymbol{\delta}_2 \\ \mathbf{0} \end{bmatrix} + \cdots + p_g \begin{bmatrix} \boldsymbol{\delta}_g \\ \mathbf{0} \end{bmatrix}. \tag{4.12}$$

■

If we use OLS to model mixture linear regression models, we will have unbiased estimation for the restricted components. The expectations of the OLS estimator of unrestricted parameters are weighted sum of the true parameters, with the mixing proportions the corresponding weights.

4.3 EM Initialization For Mixture Intercept Model

Suppose we have the following mixture intercept model (MIM):

$$y = \begin{cases} \alpha + \boldsymbol{x}^T\boldsymbol{\beta} + \varepsilon & \text{with probability } p_1, \\ \alpha + \delta_2 + \boldsymbol{x}^T\boldsymbol{\beta} + \varepsilon & \text{with probability } p_2, \\ \dots & \dots \\ \alpha + \delta_g + \boldsymbol{x}^T\boldsymbol{\beta} + \varepsilon & \text{with probability } p_g. \end{cases} \tag{4.13}$$

This can be changed into the following mixture error term model:

$$y = \begin{cases} \alpha + \overline{\delta} + \boldsymbol{x}^T\boldsymbol{\beta} - \overline{\delta} + \varepsilon & \text{with probability } p_1, \\ \alpha + \overline{\delta} + \boldsymbol{x}^T\boldsymbol{\beta} + \delta_2 - \overline{\delta} + \varepsilon & \text{with probability } p_2, \\ \dots & \dots \\ \alpha + \overline{\delta} + \boldsymbol{x}^T\boldsymbol{\beta} + \delta_g - \overline{\delta} + \varepsilon & \text{with probability } p_g, \end{cases} \tag{4.14}$$

where

$$\overline{\delta} = p_2\delta_2 + \ldots + p_g\delta_g.$$

According to Theorem 4.1.1, the OLS estimator $\hat{\boldsymbol{\theta}} = (X^TX)^{-1}X^T\boldsymbol{y}$ is an unbiased consistent estimator which converges to

$$\boldsymbol{\theta} = \begin{pmatrix} \alpha + \overline{\delta} \\ \boldsymbol{\beta} \end{pmatrix} \tag{4.15}$$

and the corresponding residual is approximately

$$e = \begin{cases} -\overline{\delta} + \varepsilon & \text{when observation comes from component 1,} \\ \delta_2 - \overline{\delta} + \varepsilon & \text{when observation comes from component 2,} \\ \dots & \dots \\ \delta_g - \overline{\delta} + \varepsilon & \text{when observation comes from component g,} \end{cases} \tag{4.16}$$

which can be approximated with a mixture normal distribution. Based on this mixture residual distribution, we have the following EM initialization procedure for the mixture intercept model (supposing that we have an initial value for the mixing proportion $\boldsymbol{p}$):

Procedure 4.3.1 (**EM Initialization Procedure for Mixture Intercept Model**).

- *Estimate the OLS parameters from the data;*
- *Calculate the residuals;*
- *Sort residuals;*
- *Split the sample into g groups according* $\boldsymbol{p}$ *and their residual rank;*
- *Calculate initial values for other parameters accordingly.* ■

4.4 EM Initialization For Mixture Slope Model

Suppose we have the following mixture slope model (MSM):

$$y = \begin{cases} \alpha + \gamma x_t + \boldsymbol{x}_c^T \boldsymbol{\beta} + \varepsilon & \text{with probability } p_1, \\ \alpha + (\gamma + \delta_2) x_t + \boldsymbol{x}_c^T \boldsymbol{\beta} + \varepsilon & \text{with probability } p_2, \\ \cdots & \cdots \\ \alpha + (\gamma + \delta_g) x_t + \boldsymbol{x}_c^T \boldsymbol{\beta} + \varepsilon & \text{with probability } p_g. \end{cases} \tag{4.17}$$

This can be changed into following model:

$$y = \begin{cases} \alpha + (\gamma + \overline{\delta}) x_t + \boldsymbol{x}_c^T \boldsymbol{\beta} - \overline{\delta} x_t + \varepsilon & \text{with probability } p_1, \\ \alpha + (\gamma + \overline{\delta}) x_t + \boldsymbol{x}_c^T \boldsymbol{\beta} + (\delta_2 - \overline{\delta}) x_t + \varepsilon & \text{with probability } p_2, \\ \cdots & \cdots \\ \alpha + (\gamma + \overline{\delta}) x_t + \boldsymbol{x}_c^T \boldsymbol{\beta} + (\delta_g - \overline{\delta}) x_t + \varepsilon & \text{with probability } p_g. \end{cases} \tag{4.18}$$

If we use OLS to model (4.18). Then, the "scaled residual" ($e' = e/x_t$) for model (4.18) will approximately to be

$$e' = \begin{cases} -\overline{\delta} & \text{when observation comes from component 1,} \\ \delta_2 - \overline{\delta} & \text{when observation comes from component 2,} \\ \dots & \dots \\ \delta_g - \overline{\delta} & \text{when observation comes from component g.} \end{cases} \tag{4.19}$$

We have the following EM initialization procedure for mixture slope model (suppose we already have initial value for $\boldsymbol{p}$):

Procedure 4.4.1 (**EM Initialization Procedure for Mixture Slope Model**)**.**

- *Estimate OLS parameters from data;*
- *Calculate the "scaled residuals";*
- *Sort "scaled residuals";*
- *Split the sample into g groups according* $\boldsymbol{p}$ *and their rank of "scaled residuals";*
- *Calculate initial values for other parameters accordingly.* ■

In applications we may need to modify procedure 4.4.1 to handle extreme values of ε/x_t caused by x_t near 0.

Chapter 5

Power and Sample Size Calculations for Three Two-Component Mixture Models

5.1 Motivation and LRTS Decomposition

In this chapter, we study an approximate approach to perform power and sample size calculations for detecting some two-component mixture models using the likelihood ratio test statistic (LRTS).

Suppose we have a sample of observations $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, and we want to test:

$$\begin{aligned} \mathbf{H_0} &: y_i \overset{iid}{\sim} f(y_i; \boldsymbol{\theta}_0) \triangleq g_0(y_i; \boldsymbol{\psi}_0), \\ \mathbf{H_A} &: y_i \overset{iid}{\sim} pf(y_i; \boldsymbol{\theta}_1) + (1-p)f(y_i; \boldsymbol{\theta}_2) \triangleq g_A(y_i; \boldsymbol{\psi}_A), \end{aligned} \tag{5.1}$$

where $\boldsymbol{\psi}_0 = \boldsymbol{\theta}_0$, $\boldsymbol{\psi}_A = (p, \boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$. The LRTS is:

$$\lambda(\widehat{\boldsymbol{\psi}}; \boldsymbol{y}) = 2\{\ell(\widehat{\boldsymbol{\psi}}_A; \boldsymbol{y}) - \ell(\widehat{\boldsymbol{\psi}}_0; \boldsymbol{y})\}, \tag{5.2}$$

in which,

$$\ell(\widehat{\boldsymbol{\psi}}_j; \boldsymbol{y}) = \sum_{i=1}^{n} \log g_j(y_i; \widehat{\boldsymbol{\psi}}_j), \qquad (j \in \{0, A\}), \tag{5.3}$$

where $\boldsymbol{\psi} = (\boldsymbol{\psi}_0^T, \boldsymbol{\psi}_A^T)^T$, and $\widehat{\boldsymbol{\psi}}_0$, $\widehat{\boldsymbol{\psi}}_A$ and $\widehat{\boldsymbol{\psi}}$ are the MLEs of their corresponding parameters.

In order to calculate the power and sample size for detecting two-component mixture distribution against homogenous distribution, we need two LRTS distributions:

- LRTS distribution when $\mathbf{H_0}$ is true, which we call the *null LRTS distribution*;
- LRTS distribution when $\mathbf{H_A}$ is true, which we call the *alternative LRTS distribution.*

We estimate the null LRTS distribution through simulations in the next chapter.

If we have a two-component normal mixture model

$$y_i \overset{iid}{\sim} pN(\mu_1, \sigma_A^2) + (1-p)N(\mu_2, \sigma_A^2), \tag{5.4}$$

and we mistakenly assume it is a homogenous normal model

$$y_i \overset{iid}{\sim} N(\mu_0, \sigma_0^2), \tag{5.5}$$

then for the MLEs $\hat{\mu}_0$ and $\hat{\sigma}_0^2$, we have

$$\mathbf{E}(\hat{\mu}_0) = p\mu_1 + (1-p)\mu_2 \triangleq \mu_0, \tag{5.6}$$

$$\mathbf{E}(\hat{\sigma}_0^2) = \sigma_A^2 + p(1-p)(\mu_1 - \mu_2)^2 \triangleq \sigma_0^2. \tag{5.7}$$

With the real two-component mixture model (5.4) and corresponding misspecified homogenous model (5.5), (5.6) and(5.7), the alternative LRTS equation (5.2) for this normal mixture detecting problem can be decomposed as:

$$\begin{aligned}\lambda(\widehat{\boldsymbol{\psi}};\boldsymbol{y}) &= 2\{\ell(\widehat{\boldsymbol{\psi}}_A;\boldsymbol{y})-\ell(\widehat{\boldsymbol{\psi}}_0;\boldsymbol{y})\}\\ &= 2\{\ell(\boldsymbol{\psi}_A;\boldsymbol{y})-\ell(\boldsymbol{\psi}_0;\boldsymbol{y})\}+2\{\ell(\widehat{\boldsymbol{\psi}}_A;\boldsymbol{y})-\ell(\boldsymbol{\psi}_A;\boldsymbol{y})\}\\ &\quad -2\{\ell(\widehat{\boldsymbol{\psi}}_0;\boldsymbol{y})-\ell(\boldsymbol{\psi}_0;\boldsymbol{y})\}. \qquad (5.8)\end{aligned}$$

In the decomposition equation (5.8), $2\{\ell(\boldsymbol{\psi}_A;\boldsymbol{y})-\ell(\boldsymbol{\psi}_0;\boldsymbol{y})\}$ is easily computed once we have the parameters of p, μ_1, μ_2 and σ_A, which are the design parameters for power and sample size calcualtion problems. Later on, we will call $2\{\ell(\boldsymbol{\psi}_A;\boldsymbol{y})-\ell(\boldsymbol{\psi}_0;\boldsymbol{y})\}$ the *nominal alternative LRTS*. We conjecture that $\{\ell(\boldsymbol{\psi}_A;\boldsymbol{y})-\ell(\boldsymbol{\psi}_0;\boldsymbol{y})\}$ is independent of $\{\ell(\widehat{\boldsymbol{\psi}}_A;\boldsymbol{y})-\ell(\boldsymbol{\psi}_A;\boldsymbol{y})\}-\{\ell(\widehat{\boldsymbol{\psi}}_0;\boldsymbol{y})-\ell(\boldsymbol{\psi}_0;\boldsymbol{y})\}$, and $2\{\ell(\widehat{\boldsymbol{\psi}}_A;\boldsymbol{y})-\ell(\boldsymbol{\psi}_A;\boldsymbol{y})\}-2\{\ell(\widehat{\boldsymbol{\psi}}_0;\boldsymbol{y})-\ell(\boldsymbol{\psi}_0;\boldsymbol{y})\}$ could be approximated by $\chi^2_{d_A-d_0}$, where d_A and d_0 are dimensions of $\boldsymbol{\psi}_A$ and $\boldsymbol{\psi}_0$ respectively. Therefore, we have a possibly useful approximation for LRTS as

$$\lambda(\widehat{\boldsymbol{\psi}};\boldsymbol{y}) \dot{\sim} 2\{\ell(\boldsymbol{\psi}_A;\boldsymbol{y})-\ell(\boldsymbol{\psi}_0;\boldsymbol{y})\}+\chi^2_{d_A-d_0}. \qquad (5.9)$$

In the following sections, for the two-component *normal mixture model (NMM)*, the *mixture intercept model (MIM)* and the *mixture slope model (MSM)*, we first provide an approach to obtain nominal alternative LRTS distributions of $2\{\ell(\boldsymbol{\psi}_A;\boldsymbol{y})-\ell(\boldsymbol{\psi}_0;\boldsymbol{y})\}$. Then we obtain power and sample size formulas based on the approximated alternative LRTS distributions from (5.9) and the simulated null LRTS distributions.

5.2 Nominal Alternative LRTS Distributions for Normal Mixture Model (NMM)

Suppose we have an observation from following NMM

$$y = \begin{cases} N(\mu_1, \sigma_A^2) & \text{with probability } p, \\ N(\mu_2, \sigma_A^2) & \text{with probability } 1-p, \end{cases} \tag{5.10}$$

which is equivalent to

$$y = \begin{cases} \mu_1 + \varepsilon & \text{with probability } p, \\ \mu_2 + \varepsilon & \text{with probability } 1-p, \end{cases} \tag{5.11}$$

where

$$\varepsilon \overset{iid}{\sim} N(0, \sigma_A^2).$$

Assume that we know the true parameters for this model. Then the corresponding log likelihood function $\ell_2(\boldsymbol{\psi}_2; y)$ for a single observation y will be

$$\begin{aligned} \ell_2(\boldsymbol{\psi}_2; y) &= \log(p\phi(y; \mu_1, \sigma_A^2) + (1-p)\phi(y; \mu_2, \sigma_A^2)) \\ &= \log\left(\frac{p}{\sqrt{2\pi}\sigma_A} e^{-\frac{(y-\mu_1)^2}{2\sigma_A^2}} + \frac{1-p}{\sqrt{2\pi}\sigma_A} e^{-\frac{(y-\mu_2)^2}{2\sigma_A^2}} \right) \\ &= \begin{cases} \log\left(\frac{p}{\sqrt{2\pi}\sigma_A} e^{-\frac{\varepsilon^2}{2\sigma_A^2}} + \frac{1-p}{\sqrt{2\pi}\sigma_A} e^{-\frac{(\varepsilon-\delta)^2}{2\sigma_A^2}} \right) \\ \qquad\qquad \text{when } y \text{ comes from component 1} \\ \log\left(\frac{p}{\sqrt{2\pi}\sigma_A} e^{-\frac{(\varepsilon+\delta)^2}{2\sigma_A^2}} + \frac{1-p}{\sqrt{2\pi}\sigma_A} e^{-\frac{\varepsilon^2}{2\sigma_A^2}} \right) \\ \qquad\qquad \text{when } y \text{ comes from component 2} \end{cases} \end{aligned} \tag{5.12}$$

where $\delta = \mu_2 - \mu_1$ and $\boldsymbol{\psi}_2 = (p, \mu_1, \mu_2, \sigma_A)^T$.

If we mistakenly assume that the data come from a homogeneous normal distribution, then according to the previous section, the underlying misspecified model is

$$y_i \overset{iid}{\sim} N(\mu_0, \sigma_0^2), \tag{5.13}$$

$$\mu_0 = p\mu_1 + (1-p)\mu_2, \tag{5.14}$$

$$\sigma_0^2 = \sigma_A^2 + p(1-p)(\mu_1 - \mu_2)^2. \tag{5.15}$$

The corresponding log likelihood function for a single observation $\ell_1(\boldsymbol{\psi}_1; y)$ is

$$\begin{aligned}
\ell_1(\boldsymbol{\psi}_1; y) &= \log\left(\frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(y-\mu_0)^2}{2\sigma_0^2}}\right) \\
&= -\log(\sqrt{2\pi}\sigma_0) - \frac{(y-\mu_0)^2}{2\sigma_0^2} \\
&= \begin{cases} -\log(\sqrt{2\pi}\sigma_0) - \dfrac{(-(1-p)\delta + \varepsilon)^2}{2\sigma_0^2} & \text{when } y \text{ comes from component 1} \\ \\ -\log(\sqrt{2\pi}\sigma_0) - \dfrac{(p\delta + \varepsilon)^2}{2\sigma_0^2} & \text{when } y \text{ comes from component 2} \end{cases}
\end{aligned} \tag{5.16}$$

where $\boldsymbol{\psi}_1 = (p, \delta, \sigma_0)^T$.

Therefore, for a single observation, we have the LRTS value:

$$
\begin{aligned}
\lambda(\boldsymbol{\psi}; y) &= \\
&= 2(\ell_2(\boldsymbol{\psi}_2; y) - \ell_1(\boldsymbol{\psi}_1; y)) \\
&= \begin{cases}
2\log\left(\frac{p}{\sqrt{2\pi}\sigma_A} e^{-\frac{\varepsilon^2}{2\sigma_A^2}} + \frac{1-p}{\sqrt{2\pi}\sigma_A} e^{-\frac{(\varepsilon-\delta)^2}{2\sigma_A^2}}\right) + \log(2\pi\sigma_0^2) + \frac{(-(1-p)\delta+\varepsilon)^2}{\sigma_0^2} \\
\qquad\qquad \text{when } y \text{ comes from component 1} \\
2\log\left(\frac{p}{\sqrt{2\pi}\sigma_A} e^{-\frac{(\varepsilon+\delta)^2}{2\sigma_A^2}} + \frac{1-p}{\sqrt{2\pi}\sigma_A} e^{-\frac{\varepsilon^2}{2\sigma_A^2}}\right) + \log(2\pi\sigma_0^2) + \frac{(p\delta+\varepsilon)^2}{\sigma_0^2} \\
\qquad\qquad \text{when } y \text{ comes from component 2}
\end{cases}
\end{aligned}
\tag{5.17}
$$

where $\boldsymbol{\psi} = (p, \mu_1, \mu_2, \sigma_A)^T$.

From (5.17) we can see that $\lambda(\boldsymbol{\psi}; y) = \lambda(p, \mu_1, \mu_2, \sigma_A; y)$ is a determinate function of y with p, μ_1, μ_2 and σ_A as parameters, and *independent and identically distributed (i.i.d.)* y gives *i.i.d.* $\lambda(\boldsymbol{\psi}; y)$. There are no closed-form formulas to describe the distribution of $\lambda(\boldsymbol{\psi}; y)$, but we can estimate its mean $\mathbf{E}(\lambda(\boldsymbol{\psi}; y))$ and standard deviation σ_λ by the sample average $\overline{\lambda}$ and sample standard deviation s_λ through the following simulation procedure:

Procedure 5.2.1.

- *Given p, μ_1, μ_2 and σ_A;*
- *Let $\delta = \mu_2 - \mu_1$ and calculate σ_0 according to (5.15);*
- *Set sample size $N = 2000$ (which is large enough to get accurate estimates of $\mathbf{E}(\lambda(\boldsymbol{\psi}; y))$ and σ_λ);*
- *Create a random sample $\boldsymbol{\varepsilon}$ from $N(0, \sigma_A^2)$;*
- *Partition $\boldsymbol{\varepsilon}$ into 2 groups with proportion of group 1 as p;*

- *Calculate the realizations of $\lambda(\boldsymbol{\psi}; y)$ according to (5.17);*
- *Calculate the average $\overline{\lambda}$ and sample standard deviation s_λ from the realizations of $\lambda(\boldsymbol{\psi}; y)$.* ■

Using the *central limit theorem*, for a sample of n observation $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, the nominal alternative LRTS distribution for detecting two-component normal mixture against homogenous normal distribution can be approximated by:

$$\begin{aligned} \lambda(\boldsymbol{\psi}; \boldsymbol{y}) &= \sum_{i=1}^{n} \lambda(\boldsymbol{\psi}; y_i) \\ &\dot{\sim} N(n\overline{\lambda}, ns_\lambda^2). \end{aligned} \tag{5.18}$$

5.3 Power and Sample Size Calculations for Normal Mixture Model

In section 6.3, we will find the null LRTS distribution $\lambda_0(n)$ through simulation for specified sample size n. Here, we assume $\lambda_0(n)$ is known. Given null and alternative LRTS distributions, the power and sample size calculations can be illustrated by Figure 5.1. We use the following procedure to calculate the power:

Procedure 5.3.1.

1. *Obtain the nominal alternative LRTS distribution $N(n\overline{\lambda}, ns_\lambda^2)$ with design parameters p, μ_1, μ_2 and σ_A according to Procedure 5.2.1;*
2. *Obtain the approximate alternative LRTS distribution $\lambda(\widehat{\boldsymbol{\psi}}; \boldsymbol{y}) \dot{\sim} N(n\overline{\lambda}, ns_\lambda^2) + \chi^2_{d_A - d_0}$;*
3. *Find $q_{1-\alpha}(n)$, the $1-\alpha$ quantile point of $\lambda_0(n)$;*

4. *Calculate the power by integrating the area enclosed by* $y = 0$, $x = q_{1-\alpha}(n)$ *and the curve of the alternative LRTS distribution* $\lambda(\widehat{\boldsymbol{\psi}}; \boldsymbol{y})$. ■

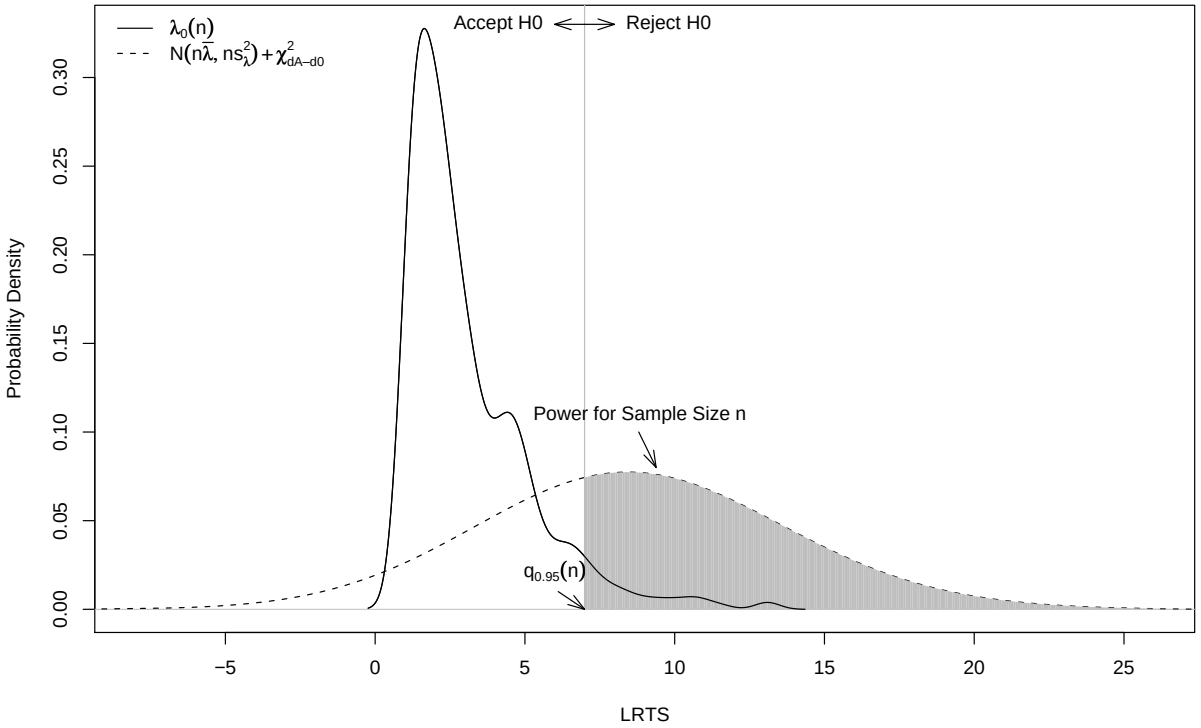

Figure 5.1: Power and sample size calculation by null and alternative LRTS. $q_{0.95}(n)$ is the 95% quantile point of the null LRTS distribution $\lambda_0(n)$ with sample size n. The shaded area represents the power.

Because the approximate alternative LRTS distribution $\lambda(\widehat{\boldsymbol{\psi}}; \boldsymbol{y}) \dot{\sim} N(n\overline{\lambda}, ns_\lambda^2) + \chi^2_{d_A - d_0}$ does not have a closed-form formula, it is impossible to have analytic solutions for power and sample size. On the other hand, from the calculation point of view, there is no difference between calculating $q_{1-\alpha}(n)$ of $\lambda_0(n)$ and calculating that of $\lambda_0(n) - \chi^2_{d_A - d_0}$. Therefore, if we can change comparing $\lambda_0(n)$ (null statistic) against $N(n\overline{\lambda}, ns_\lambda^2) + \chi^2_{d_A - d_0}$ (alternative statistic) into comparing $\lambda_0(n) - \chi^2_{d_A - d_0}$ (null statistic) against $N(n\overline{\lambda}, ns_\lambda^2)$ (alternative statistic) in the processes of power and sample size calculations, then we will have closed-form solutions because of the simple analytic form of $N(n\overline{\lambda}, ns_\lambda^2)$. To validate this change for hypothesis testing, we use the following theorem:

Theorem 5.3.2. *Given two random variables X and Y that have probability density functions $f_X(x)$ and $f_Y(y)$ respectively, and s a real number, then*

$$\Pr(X \leq Y + s) = \Pr(X - Y \leq s) \tag{5.19}$$

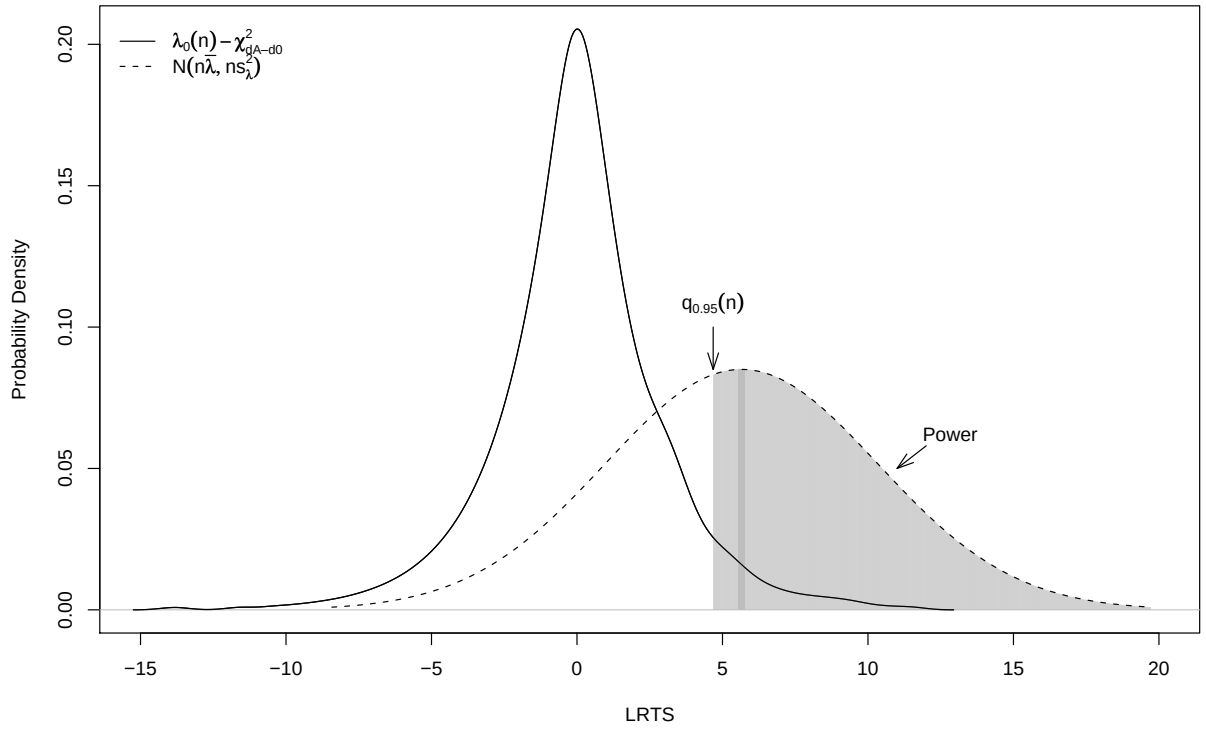

Figure 5.2: Power and sample size calculation by LRTS. $q_{0.95}(n)$ is the 95% quantile point of the distribution of $(\lambda_0(n) - \chi^2_{d_A - d_0})$. The shaded area represents the power.

With Theorem 5.3.2, we can use Figure 5.2 to carry out sample size and power calculations. When we have a specified normal mixture model and have a certain sample size n, we can first calculate $\overline{\lambda}$ and s_λ according to procedure 5.2.1, then calculate the power by following *power formula*:

$$\text{Power}(\boldsymbol{\psi}, \alpha, n) = 1 - \Phi\left(\frac{q_{1-\alpha}(n) - n\overline{\lambda}}{\sqrt{n}s_\lambda}\right), \tag{5.20}$$

where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution. If

we have a target power, then we have:

$$\begin{aligned} & \text{Power}(\boldsymbol{\psi}, \alpha, n) = 1 - \Phi\left(\tfrac{q_{1-\alpha}(n) - n\overline{\lambda}}{\sqrt{n}s_\lambda}\right) \\ \Rightarrow \quad & Z_\beta = \frac{q_{1-\alpha}(n) - n\overline{\lambda}}{\sqrt{n}s_\lambda} \\ \Rightarrow \quad & \sqrt{n}s_\lambda Z_\beta = q_{1-\alpha}(n) - n\overline{\lambda} \end{aligned} \tag{5.21}$$

where $Z_\beta = \Phi^{-1}(1 - \text{Power})$. In section 6.3, we will find that $q_{1-\alpha}(n_1) \approx q_{1-\alpha}(n_2)$ when $100 \leq n_1, n_2 \leq 1600$. Therefore, we define $q_{1-\alpha}$ as the average of $q_{1-\alpha}(n)$ for those sample sizes, and use it to equation (5.21). The sample size formula then is:

$$n = \left\lceil \frac{2q_{1-\alpha}\overline{\lambda} + \sigma_\lambda^2 Z_\beta^2 + \sqrt{\sigma_\lambda^4 Z_\beta^4 + 4q_{1-\alpha}\overline{\lambda}\sigma_\lambda^2 Z_\beta^2}}{2\overline{\lambda}^2} \right\rceil \tag{5.22}$$

where $\lceil x \rceil$ means smallest integer that greater than x.

5.4 Nominal Alternative LRTS Distributions for Mixture Intercept Model (MIM)

Suppose we have an observation from the following MIM

$$y = \begin{cases} \alpha + \boldsymbol{x}^T\boldsymbol{\beta} + \varepsilon & \text{with probability } p, \\ (\alpha + \delta) + \boldsymbol{x}^T\boldsymbol{\beta} + \varepsilon & \text{with probability } 1 - p, \end{cases} \tag{5.23}$$

$$\varepsilon \overset{iid}{\sim} N(0, \sigma^2).$$

Assume that we know the parameters for this model. Then the log likelihood function $\ell_2(\boldsymbol{\psi}_2; y)$ for a single observation y will be

$$\begin{aligned}
\ell_2(\boldsymbol{\psi}_2; y) &= \log(pf_1(y; \boldsymbol{\theta}_1, \sigma^2) + (1-p)f_2(y; \boldsymbol{\theta}_2, \sigma^2)) \\
&= \log\left(\frac{p}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\boldsymbol{x}^T\boldsymbol{\theta}_1)^2}{2\sigma^2}} + \frac{1-p}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\boldsymbol{x}^T\boldsymbol{\theta}_2)^2}{2\sigma^2}}\right) \\
&= \begin{cases} \log\left(\frac{p}{\sqrt{2\pi}\sigma} e^{-\frac{\varepsilon^2}{2\sigma^2}} + \frac{1-p}{\sqrt{2\pi}\sigma} e^{-\frac{(\varepsilon-\delta)^2}{2\sigma^2}}\right) \\ \qquad\qquad \text{when } y \text{ comes from component 1} \\ \log\left(\frac{p}{\sqrt{2\pi}\sigma} e^{-\frac{(\varepsilon+\delta)^2}{2\sigma^2}} + \frac{1-p}{\sqrt{2\pi}\sigma} e^{-\frac{\varepsilon^2}{2\sigma^2}}\right) \\ \qquad\qquad \text{when } y \text{ comes from component 2} \end{cases}
\end{aligned} \tag{5.24}$$

where $\boldsymbol{\theta}_1 = (\alpha, \boldsymbol{\beta}^T)^T$, $\boldsymbol{\theta}_2 = (\alpha, \delta, \boldsymbol{\beta}^T)^T$ and $\boldsymbol{\psi}_2 = (\alpha, \delta, \boldsymbol{\beta}^T, \sigma, p)^T$

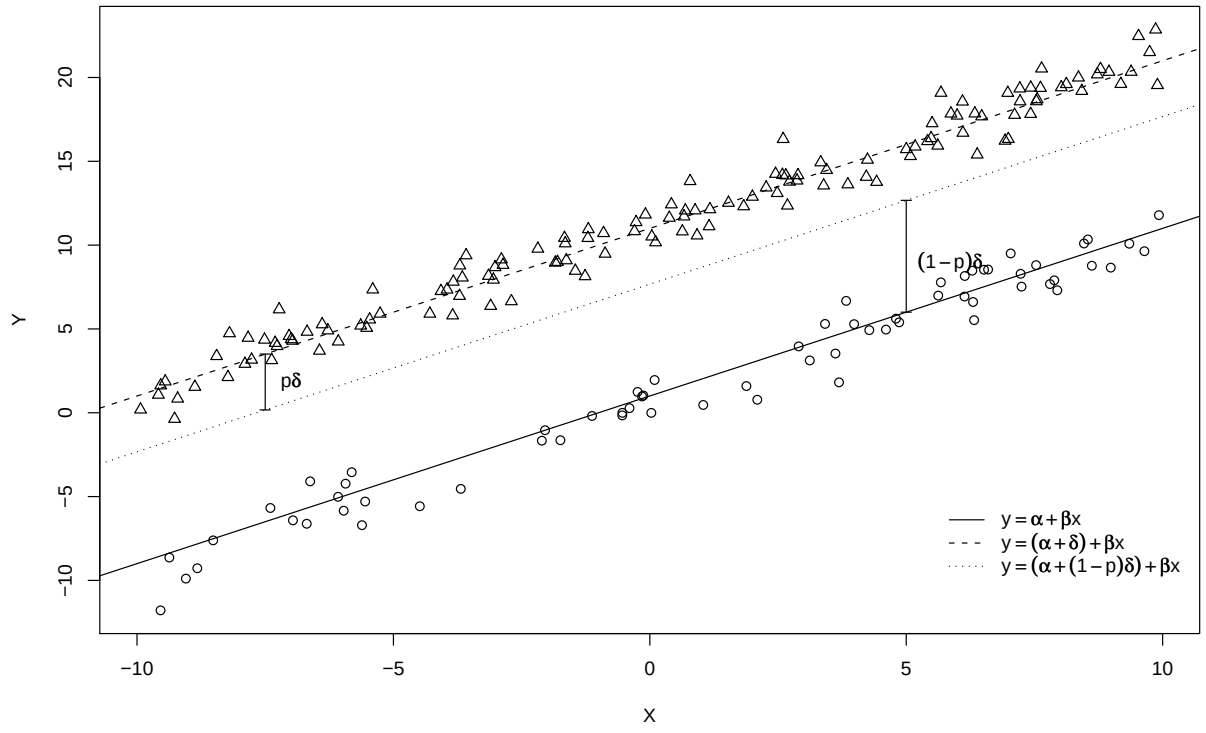

Figure 5.3: Hypothetic Plot for Single Linear Regression on Mixture Linear Model

Suppose we mistakenly assume the data come from a homogeneous linear model and use OLS to model it. Then according to model (4.1), Theorem 4.1.1, corollary 4.2.2 and figure 5.3, the corresponding misspecified model is

$$y = \alpha + (1-p)\delta + \boldsymbol{x}^T\boldsymbol{\beta} + \epsilon \tag{5.25}$$
$$\epsilon \overset{iid}{\sim} N(0, \varrho^2)$$

where

$$\varrho^2 = p(1-p)\delta^2 + \sigma^2 \tag{5.26}$$

Then the log likelihood function (also assume we have the exact value of $\boldsymbol{\psi}_1$) for a single observation $\ell_1(\boldsymbol{\psi}_1; y)$ with this model is

$$\begin{aligned}
\ell_1(\boldsymbol{\psi}_1; y) &= \log\left(\frac{1}{\sqrt{2\pi}\varrho} e^{-\frac{(y-\boldsymbol{x}^T\boldsymbol{\theta})^2}{2\varrho^2}}\right) \\
&= -\log(\sqrt{2\pi}\varrho) - \frac{(y-\boldsymbol{x}^T\boldsymbol{\theta})^2}{2\varrho^2} \\
&= \begin{cases} -\log(\sqrt{2\pi}\varrho) - \dfrac{(-(1-p)\delta+\varepsilon)^2}{2\varrho^2} & \text{when } y \text{ comes from component 1} \\ \\ -\log(\sqrt{2\pi}\varrho) - \dfrac{(p\delta+\varepsilon)^2}{2\varrho^2} & \text{when } y \text{ comes from component 2} \end{cases}
\end{aligned} \tag{5.27}$$

where $\boldsymbol{\theta} = (\alpha + (1-p)\delta, \boldsymbol{\beta}^T)^T$ and $\boldsymbol{\psi}_1 = (\alpha + (1-p)\delta, \boldsymbol{\beta}^T, \varrho)^T$.

Therefore, for a single observation, we have the log likelihood ratio:

$$
\begin{aligned}
\lambda(\boldsymbol{\psi}; y) &= \\
&= 2(\ell_2(\boldsymbol{\psi}_2; y) - \ell_1(\boldsymbol{\psi}_1; y)) \\
&= \begin{cases}
2\log\left(\dfrac{p}{\sqrt{2\pi}\sigma}e^{-\dfrac{\varepsilon^2}{2\sigma^2}} + \dfrac{1-p}{\sqrt{2\pi}\sigma}e^{-\dfrac{(\varepsilon-\delta)^2}{2\sigma^2}}\right) + \log(2\pi\varrho^2) + \dfrac{(-(1-p)\delta+\varepsilon)^2}{\varrho^2} \\
\qquad\qquad \text{when } y \text{ comes from component 1} \\
2\log\left(\dfrac{p}{\sqrt{2\pi}\sigma}e^{-\dfrac{(\varepsilon+\delta)^2}{2\sigma^2}} + \dfrac{1-p}{\sqrt{2\pi}\sigma}e^{-\dfrac{\varepsilon^2}{2\sigma^2}}\right) + \log(2\pi\varrho^2) + \dfrac{(p\delta+\varepsilon)^2}{\varrho^2} \\
\qquad\qquad \text{when } y \text{ comes from component 2}
\end{cases}
\end{aligned}
\tag{5.28}
$$

where $\boldsymbol{\psi} = (\alpha, \delta, \boldsymbol{\beta}^T, \sigma, p)^T$.

From (5.28) we can see that if we know the values of all parameters, $\lambda(\boldsymbol{\psi}; y) = \lambda(\alpha, \delta, \boldsymbol{\beta}^T, \sigma, p; y)$ is a completely specified function of y with σ, δ and p as parameters. Since y is *independent and identically distributed (i.i.d.)*, $\lambda(\boldsymbol{\psi}; y)$ is *i.i.d.* as well. Therefore, we can estimate its mean $\mathbf{E}(\lambda(\boldsymbol{\psi}; y))$ and standard deviation σ_λ by sample average $\overline{\lambda}$ and sample standard deviation s_λ through the following simulation procedure:

Procedure 5.4.1.

- *Given σ, δ, p;*
- *Set sample size $N = 2000$ (which is large enough to get accurate estimates on $\mathbf{E}(\lambda(\boldsymbol{\psi}; y))$ and σ_λ);*
- *Calculate ϱ according to (5.26);*
- *Create a random sample $\boldsymbol{\varepsilon}$ from $N(0, \sigma^2)$;*
- *Partition $\boldsymbol{\varepsilon}$ into 2 groups with proportion of group 1 as p;*

- *Calculate the realizations of* $\lambda(\boldsymbol{\psi}; y)$ *according to (5.28);*
- *Calculate* $\overline{\lambda}$ *and* s_λ *from the realizations of* $\lambda(\boldsymbol{\psi}; y)$. ■

Using the *central limit theorem*, for a sample of n observation $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, the nominal alternative LRTS distribution for detecting two-component mixture intercept against homogenous intercept model can be approximated by:

$$\begin{aligned} \lambda(\boldsymbol{\psi}; \boldsymbol{y}) &= \sum_{i=1}^{n} \lambda(\boldsymbol{\psi}; y_i) \\ &\dot{\sim} N(n\overline{\lambda}, n\sigma_\lambda^2). \end{aligned} \tag{5.29}$$

5.5 Power and Sample Size Calculations for Mixture Intercept Model

Following the same derivation in section 5.3, the power and sample size formulas are the same as (5.20) and (5.22).

5.6 Nominal Alternative LRTS Distributions for Mixture Slope Model (MSM)

Suppose we have an observation from the following MSM

$$y = \begin{cases} \alpha + \gamma x_t + \boldsymbol{x}_c^T \boldsymbol{\beta} + \varepsilon & \text{with probability } p, \\ \alpha + (\gamma + \delta) x_t + \boldsymbol{x}_c^T \boldsymbol{\beta} + \varepsilon & \text{with probability } 1 - p, \end{cases} \tag{5.30}$$

$$\varepsilon \overset{iid}{\sim} N(0, \sigma^2).$$

If we know the parameters for this model, then the log likelihood function $\ell_2(\boldsymbol{\psi}_2; y, x_t)$ for a single observation y will be

$$\ell_2(\boldsymbol{\psi}_2; y, x_t) = \begin{cases} \log\left(\frac{p}{\sqrt{2\pi}\sigma} e^{-\frac{\varepsilon^2}{2\sigma^2}} + \frac{1-p}{\sqrt{2\pi}\sigma} e^{-\frac{(\varepsilon-\delta x_t)^2}{2\sigma^2}}\right) \\ \qquad \text{when } y \text{ comes from component 1,} \\ \log\left(\frac{p}{\sqrt{2\pi}\sigma} e^{-\frac{(\varepsilon+\delta x_t)^2}{2\sigma^2}} + \frac{1-p}{\sqrt{2\pi}\sigma} e^{-\frac{\varepsilon^2}{2\sigma^2}}\right) \\ \qquad \text{when } y \text{ comes from component 2,} \end{cases}$$

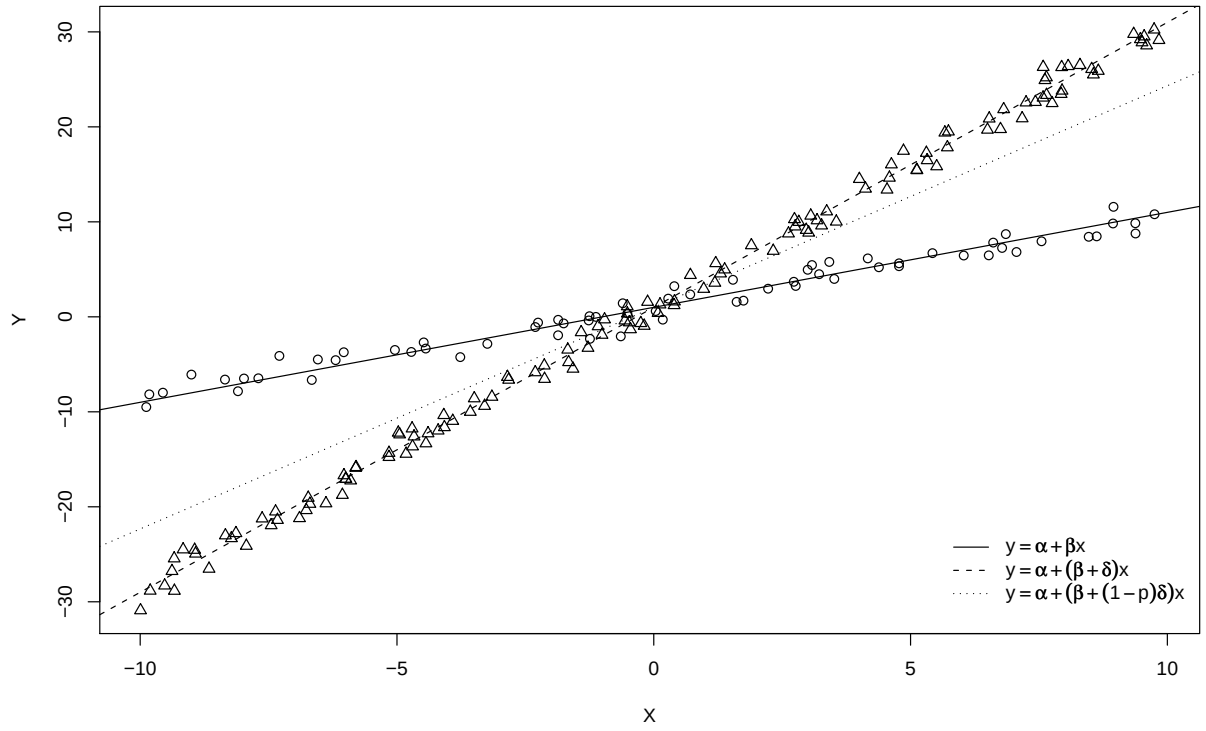

Figure 5.4: Hypothetic Plot for Mixture Linear Model with Mixture Slopes

Suppose we mistakenly assume that the data come from a homogeneous linear model and use OLS to model it. Then according to model (5.30), corollary (4.2.2) and figure (5.4), the corresponding misspecified model is

$$\hat{y} = \alpha + ((1-p)\delta + \gamma)x_t + \boldsymbol{x}_c^T \boldsymbol{\beta} \tag{5.31}$$

and the residual $e = y - \hat{y}$ against this model is

$$e = \begin{cases} -(1-p)\delta x_t + \varepsilon & \text{when } y \text{ comes from component 1,} \\ p\delta x_t + \varepsilon & \text{when } y \text{ comes from component 2.} \end{cases} \tag{5.32}$$

Then

$$\begin{aligned} \mathbf{E}[e] &= p\mathbf{E}[-(1-p)\delta x_t + \varepsilon] + (1-p)\mathbf{E}[p\delta x_t + \varepsilon] \\ &= -p(1-p)\delta\mathbf{E}[x_t] + (1-p)p\delta\mathbf{E}[x_t] \\ &= 0, \end{aligned} \tag{5.33}$$

and

$$\begin{aligned} \varrho^2 &= Var[e] \\ &= p\mathbf{E}\big[(-(1-p)\delta x_t + \varepsilon)^2\big] + (1-p)\mathbf{E}\big[(p\delta x_t + \varepsilon)^2\big] \\ &= p\big((1-p)^2\delta^2\mathbf{E}[x_t^2] + \sigma^2\big) + (1-p)\big(p^2\delta^2\mathbf{E}[x_t^2] + \sigma^2\big) \\ &= \sigma^2 + p(1-p)\delta^2\mathbf{E}[x_t^2]. \end{aligned} \tag{5.34}$$

The log likelihood function $\ell_1(\boldsymbol{\psi}_1; y, x_t)$ corresponding to the misspecified homogeneous linear model is

$$\ell_1(\boldsymbol{\psi}_1; y, x_t) = \begin{cases} -\log(\sqrt{2\pi}\varrho) - \dfrac{(-(1-p)\delta x_t + \varepsilon)^2}{2\varrho^2} \\ \qquad\qquad \text{when } y \text{ comes from component 1,} \\ -\log(\sqrt{2\pi}\varrho) - \dfrac{(p\delta x_t + \varepsilon)^2}{2\varrho^2} \\ \qquad\qquad \text{when } y \text{ comes from component 2.} \end{cases}$$

Therefore, the LRTS for a single observation is:

$$\begin{aligned}
\lambda(\boldsymbol{\psi}; y, x_t) &= 2(\ell_2(\boldsymbol{\psi}_2; y, x_t) - \ell_1(\boldsymbol{\psi}_1; y, x_t)) \\
&= \begin{cases}
2\log\left(\frac{p}{\sqrt{2\pi}\sigma}e^{-\frac{\varepsilon^2}{2\sigma^2}} + \frac{1-p}{\sqrt{2\pi}\sigma}e^{-\frac{(\varepsilon-\delta x_t)^2}{2\sigma^2}}\right) \\
+\log(2\pi\varrho^2) + \frac{(-(1-p)\delta x_t + \varepsilon)^2}{\varrho^2} \\
\qquad\qquad \text{when } y \text{ comes from component 1,} \\
2\log\left(\frac{p}{\sqrt{2\pi}\sigma}e^{-\frac{(\varepsilon+\delta x_t)^2}{2\sigma^2}} + \frac{1-p}{\sqrt{2\pi}\sigma}e^{-\frac{\varepsilon^2}{2\sigma^2}}\right) \\
+\log(2\pi\varrho^2) + \frac{(p\delta x_t + \varepsilon)^2}{\varrho^2} \\
\qquad\qquad \text{when } y \text{ comes from component 2.}
\end{cases}
\end{aligned} \tag{5.35}$$

From (5.35) we can see that if we know the exact value on all the parameters, $\lambda(\boldsymbol{\psi}; y, x_t) = \lambda(\alpha, \gamma, \delta, \boldsymbol{\beta}^T, \sigma, p; y, x_t)$ is a completely specified function of y and x_t with σ, δ and p as parameters. Since y is *independent and identically distributed (i.i.d.)*, $\lambda(\boldsymbol{\psi}; y, x_t)$ is *i.i.d.* as well. Therefore, the mean $\overline{\lambda}$ and standard deviation σ_λ of $\lambda(\boldsymbol{\psi}; y, x_t)$ can be calculated by procedure 5.6.1.

Procedure 5.6.1.

- *Given σ, δ, p;*
- *Set sample size $N = 2000$ (which is large enough to get accurate estimates on $\mathbf{E}(\lambda(\boldsymbol{\psi}; y))$ and σ_λ);*
- *Calculate ϱ according to (5.34);*
- *Create two independent random samples $\boldsymbol{\varepsilon}$ from $N(0, \sigma^2)$ and x_t from certain distribution (decided by design problems);*

- *Partition* $\boldsymbol{\varepsilon}$ *and* x_t *into* 2 *groups with proportion of group* 1 *as* p*;*
- *Calculate the realizations of* $\lambda(\boldsymbol{\psi}; y, x_t)$ *according to (5.35);*
- *Calculate* $\overline{\lambda}$ *and* s_λ *from the realizations of* $\lambda(\boldsymbol{\psi}; y, x_t)$*.* ■

Using the *central limit theorem*, for a sample of n observations with $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ and $\boldsymbol{x}_t = (x_{t_1}, \ldots, x_{t_n})^T$, the nominal alternative LRTS distribution for detecting two-component mixture slope against homogenous slope model can be approximated by:

$$\begin{aligned} \lambda(\boldsymbol{\psi}; \boldsymbol{y}, \boldsymbol{x}_t) &= \sum_{i=1}^{n} \lambda(\boldsymbol{\psi}; y_i, x_{t_i}) \\ &\dot{\sim} N(n\overline{\lambda}, n\sigma_\lambda^2). \end{aligned} \tag{5.36}$$

5.7 Power and Sample Size Calculations for Mixture Slope Model

Following the same derivation in section 5.3, the power and sample size formulas are the same as (5.20) and (5.22).

Chapter 6

Simulation Study

6.1 Models and Tasks

The purpose of this chapter is to find the best implementation strategies for solving *restricted mixture linear regression model (RMLRM)* estimation and statistical inference problems; and to verify some proposed methods from previous chapters. Our main interests are in the following three mixture models:

Normal Mixture Model (NMM)

$$y \sim \begin{cases} N(\mu_1, \sigma^2) & \text{with probability } p, \\ N(\mu_2, \sigma^2) & \text{with probability } 1-p. \end{cases} \tag{6.1}$$

Mixture Intercept Model (MIM)

$$y \sim \begin{cases} \alpha + \gamma x_t + \beta x_c + \varepsilon & \text{with probability } p, \\ (\alpha + \delta) + \gamma x_t + \beta x_c + \varepsilon & \text{with probability } 1-p. \end{cases} \tag{6.2}$$

Mixture Slope Model (MSM)

$$y \sim \begin{cases} \alpha + \gamma x_t + \beta x_c + \varepsilon & \text{with probability } p, \\ \alpha + (\gamma + \delta) x_t + \beta x_c + \varepsilon & \text{with probability } 1 - p. \end{cases} \tag{6.3}$$

In order to show that RMLRM is more powerful to detect mixtures for the data from MIM and MSM, the following *unrestricted mixture linear regression model (UMLRM)*

$$y \sim \begin{cases} \alpha_1 + \gamma_1 x_t + \beta_1 x_c + \varepsilon & \text{with probability } p, \\ \alpha_2 + \gamma_2 x_t + \beta_2 x_c + \varepsilon & \text{with probability } 1 - p \end{cases} \tag{6.4}$$

is also included.

In the following sections, we conduct simulations to

- investigate the ideal EM implementation strategies to find MLEs for NMM, MIM and MSM.
- investigate the empirical null distributions for certain NMM, MIM, MSM and UMLRM models.
- study the power to detect a mixture intercept model by NMM, MIM and UMLRM models.
- study the power to detect a mixture slope model by MSM and UMLRM models.
- verify and obtain application guidelines for the power and sample size formulas (5.20) and (5.22) for NMM, MIM and MSM models.

6.2 Pilot Study

In order to have correct null and alternative LRTS, statisticians typically run the EM algorithm multiple times with different starting points, and set proper stopping criterion and maximum number of iterations for each EM run. By doing this, we hope that each EM run will reach a local maximum, and the best solution from all the local maximums will be the global maximum with a high probability.

For NMM, MIM and MSM, we conduct pilot studies to explore the EM implementation strategies on number of EM runs, starting point selection, stopping criteria and maximum number of iterations. We confirm that our EM procedures have a high probability of finding the global maximum for LRTS under certain implementation conditions. Because of the similarity of the results between NMM and MIM, we only report the pilot study results for NMM and MIM.

6.2.1 Pilot Study for Normal Mixture Model (NMM)

For NMM with sample sizes 100 (1600), stopping criterion $10e^{-8}$ ($10e^{-12}$) and maximum number of iterations 2000 (5000), we carry out 25 runs of the following procedure (Procedure 6.2.1) to use a two-component normal mixture model (6.1) to detect mixtures from the data from a standard normal distribution ($N(0,1)$).

Procedure 6.2.1 (**Pilot Study Procedure for NMM**).

1: *Create a random sample* $\boldsymbol{x}$ *with sample size* n *from* $N(0,1)$*;*

2: *Obtain* 19 *starting values* p_0 *for mixture proportion* p *as* $0.05, 0.10, \ldots, 0.95$*;*

3: *Obtain* 250 *random starting values* p_0 *for mixture proportion* p *by sampling from a uni-*

form $U(0,1)$ distribution;

4: *For every p_0 from step 2 and 3, calculate the starting values for the remaining three parameters according to:*

$$\mu_{10} = \sum_{i=1}^{m} x_{(i)}/m$$

$$\mu_{20} = \sum_{i=m+1}^{n} x_{(i)}/(n-m)$$

$$\sigma_0^2 = \left[\sum_{i=1}^{m}(x_{(i)} - \mu_{10})^2 + \sum_{i=m+1}^{n}(x_{(i)} - \mu_{20})^2\right] \Big/ (n-2)$$

where $x_{(i)}$ denotes the ith sample order statistic and m is the integer part of np_0;

5: *Carry out EM algorithm on each starting point, and save final results $\hat{p}$, $\hat{\mu}_1$, $\hat{\mu}_2$ and $\hat{\sigma}$.*

■

In appendix A, we show the figures of pilot study results for four samples with sample size 1600. These four samples give us the following three typical convergence patterns of the LRTS results from multiples runs of EM on a single sample:

Pattern 1: The LRTS results for all *random starting points (RSPs)* converge to very few domains of convergence. The global minimum and global maximum are two convergence domains. This pattern mostly occurs when the global maximum of LRTS greater than 2.0. It is demonstrated by the results from sample 5 (figure A.5 and figure A.6);

Pattern 2: The LRTS results for most *random starting points (RSPs)* converge to a very few domains of convergence, and a small proportion of LRTS results do not converge. The global minimum and global maximum are two convergence domains. This pattern usually occurs when the global maximum of LRTS ranges between 1.0 and 2.0. It is demonstrated by the results from sample 2 (figure A.1 and figure A.2);

Pattern 3: The LRTS results for all *random starting points (RSPs)* do not converge, or the global maximum is not a convergence domain to which significant number of EM runs converge. This pattern mostly occurs when the maximum LRTS is less than 1.0 and is demonstrated by the results from sample 6 and sample 17 (figure A.9, figure A.10 figure A.13 and figure A.14).

For the NMM pilot study with sample size 1600, among the 25 simulations, there are 7 simulations that have Pattern 1; 8 simulations with Pattern 2; and 8 simulations with Pattern 3. The corresponding ranges of their final maximal LRTS are $[1.02, 4.85]$, $[0.93, 2.20]$ and $[0.09, 1.21]$. For the NMM pilot study with sample size 100, among 25 simulations, there are 14 simulations that converge with Pattern 1; 11 simulations with Pattern 2; and no simulations with Pattern 3. The corresponding ranges of final maximal LRTS are $[0.88, 10.89]$ for pattern 1 and $[0.30, 2.63]$ for Pattern 2 respectively. Clearly, it is much more difficult to find the global maximum for samples with larger sample size. Therefore we need to set more stringent stopping criteria and increase the maximum number of iterations as sample size increases.

From those pilot study results, we also find that the simulations with 19 predetermined starting point $p_0 \in \{0.05, 0.10, \ldots, 0.95\}$ also have a high probability of finding the largest observed LRTS. To confirm this and find reasonable stopping criterion and maximum number of iterations, for the same 25 random samples with sample size 1600, we use $p_0 \in \{0.05, 0.10, \ldots, 0.95\}$ and several combinations of various stopping criteria and maximum number of iterations carry out EM optimization again. For every sample, we find the five highest LRTS results under every simulation condition. All the results are listed in Appendix A. By comparing the LRTS results obtained from the stringent simulation condition (250 RSPs, stopping criterion $10e^{-12}$ and maximum number of iterations 5000) and the most relaxed simulation condition (19 fixed starting points, stopping criterion $10e^{-4}$

and maximum number of iterations 1000), we find the 0, 25, 50, 75 and 100 percentiles for the LRTS difference are 0.0009, 0.0026, 0.0071, 0.0163 and 0.0989 respectively. The largest LRTS difference for LRTS ≥ 3.0 is only 0.0026.

6.2.2 Pilot Study for Mixture Slope Model (MSM)

Parallel to the pilot study for normal mixture model, we also carry out a pilot study of the mixture slope model (MSM) (6.3) to detect mixture slopes (respect to x_t) with the following null homogeneous linear model

$$y = 1 + 1x_t + 1x_c + \varepsilon, \tag{6.5}$$

$$\varepsilon \overset{iid}{\sim} N(0,1), \quad x_t \overset{iid}{\sim} U(0,10), \quad x_c \overset{iid}{\sim} Bernoulli(0.5),$$

according to procedure 6.2.2. In this pilot study, we also run the pilot study procedure 25 times, and set sample size to 100, stopping criterion to $10e^{-10}$ and the maximum number of iterations to 5000.

Procedure 6.2.2 (**Pilot Study Procedure for MSM**).

1: *Create a random sample with sample size* 100 *according to null model (6.5);*

2: *Obtain* 19 *starting values* p_0 *for mixture proportion* p *as* $0.05, 0.10, \ldots, 0.95$. *For every* p_0 *created in this step, calculate the starting values for the remaining parameters according to procedure 4.4.1;*

3: *Obtain* 250 *random starting values* p_0 *for mixture proportion* p *by sampling from a uniform* $U(0,1)$ *distribution. For every* p_0 *created in this step, randomly split the sample into* 2 *groups according to this* p_0*, calculate the starting values for the remaining parameters accordingly;*

4: *Carry out EM algorithm on each starting point, and save final results* $\hat{p}$, $\hat{\alpha}$, $\hat{\gamma}$, $\hat{\beta}$, $\hat{\delta}$ *and* $\hat{\sigma}$. ■

Among the 25 simulations, there are 16 simulations with Pattern 1; 9 simulations with Pattern 2; and no simulations with Pattern 3.

Parallel to the study of the normal mixture model, for these same 25 random samples, we also use $p_0 \in \{0.05, 0.10, \ldots, 0.95\}$ and several combinations of various stopping criteria and maximum number of iterations carry out EM optimization. Results from these simulations are listed in Appendix A. By comparing the LRTS results obtained from the stringent simulation condition (250 RSPs, stopping criterion $10e^{-10}$ and maximum number of iterations 5000) and the most relaxed simulation condition (19 fixed starting points, stopping criterion $10e^{-4}$ and maximum number of iterations 1000), we find the largest LRTS difference is 0.0028.

6.2.3 Conclusion from the Pilot Studies

From our pilot studies we can conclude that:

- The difficulty for finding global maximum increases with sample size. Setting more stringent stopping criteria and increasing the maximum number of iterations seems to increase the chance of finding global maximum, especially for those cases with larger sample size.

- The difficulty of finding the global maximum also depends on the true LRTS. The EM procedure seems to have a high probability of finding global maximum for those samples with large LRTS. Therefore, the simulated null LRTS distribution have high accuracy for larger critical values.

- Running the EM procedure multiple times with fixed starting point of $p_0 \in \{0.05, 0.10, \ldots, 0.95\}$ is a efficient and reliable strategy to simulate the distribution of LRTS at high percentile points.

Based on the considerations of computation time and reliability for obtaining accuracy LRTS estimation, we will set the stopping criterion to 10^{-4} and the maximum number of iterations to 1000. For some large sample cases, we will use more stringent choices and will mention those conditions specifically.

6.3 Null LRTS Distributions

As mentioned in the literature review, simulation is needed to estimate the null LRTS distributions for detecting mixture models. We carry out extensive simulations here to investigate the null LRTS distributions for all models listed in section 6.1. Intuitively, for every homogenous linear model with predetermined structure, changing regression parameters α, γ, β and standard deviation σ only changes the orientation and dispersion of the whole distribution of the sample data but not the geometric relationship among observations. Therefore, null distributions should not change much under different parameter values. To verify this for our mixture models, we carried out several simulation studies using different parameter settings and sample size 200 (data not shown). The null distributions do not appear to change with different parameter values. Accordingly, we set α, γ, β and σ all equal 1 for all the linear mixture models.

For every mixture model, using the EM implementation strategy listed in section 6.2.3, we ran 1000 simulations for sample sizes 100, 200, 400, 800, 1600 to obtain empirical null LRTS distributions for each model listed in section 6.1. For NMM, we also ran 240 simulations for sample size 5000 (use stopping criterion $10e^{-6}$ and maximum number of iterations

1000) to obtain its empirical null LRTS distribution.

Table 6.1: Percentages of LRTS whose value are less than -0.01 in simulated null LRTS distributions for normal mixture model (NMM), mixture intercept model (MIM), mixture slope model (MSM) and unrestricted mixture linear regression model (UMLRM)

Sample Size	NMM	MIM	MSM	UMLRM
100	0	0	0	0
200	0	0	0	0
400	0	0	0	0
800	0	0.1	0	0
1600	0	0.1	1.4	0

Table 6.2: Percentages of LRTS whose absolute value are less than 0.01 in simulated null LRTS distributions for normal mixture model (NMM), mixture intercept model (MIM), mixture slope model (MSM) and unrestricted mixture linear regression model (UMLRM)

Sample Size	NMM	MIM	MSM	UMLRM
100	0.4	0.5	31.7	0
200	1.0	0.6	31.7	0
400	1.8	0.8	28.0	0
800	1.4	1.1	30.9	0
1600	0.7	1.0	33.8	0

Table 6.1 shows the percentages of LRTS that are less than -0.01 in every simulated null LRTS distribution. Table 6.2 shows the percentages of LRTS whose absolute values are less than 0.01. Taking round-off errors and the limited numbers of EM iterations into consideration, we conclude that:

- All the null LRTS are nonnegative;
- The values of the null LRTS for NMM, MIM and UMLRM are largely > 0.01. There are around 30% of values for MSM ≤ 0.01.

The empirical null LRTS distributions for different models with various sample sizes are shown in figure 6.1, figure 6.2, figure 6.3, figure 6.4. Figure 6.5 shows the empirical null LRTS distributions of various models with sample size 1600 in a single graph. Table 6.3, table 6.4, table 6.5, table 6.6 and table 6.7 provide the important percentiles, means and variances for these distributions. For the finite sample sizes used in our simulation, we find that the changes among null LRTS distributions for the same model with different sample sizes do not have a predictable trend. Since the dependence of the null LRTS distribution on the specific sample size is not clear, we use the null LRTS distribution which is obtained with the same sample size as the target data set when we study the power to detect mixture models.

For NMM, we can not detect the $\log\log(n)$ rate of divergence of LRTS with sample size 5000, which agrees the finding of Liu and Shao [18]. In our simulation, the null distribution for sample size 5000 has the largest 2.5% and 5.0% percentiles, but smallest 95.0% and 97.5% ones among all the NMM null distributions. It might not be reliable since we only use 240 replicates in sample size 5000 case, but it may be worthy of further study.

Table 6.3: Percentiles, means and variances for empirical null LRTS distributions of normal mixture model (NMM) with different sample sizes

Percentile	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 1600$	$n = 5000$
2.5%	0.0611	0.0468	0.0325	0.0253	0.0514	0.1032
5.0%	0.1041	0.1139	0.1620	0.1332	0.0922	0.1646
25%	0.6587	0.5906	0.6760	0.6118	0.6058	0.7212
50%	1.4352	1.4401	1.5219	1.4783	1.4099	1.3894
75%	2.6693	2.8424	2.9483	2.8473	3.0368	2.6493
95%	6.5039	5.8002	6.2598	5.9761	5.9913	5.0379
97.5%	8.0411	7.0062	7.4161	7.7415	7.4168	6.1532
Mean	2.0996	2.0128	2.1521	2.0769	2.0657	1.8946
Variance	4.6722	3.6299	4.4795	4.1460	4.0968	2.7603

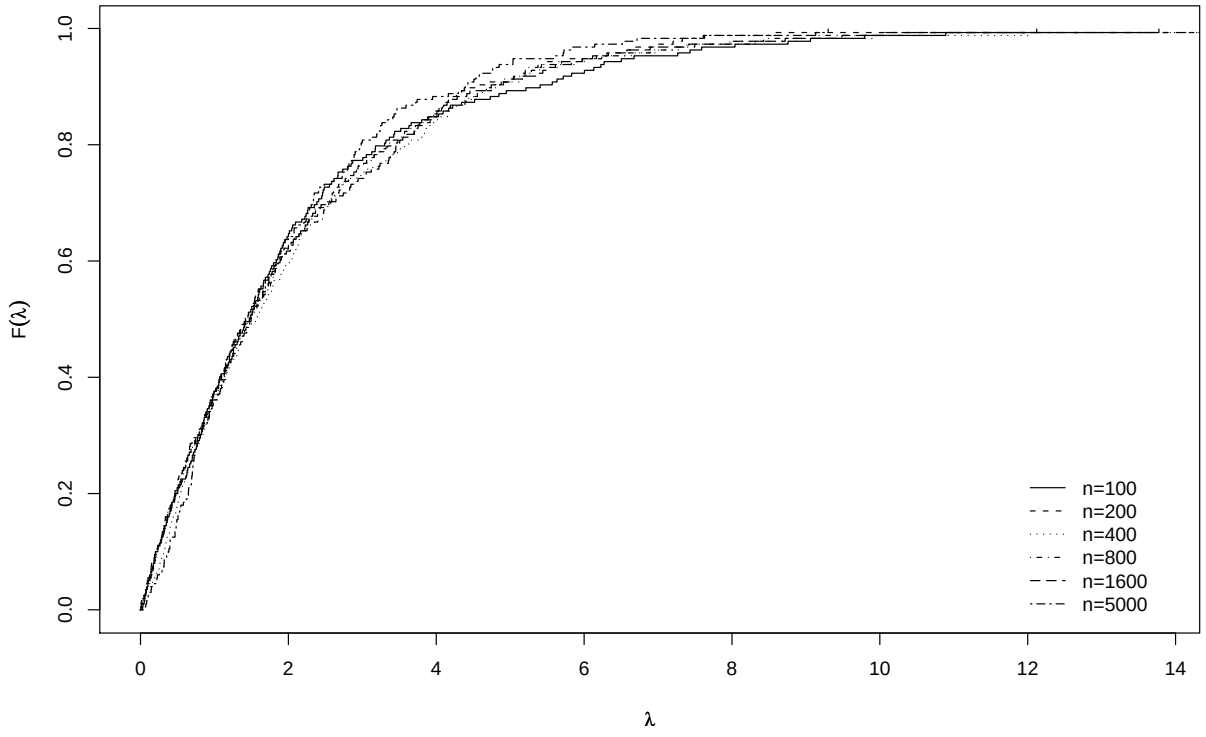

Figure 6.1: Null LRTS distribution for normal mixture model (NMM) with different sample sizes

Table 6.4: Percentiles, means and variances for empirical null LRTS distributions of mixture intercept model (MIM) with different sample sizes

Percentile	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 1600$
2.5%	0.0620	0.0402	0.0500	0.0297	0.0352
5.0%	0.1345	0.1106	0.1522	0.1076	0.0881
25%	0.6911	0.5858	0.7072	0.6025	0.5701
50%	1.5816	1.4769	1.5774	1.5030	1.3783
75%	3.1635	2.9690	3.0566	2.8926	2.9930
95%	7.0375	6.4751	6.1997	6.0230	6.0419
97.5%	8.7938	7.3317	7.9590	7.7486	7.5256
Mean	2.3123	2.1075	2.2048	2.0897	2.0524
Variance	5.3395	4.1771	4.7504	4.2637	4.1338

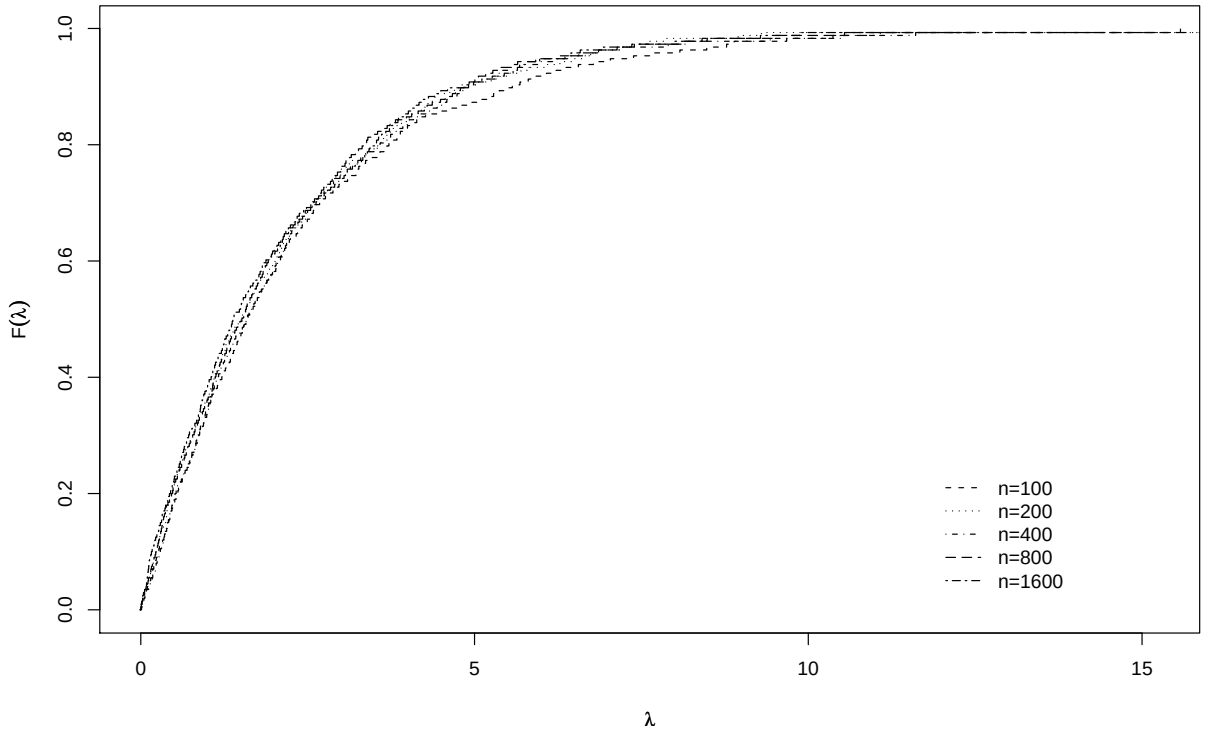

Figure 6.2: Null LRTS distribution for mixture intercept model (MIM) with different sample sizes

Table 6.5: Percentiles, means and variances for empirical null LRTS distributions of mixture slope model (MSM) with different sample sizes

Percentile	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 1600$
2.5%	-0.0019	-0.0030	-0.0042	-0.0054	-0.0086
5.0%	-0.0014	-0.0022	-0.0029	-0.0042	-0.0074
25%	-0.0003	-0.0005	-0.0005	-0.0011	-0.0021
50%	0.4866	0.5108	0.5414	0.4655	0.3380
75%	2.0499	1.8536	1.8808	1.7776	1.6134
95%	5.1428	4.7321	5.0884	5.4740	5.1524
97.5%	6.5122	5.9085	6.9499	7.0455	6.6229
Mean	1.3547	1.2564	1.2929	1.2698	1.1688
Variance	3.9367	3.1974	3.4990	3.4817	3.4178

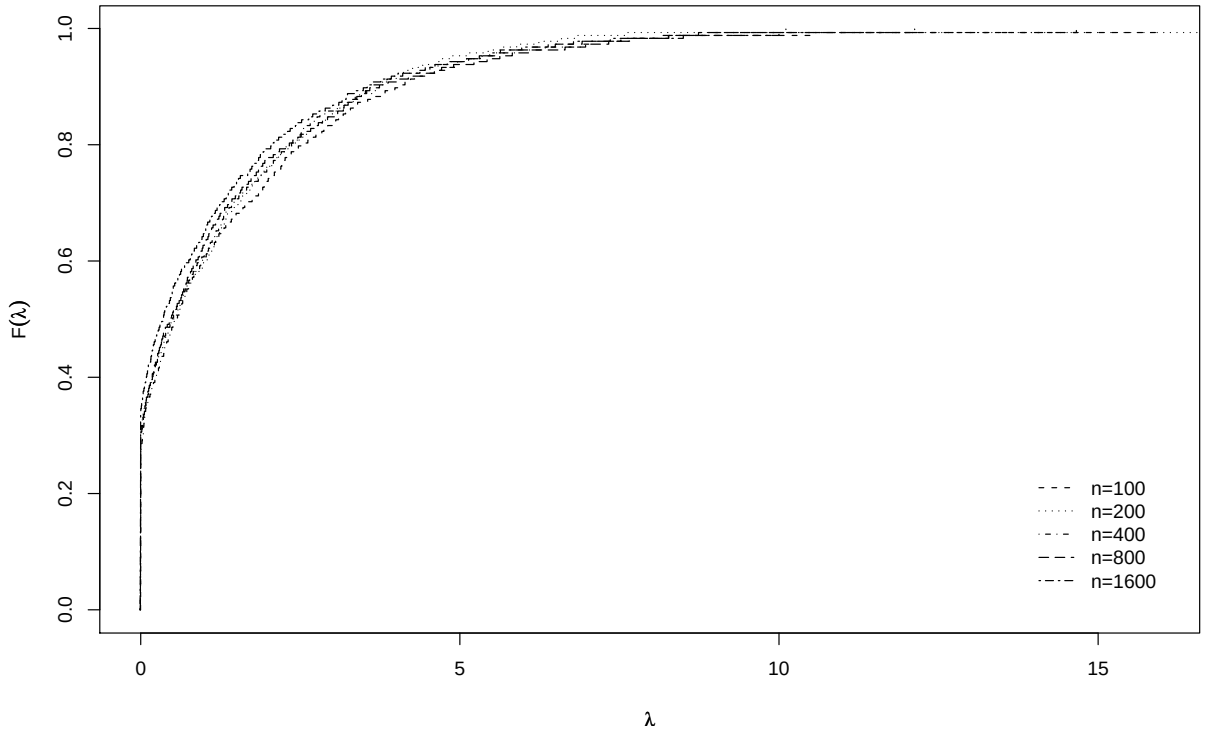

Figure 6.3: Null LRTS distribution for mixture slope model (MSM) with different sample sizes

Table 6.6: Percentiles, means and variances for empirical null LRTS distributions of unrestricted mixture linear regression model (UMLRM) with different sample sizes

Percentile	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 1600$
2.5%	1.4634	1.4479	1.4854	1.2264	1.2741
5.0%	1.9008	1.7665	1.7841	1.7007	1.5827
25%	3.5425	3.3675	3.3295	3.4014	3.3003
50%	5.2796	5.0736	5.0061	5.0934	4.9240
75%	7.8496	7.3431	7.1019	7.1536	7.0774
95%	12.6914	11.2815	11.4323	10.7234	11.4324
97.5%	14.6111	12.6344	13.1646	12.1602	12.5947
Mean	6.0735	5.6326	5.5775	5.5054	5.4992
Variance	12.1279	9.4450	9.2758	7.9565	9.2351

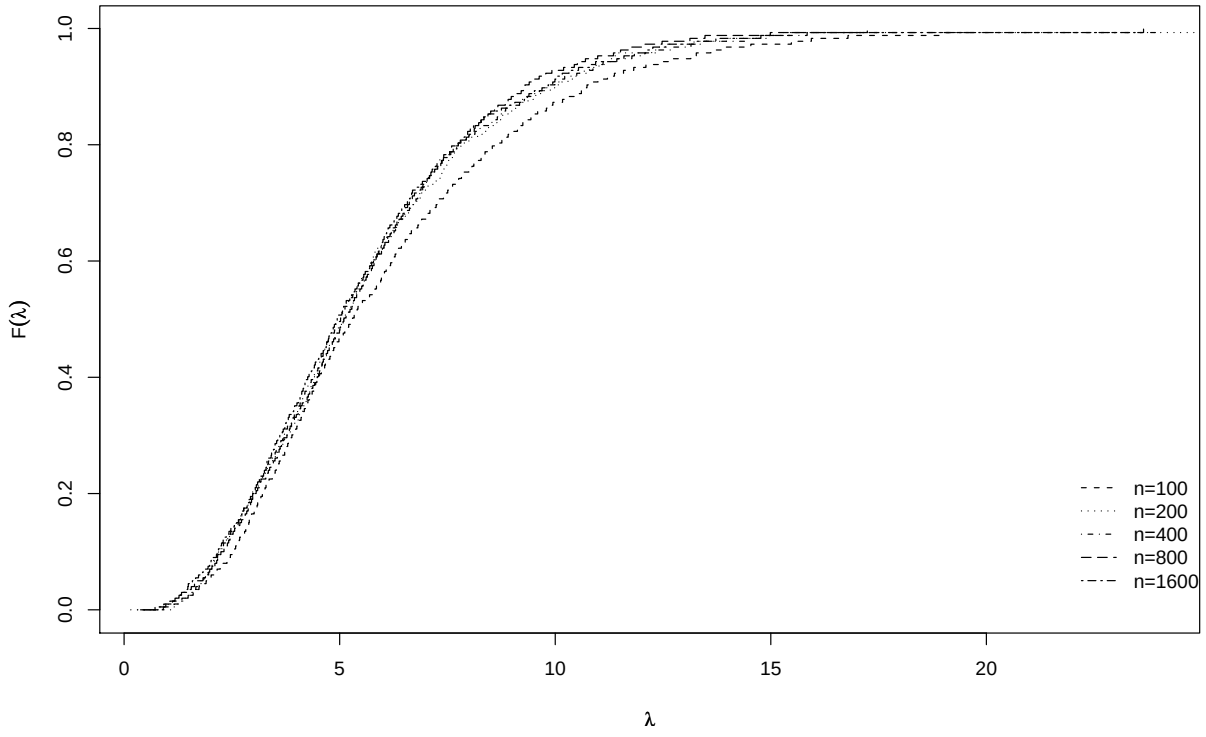

Figure 6.4: Null LRTS distribution for unrestricted mixture linear regression model (UMLRM) with different sample sizes

Table 6.7: Percentiles, means and variances for the empirical null LRTS distributions for different mixture models with sample size 1600

Percentile	MSM	NMM	MIM	UMLRM
2.5%	-0.0086	0.0514	0.0352	1.2741
5.0%	-0.0074	0.0922	0.0881	1.5827
25%	-0.0021	0.6058	0.5701	3.3003
50%	0.3380	1.4099	1.3783	4.9240
75%	1.6134	3.0368	2.9930	7.0774
95%	5.1524	5.9913	6.0419	11.4324
97.5%	6.6229	7.4168	7.5256	12.5947
Mean	1.1688	2.0657	2.0524	5.4992
Variance	3.4178	4.0968	4.1338	9.2351

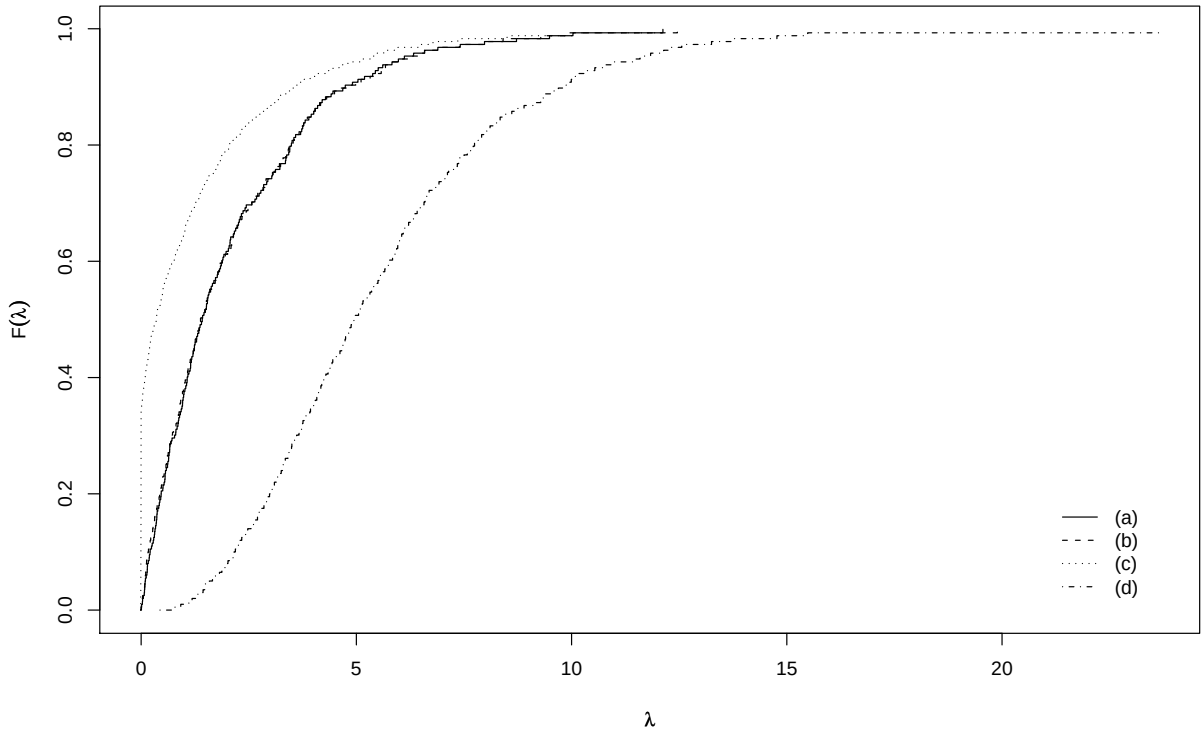

Figure 6.5: Null LRTS distributions for 4 models with sample size 1600. (a) normal mixture model (NMM), (b) mixture intercept model (MIM), (c) mixture slope model (MSM), (d) unrestricted mixture linear regression model (UMLRM)

6.4 Power to Detect Mixture Intercept Models

Consider the following mixture intercept model:

$$y \sim \begin{cases} \alpha + \gamma x_t + \beta x_c + \varepsilon & \text{with probability } p, \\ (\alpha + \delta) + \gamma x_t + \beta x_c + \varepsilon & \text{with probability } 1 - p, \end{cases} \tag{6.6}$$

$$x_t \overset{iid}{\sim} U(-10, 10),$$
$$x_c \overset{iid}{\sim} Bernoulli(0.5),$$
$$\varepsilon \overset{iid}{\sim} N(0, \sigma^2).$$

This model is equivalent to the mixture error term model which is described in Section 4.3. We study the power to detect this model under the following parameter settings:

$$\alpha = 1, \quad \beta = 0.7, \quad \sigma = 1,$$
$$p \in \{0.15, 0.5\},$$
$$\gamma \in \{0.125, 0.25, 0.5, 1\},$$
$$\delta \in \{1, 2, 3, 4\},$$
$$n \in \{200, 400, 800, 1600\},$$

where n is the sample size.

The three mixture intercept detection methods used here are:

Normal mixture model (NMM): First use ordinary least squares (OLS) to fit the data. Then test whether the residuals are mixture of two normal distributions.

Mixture intercept model (MIM): Use MIM to detect whether data come from a mix-

ture model with two mixture intercepts;

Unrestricted mixture linear regression model (UMLRM): Use UMLRM to detect whether data come from mixture of two linear models.

In this simulation study, for every parameter setting listed in the result tables, we first randomly create 200 samples according to the mixture intercept model. Then for every sample, we use our EM implementation strategy to perform mixture detection with the three mixture intercept detection methods. All the power results with significance level $\alpha = 0.05$ are listed in table 6.8, table 6.10 and table 6.12. The results with significance level $\alpha = 0.01$ are listed in table 6.9, table 6.11 and table 6.13. In some of these tables, we have a column labeled as 'Power Estimated from Formula (5.20)'. In those columns, we list powers that are estimated from Formula (5.20). Note that the estimated powers are close to simulated ones in the majority of cases with separation $\delta \geq 2$.

Table 6.14, table 6.15, table 6.16 and table 6.17 show the power comparisons under every simulation condition by using McNemar's Test. With those tables, we conclude:

- For sample size 200, NMM usually outperforms MIM. When separation $\delta = 2$, most p-values from McNemar's test are significant.
- For sample size 400, MIM sometimes outperforms NMM, and there are only 3 cases in which MIM significantly outperforms NMM.
- For sample sizes 800 and 1600, the difference of mixture detecting power between NMM and MIM is minor.
- MIM outperforms UMLRM in most simulations, and in almost half times, MIM significantly outperforms UMLRM. In only one case with sample size $n = 800$, mixture

proportion $p = 0.5$, separation $\delta = 1$ and slope $\gamma = 1$, UMLRM significantly outperforms MIM.

- Based on the preceding comparisons, NMM seems to be a good choice to detect mixture slopes.

- For every method, power is a function of the separation δ, mixture proportion p and sample size n but is not sensitive to the regression parameter γ.

- For every method, power changes with p. The power when $p = .15$ is generally greater than those of $p = .50$ when both powers are in the range from 30% to 100%. This trend is also predicted by our power estimation formula (5.20).

Table 6.8: Power (in percent) of log likelihood ratio test to detect mixture intercept by normal mixture model (NMM) method at .05 level of significance based on 200 replicates for each alternative.

Sample			By Simulation				Power
			γ				Estimated from
Size	p	δ	0.125	0.25	0.5	1.0	Formula (5.20)
200	0.15	1	7.5	9.0	8.0	3.5	5.39e-06
		2	54.0	55.5	57.5	61.0	59.65
		3	100.0	100.0	99.0	100.0	99.02
	0.50	1	6.0	10.5	6.0	9.0	1.51e-22
		2	46.5	43.5	42.0	43.5	46.59
		3	98.5	100.0	99.0	99.5	98.97
400	0.15	1	4.5	7.0	9.0	5.5	6.11e-03
		2	82.5	83.5	81.5	83.0	83.80
		3	100.0	100.0	100.0	100.0	99.98
	0.50	1	5.0	7.0	4.5	7.0	8.46e-12
		2	72.5	70.0	71.0	70.0	76.14
		3	100.0	100.0	100.0	100.0	99.98
800	0.15	1	5.5	14.5	9.5	10.0	0.70
		2	99.5	99.0	99.5	99.0	97.22
	0.50	1	5.5	5.5	9.0	6.5	9.69e-05
		2	95.0	96.5	97.5	96.0	95.34
1600	0.15	1	13.5	13.5	13.5	13.5	6.70
		2	100.0	100.0	100.0	100.0	99.89
	0.50	1	6.5	6.5	6.5	6.5	0.15
		2	100.0	100.0	100.0	100.0	99.76

Table 6.9: Power (in percent) of log likelihood ratio test to detect mixture intercept by normal mixture model (NMM) method at .01 level of significance based on 200 replicates for each alternative.

Sample			By Simulation				Power
			γ				Estimated from
Size	p	δ	0.125	0.25	0.5	1.0	Formula (5.20)
200	0.15	1	4.5	3.5	2.5	1.5	8.97e-15
		2	34.0	41.0	40.5	42.0	39.53
		3	99.5	99.5	99.0	99.0	98.21
	0.50	1	2.5	5.5	2.5	1.5	5.10e-54
		2	30.0	30.0	28.0	27.5	22.73
		3	97.5	99.5	98.5	97.5	97.99
400	0.15	1	1.0	0.5	1.0	0.0	8.97e-12
		2	60.5	58.5	49.5	58.0	64.50
		3	100.0	100.0	100.0	100.0	99.95
	0.50	1	1.0	0.5	0.0	0.5	2.53e-43
		2	40.0	41.0	38.0	32.0	46.53
		3	100.0	100.0	100.0	100.0	99.95
800	0.15	1	2.0	4.0	3.5	3.0	2.15e-04
		2	94.5	97.0	94.0	96.5	93.83
	0.50	1	2.0	1.0	3.0	1.0	5.15e-17
		2	84.5	88.5	90.0	89.5	88.39
1600	0.15	1	7.0	7.0	7.0	7.0	0.13
		2	100.0	100.0	100.0	100.0	99.74
	0.50	1	1.5	1.5	1.5	1.5	2.83e-07
		2	99.5	99.5	99.5	99.5	99.35

Table 6.10: Power (in percent) of log likelihood ratio test to detect mixture intercept by mixture intercept model (MIM) method at .05 level of significance based on 200 replicates for each alternative.

Sample			By Simulation				Power
			γ				Estimated from
Size	p	δ	0.125	0.25	0.5	1.0	Formula (5.20)
200	0.15	1	6.5	7.5	7.0	3.5	2.73e-07
		2	51.5	53.0	54.0	57.5	56.11
		3	100.0	100.0	99.5	99.5	98.91
	0.50	1	5.5	9.0	5.5	7.5	3.66e-27
		2	43.0	40.0	42.5	39.0	41.95
		3	98.5	100.0	99.5	99.5	98.84
400	0.15	1	4.5	7.5	10.0	6.5	3.90e-03
		2	83.5	85.0	82.0	83.5	83.33
		3	100.0	100.0	100.0	100.0	99.98
	0.50	1	5.5	7.0	4.5	7.5	1.75e-12
		2	75.5	71.0	72.5	72.5	75.37
		3	100.0	100.0	100.0	100.0	99.98
800	0.15	1	5.5	15.0	9.5	10.0	0.62
		2	99.5	99.0	99.5	99.0	97.17
	0.50	1	5.5	5.5	9.5	6.5	6.61e-05
		2	95.0	96.5	97.5	96.0	95.25
1600	0.15	1	13.5	13.5	13.5	13.5	6.34
		2	100.0	100.0	100.0	100.0	99.88
	0.50	1	6.5	6.5	6.5	6.5	0.13
		2	100.0	100.0	100.0	100.0	99.76

Table 6.11: Power (in percent) of log likelihood ratio test to detect mixture intercept by mixture intercept model (MIM) method at .01 level of significance based on 200 replicates for each alternative.

Sample			By Simulation				Power
			γ				Estimated from
Size	p	δ	0.125	0.25	0.5	1.0	Formula (5.20)
200	0.15	1	2.0	4.0	1.5	0.5	1.52e-17
		2	31.5	38.5	35.5	38.5	34.70
		3	99.5	99.0	99.0	99.0	97.93
	0.50	1	2.5	3.0	1.5	0.0	5.24e-64
		2	25.0	27.5	26.0	24.5	18.05
		3	97.5	99.0	98.5	97.5	97.65
400	0.15	1	1.5	0.5	1.0	0.0	5.91e-13
		2	61.5	59.0	51.0	59.5	62.17
		3	100.0	100.0	100.0	100.0	99.95
	0.50	1	1.0	0.5	0.0	0.5	1.42e-47
		2	40.0	41.5	40.0	32.0	43.35
		3	100.0	100.0	100.0	100.0	99.94
800	0.15	1	2.0	3.0	3.5	3.0	1.63e-04
		2	94.5	97.0	93.5	96.5	93.71
	0.50	1	2.0	1.0	2.5	0.5	1.95e-17
		2	85.0	87.5	89.5	89.5	88.14
1600	0.15	1	7.0	7.0	7.0	7.0	0.18
		2	100.0	100.0	100.0	100.0	99.75
	0.50	1	1.5	1.5	1.5	1.5	7.79e-07
		2	100.0	100.0	100.0	100.0	99.38

Table 6.12: Power (in percent) of log likelihood ratio test to detect mixture intercept by unrestricted mixture linear regression model (UMLRM) method at .05 level of significance based on 200 replicates for each alternative.

Sample			γ			
Size	p	δ	0.125	0.25	0.5	1.0
200	0.15	1	6.5	7.5	3.5	3.0
		2	35.5	37.5	36.5	43.0
		3	99.5	99.0	99.0	98.0
	0.50	1	5.0	6.0	4.0	4.5
		2	24.5	28.5	25.5	25.5
		3	97.5	98.5	98.0	97.0
400	0.15	1	7.0	7.0	5.0	6.0
		2	70.0	70.0	65.0	65.5
		3	100.0	100.0	100.0	100.0
	0.50	1	3.5	2.0	3.0	5.5
		2	52.5	48.0	50.0	44.0
		3	100.0	100.0	100.0	100.0
800	0.15	1	9.0	11.0	9.5	8.0
		2	95.5	97.0	93.5	97.0
	0.50	1	5.5	5.0	5.5	4.5
		2	87.0	88.5	91.5	90.0
1600	0.15	1	10.5	10.5	10.5	10.5
		2	100.0	100.0	100.0	100.0
	0.50	1	2.5	2.5	2.5	2.5
		2	99.5	99.5	99.5	99.5

Table 6.13: Power (in percent) of log likelihood ratio test to detect mixture intercept by unrestricted mixture linear regression model (UMLRM) method at .01 level of significance based on 200 replicates for each alternative.

Sample			γ			
Size	p	δ	0.125	0.25	0.5	1.0
200	0.15	1	1.5	3.5	0.0	0.5
		2	22.0	22.0	18.5	26.0
		3	96.0	96.0	95.0	93.5
	0.50	1	0.5	1.0	0.5	0.5
		2	11.0	10.0	12.5	12.5
		3	89.5	92.5	95.0	95.5
400	0.15	1	2.0	1.5	2.5	1.5
		2	48.5	49.5	39.0	48.5
		3	100.0	100.0	100.0	100.0
	0.50	1	1.5	0.5	0.5	2.0
		2	29.0	29.5	33.0	26.5
		3	100.0	100.0	100.0	100.0
800	0.15	1	3.0	4.0	5.0	3.0
		2	93.0	94.0	87.5	94.5
	0.50	1	2.0	0.0	0.5	3.0
		2	73.5	79.0	81.0	82.0
1600	0.15	1	6.0	6.0	6.0	6.0
		2	99.5	99.5	99.5	99.5
	0.50	1	2.0	2.0	2.0	2.0
		2	99.5	99.5	99.5	99.5

Table 6.14: Power Comparison between mixture intercept model (MIM) method and normal mixture model (NMM) method for detecting mixture intercept at .05 level of significance. The numbers in each cell represents (number of simulations that are detected by MIM but not NMM, number of simulations that are detected by NMM but not MIM, p-value of McNemar tset (minus p-value means NMM outperforms MIM)). The -,-,- represents 0, 0, 1, that is NMM and MIM give the same results.

Sample			γ			
Size	p	δ	0.125	0.25	0.5	1.0
200	0.15	1	0, 2, -0.16	0, 3, -0.08	1, 3, -0.32	-, -, -
		2	**0, 5, -0.03**	**0, 5, -0.03**	**0, 7, -0.01**	**1, 8, -0.02**
		3	-, -, -	-, -, -	1, 0, 0.32	0, 1, -0.32
	0.50	1	0, 1, -0.32	0, 3, -0.08	0, 1, -0.32	0, 3, -0.08
		2	**0, 7, -0.01**	**0, 7, -0.01**	4, 3, 0.71	**1, 10, -0.01**
		3	-, -, -	-, -, -	1, 0, 0.32	-, -, -
400	0.15	1	-, -, -	1, 0, 0.32	2, 0, 0.16	2, 0, 0.16
		2	2, 0, 0.16	3, 0, 0.08	1, 0, 0.32	1, 0, 0.32
		3	-, -, -	-, -, -	-, -, -	-, -, -
	0.50	1	1, 0, 0.32	-, -, -	-, -, -	1, 0, 0.32
		2	**6, 0, 0.01**	2, 0, 0.16	3, 0, 0.08	**5, 0, 0.03**
		3	-, -, -	-, -, -	-, -, -	-, -, -
800	0.15	1	-, -, -	1, 0, 0.32	-, -, -	-, -, -
		2	-, -, -	-, -, -	-, -, -	-, -, -
	0.50	1	-, -, -	-, -, -	1, 0, 0.32	-, -, -
		2	-, -, -	-, -, -	-, -, -	-, -, -
1600	0.15	1	-, -, -	-, -, -	-, -, -	-, -, -
		2	-, -, -	-, -, -	-, -, -	-, -, -
	0.50	1	-, -, -	-, -, -	-, -, -	-, -, -
		2	-, -, -	-, -, -	-, -, -	-, -, -

Table 6.15: Power Comparison between mixture intercept model (MIM) method and normal mixture model (NMM) method for detecting mixture intercept at .01 level of significance. The numbers in each cell represents (number of simulations that are detected by MIM but not NMM, number of simulations that are detected by NMM but not MIM, p-value of McNemar tset (minus p-value means NMM outperforms MIM)). The -,-,- represents 0, 0, 1, that is NMM and MIM give the same results.

Sample			γ			
Size	p	δ	0.125	0.25	0.5	1.0
200	0.15	1	**0, 5, -0.03**	1, 0, 0.32	0, 2, -0.16	0, 2, -0.16
		2	**0, 5, -0.03**	**0, 5, -0.03**	**0, 10, -0.0**	**0, 7, -0.01**
		3	-, -, -	0, 1, -0.32	-, -, -	-, -, -
	0.50	1	-, -, -	**0, 5, -0.03**	0, 2, -0.16	0, 3, -0.08
		2	**2, 12, -0.01**	1, 6, -0.06	1, 5, -0.1	**0, 6, -0.01**
		3	-, -, -	0, 1, -0.32	-, -, -	1, 1, 1
400	0.15	1	1, 0, 0.32	-, -, -	-, -, -	-, -, -
		2	2, 0, 0.16	1, 0, 0.32	4, 1, 0.18	3, 0, 0.08
	0.50	1	-, -, -	-, -, -	-, -, -	-, -, -
		2	1, 1, 1	1, 0, 0.32	**4, 0, 0.05**	1, 1, 1
800	0.15	1	-, -, -	0, 2, -0.16	-, -, -	-, -, -
		2	-, -, -	-, -, -	0, 1, -0.32	-, -, -
	0.50	1	-, -, -	-, -, -	0, 1, -0.32	0, 1, -0.32
		2	1, 0, 0.32	0, 2, -0.16	0, 1, -0.32	-, -, -
1600	0.15	1	-, -, -	-, -, -	-, -, -	-, -, -
		2	-, -, -	-, -, -	-, -, -	-, -, -
	0.50	1	-, -, -	-, -, -	-, -, -	-, -, -
		2	1, 0, 0.32	1, 0, 0.32	1, 0, 0.32	1, 0, 0.32

Table 6.16: Power Comparison between mixture intercept model (MIM) method and unrestricted mixture linear regression model (UMLRM) method for detecting mixture intercept at .05 level of significance. The numbers in each cell represents (number of simulations that are detected by MIM but not UMLRM, number of simulations that are detected by UMLRM but not MIM, p-value of McNemar tset (minus p-value means UMLRM outperforms MIM)). The -,-,- represents 0, 0, 1, that is MIM and UMLRM give the same results.

Sample			γ			
Size	p	δ	0.125	0.25	0.5	1.0
200	0.15	1	7, 7, 1	8, 8, 1	12, 5, 0.09	6, 5 , 0.76
		2	**35, 3, 0**	**38, 7, 0**	**40, 5, 0**	**34, 5 , 0**
		3	1, 0, 0.32	2, 0, 0.16	1, 0, 0.32	4, 1 , 0.18
	0.50	1	7, 6, 0.78	10, 4, 0.11	8, 5, 0.41	14, 8 , 0.2
		2	**40, 3, 0**	**31, 8, 0**	**48, 4, 0**	**30, 3 , 0**
		3	2, 0, 0.16	3, 0, 0.08	4, 1, 0.18	**5, 0 , 0.03**
400	0.15	1	2, 7, -0.1	9, 8, 0.81	**15, 5, 0.03**	13, 12, 0.84
		2	**31, 4, 0**	**33, 3, 0**	**40, 6, 0**	**38, 2 , 0**
	0.50	1	8, 4, 0.25	**12, 2, 0.01**	7, 4, 0.37	10, 6 , 0.32
		2	**48, 2, 0**	**47, 1, 0**	**47, 2, 0**	**58, 1 , 0**
800	0.15	1	6, 13, -0.11	20, 12, 0.16	8, 8, 1	12, 8, 0.37
		2	**8, 0, 0**	**4, 0, 0.05**	**12, 0, 0**	**4, 0 , 0.05**
	0.50	1	5, 5, 1	5, 4, 0.74	13, 5, 0.06	8, 4 , 0.25
		2	**16, 0, 0**	**16, 0, 0**	**12, 0, 0**	**12, 0 , 0**
1600	0.15	1	11, 5, 0.13	11, 5, 0.13	11, 5, 0.13	11, 5, 0.13
		2	-, -, -	-, -, -	-, -, -	-, -, -
	0.50	1	**9, 1, 0.01**	**9, 1, 0.01**	**9, 1, 0.01**	**9, 1, 0.01**
		2	1, 0, 0.32	1, 0, 0.32	1, 0, 0.32	1, 0, 0.32

Table 6.17: Power Comparison between mixture intercept model (MIM) method and unrestricted mixture linear regression model (UMLRM) method for detecting mixture intercept at .01 level of significance. The numbers in each cell represents (number of simulations that are detected by MIM but not UMLRM, number of simulations that are detected by UMLRM but not MIM, p-value of McNemar tset (minus p-value means UMLRM outperforms MIM)). The -,-,- represents 0, 0, 1, that is MIM and UMLRM give the same results.

Sample			γ			
Size	p	δ	0.125	0.25	0.5	1.0
200	0.15	1	3, 2, 0.65	4, 3, 0.71	3, 0, 0.08	1, 1, 1
		2	**22, 3, 0**	**36, 3, 0**	**36, 2, 0**	**32, 7, 0**
		3	**8, 1, 0.02**	**6, 0, 0.01**	**8, 0, 0**	**11, 0, 0**
	0.50	1	5, 1, 0.1	6, 2, 0.16	3, 1, 0.32	0, 1, -0.32
		2	**30, 2, 0**	**38, 3, 0**	**31, 4, 0**	**26, 2, 0**
		3	**16, 0, 0**	**13, 0, 0**	**7 , 0, 0.01**	**4 , 0, 0.05**
400	0.15	1	2, 3, -0.65	1, 3, -0.32	1, 4, -0.18	0, 3, -0.08
		2	**28, 2, 0**	**22, 3, 0**	**31, 7, 0**	**22, 0, 0**
	0.50	1	1, 2, -0.56	1, 1, 1	0, 1, -0.32	0, 3, -0.08
		2	**28, 6, 0**	**25, 1, 0**	**20, 6, 0.01**	**16, 5, 0.02**
800	0.15	1	2, 4, -0.41	3, 5, -0.48	3, 6, -0.32	1, 1, 1
		2	4, 1, 0.18	**6, 0, 0.01**	**12, 0, 0**	5 , 1, 0.1
	0.50	1	2, 2, 1	2, 0, 0.16	**4, 0, 0.03**	**0, 5, -0.03**
		2	**25, 2, 0**	**18, 1, 0**	**18, 1, 0**	**16, 1, 0**
1600	0.15	1	4, 2, 0.41	4, 2, 0.41	4, 2, 0.41	4, 2, 0.41
		2	1, 0, 0.32	1, 0, 0.32	1, 0, 0.32	1, 0, 0.32
	0.50	1	1, 2, -0.56	1, 2, -0.56	1, 2, -0.56	1, 2, -0.56
		2	1, 0, 0.32	1, 0, 0.32	1, 0, 0.32	1, 0, 0.32

6.5 Power to Detect Mixture Slope Models

Consider the following mixture slope model (MSM):

$$y \sim \begin{cases} \alpha + \gamma x_t + \beta x_c + \varepsilon & \text{with probability } p, \\ \alpha + (\gamma + \delta) x_t + \beta x_c + \varepsilon & \text{with probability } 1 - p, \end{cases} \tag{6.7}$$

$$x_t \overset{iid}{\sim} U(0, 10),$$
$$x_c \overset{iid}{\sim} Bernoulli(0.5),$$
$$\varepsilon \overset{iid}{\sim} N(0, \sigma^2).$$

We study the power to detect the mixture slope for this model under the following parameter settings:

$$\alpha = 1, \quad \gamma = 1, \quad \sigma = 1,$$
$$p \in \{0.15, 0.5\},$$
$$\beta \in \{0.25, 0.5, 1\},$$
$$\delta \in \{0.10, 0.20\},$$
$$n \in \{200, 400, 800, 1600\}.$$

We use the following two methods to detect mixture slope models:

Mixture Slope Model (MSM) Use MSM to detect whether data come from a mixture model with two different slopes.

Unrestricted Mixture Linear Regression Model (UMLRM) Use UMLRM to detect whether data come from mixture of two linear models.

For this simulation study, we follow the same approach as in the power study for detecting mixture intercept models. The power results for significance level $\alpha = 0.05$ are listed in table 6.18 and table 6.20. The results for significance level $\alpha = 0.01$ are listed in table 6.19 and table 6.21. Table 6.22 and table 6.23 show the power comparisons for every simulation condition using McNemar's Test. With these tables, we conclude:

- MSM significantly outperforms UMLRM in most circumstances. That is, MSM is a more powerful method than UMLRM for detecting a mixture slope model. The power gains from MSM over UMLRM happen when the powers of MSM are around 20% - 90%, and typical power gains are over 10%.

- For both methods, power is a function of separation δ, mixture proportion p and sample size n, but is insensitive to the regression parameter γ.

Table 6.18: Power (in percent) of log likelihood ratio test to detect mixture slope by mixture slope model (MSM) method at .05 level of significance based on 200 replicates for each alternative.

Sample			By Simulation			Power Estimated
Size	p	δ	$\gamma = 0.125$	$\gamma = 0.5$	$\gamma = 1.0$	from Formula 5.20
200	0.15	0.10	6.5	11.5	11.0	2.84e-03
		0.20	41.5	39.0	38.5	39.20
	0.50	0.10	16.0	15.0	14.0	1.11
		0.20	58.0	65.0	67.5	63.95
400	0.15	0.10	10.0	7.5	10.5	0.18
		0.20	52.5	63.0	56.5	60.40
	0.50	0.10	16.0	15.5	18.0	6.31
		0.20	93.0	89.0	87.0	84.58
800	0.15	0.10	10.5	10.5	9.0	2.28
		0.20	81.5	82.5	80.0	80.66
	0.50	0.10	24.0	23.0	22.0	19.15
		0.20	99.5	99.5	99.5	96.63
1600	0.15	0.10	21.0	21.0	21.0	15.16
		0.20	99.0	99.0	99.0	94.98
	0.50	0.10	46.5	46.5	46.5	44.63
		0.20	100.0	100.0	100.0	99.82

Table 6.19: Power (in percent) of log likelihood ratio test to detect mixture slope by mixture slope model (MSM) method at .01 level of significance based on 200 replicates for each alternative.

Sample			By Simulation			Power Estimated
Size	p	δ	$\gamma = 0.125$	$\gamma = 0.5$	$\gamma = 1.0$	from Formula 5.20
200	0.15	0.10	2.5	5.5	3.5	1.05e-10
		0.20	24.5	24.0	25.5	16.47
	0.50	0.10	6.0	6.5	6.0	1.86e-03
		0.20	39.5	44.0	47.0	42.25
400	0.15	0.10	3.0	1.5	3.0	2.53e-07
		0.20	32.0	36.0	30.0	33.61
	0.50	0.10	4.0	6.0	6.5	4.37e-02
		0.20	75.5	71.0	67.5	68.34
800	0.15	0.10	2.5	3.0	2.0	4.73e-03
		0.20	67.5	68.5	66.0	66.29
	0.50	0.10	10.0	12.0	9.5	2.08
		0.20	96.0	98.0	95.5	93.03
1600	0.15	0.10	8.0	8.0	8.0	0.46
		0.20	96.0	96.0	96.0	89.88
	0.50	0.10	24.0	24.0	24.0	13.56
		0.20	100.0	100.0	100.0	99.57

Table 6.20: Power (in percent) of log likelihood ratio test to detect mixture slope by unrestricted mixture linear regression model (UMLRM) method at .05 level of significance based on 200 replicates for each alternative.

Sample			γ		
Size	p	δ	0.125	0.5	1.0
200	0.15	0.10	6.0	4.0	4.0
		0.20	16.0	17.0	19.5
	0.50	0.10	4.5	8.0	6.0
		0.20	29.0	35.5	34.0
400	0.15	0.10	5.5	5.0	5.5
		0.20	29.0	34.5	28.0
	0.50	0.10	6.0	7.0	6.5
		0.20	70.5	69.0	65.0
800	0.15	0.10	5.0	7.0	4.0
		0.20	65.0	67.5	65.5
	0.50	0.10	12.0	14.0	10.5
		0.20	97.0	98.5	94.5
1600	0.15	0.10	9.0	9.0	9.0
		0.20	96.0	96.0	96.0
	0.50	0.10	24.0	24.0	24.0
		0.20	100.0	100.0	100.0

Table 6.21: Power (in percent) of log likelihood ratio test to detect mixture slope by unrestricted mixture linear regression model (UMLRM) method at .01 level of significance based on 200 replicates for each alternative.

Sample			γ		
Size	p	δ	0.125	0.5	1.0
200	0.15	0.10	2.5	1.0	1.0
		0.20	7.5	7.0	8.0
	0.50	0.10	0.5	1.5	2.5
		0.20	16.5	17.5	14.0
400	0.15	0.10	1.5	0.5	1.5
		0.20	17.5	18.0	15.0
	0.50	0.10	1.0	3.0	0.5
		0.20	46.0	48.0	45.5
800	0.15	0.10	2.5	3.5	1.5
		0.20	53.0	54.5	50.5
	0.50	0.10	6.0	6.0	2.5
		0.20	89.5	93.0	91.0
1600	0.15	0.10	3.5	3.5	3.5
		0.20	87.0	87.0	87.0
	0.50	0.10	12.0	12.0	12.0
		0.20	100.0	100.0	100.0

Table 6.22: Power Comparison between mixture slope model (MSM) method and unrestricted mixture linear regression model (UMLRM) method for detecting mixture slope at .05 level of significance. The numbers in each cell represents (number of simulations that are detected by MSM but not UMLRM, number of simulations that are detected by UMLRM but not MSM, p-value of McNemar tset (minus p-value means UMLRM outperforms MSM)). The -,-,- represents 0, 0, 1, that is MSM and UMLRM give the same results.

Sample			γ		
Size	p	δ	0.125	0.5	1.0
200	0.15	0.10	9, 8, 0.81	**21, 6, 0**	**19, 5, 0**
		0.20	**52, 1, 0**	**45, 1, 0**	**41, 3, 0**
	0.50	0.10	**27, 4, 0**	**24, 10, 0.02**	**22, 6, 0**
		0.20	**60, 2, 0**	**61, 2, 0**	**67, 0, 0**
400	0.15	0.10	**15, 6, 0.05**	13, 8, 0.28	**15, 5, 0.03**
		0.20	**52, 5, 0**	**60, 3, 0**	**60, 3, 0**
	0.50	0.10	**24, 4, 0**	**21, 4, 0**	**28, 5, 0**
		0.20	**45, 0, 0**	**40, 0, 0**	**45, 1, 0**
800	0.15	0.10	**14, 3, 0.01**	14, 7, 0.13	**13, 3, 0.01**
		0.20	**34, 1, 0**	**33, 3, 0**	**31, 2, 0**
	0.50	0.10	**28, 4, 0**	**26, 8, 0**	**23, 0, 0**
		0.20	**5, 0, 0.03**	2, 0, 0.16	**10, 0, 0**
1600	0.15	0.10	**29, 5, 0**	**29, 5, 0**	**29, 5, 0**
		0.20	**6, 0, 0.01**	**6, 0, 0.01**	**6, 0, 0.01**
	0.50	0.10	**46, 1, 0**	**46, 1, 0**	**46, 1, 0**
		0.20	-, -, -	-, -, -	-, -, -

Table 6.23: Power Comparison between mixture slope model (MSM) method and unrestricted mixture linear regression model (UMLRM) method for detecting mixture slope at .01 level of significance. The numbers in each cell represents (number of simulations that are detected by MSM but not UMLRM, number of simulations that are detected by UMLRM but not MSM, p-value of McNemar tset (minus p-value means UMLRM outperforms MSM)). The -,-,- represents 0, 0, 1, that is MSM and UMLRM give the same results.

Sample			γ		
Size	p	δ	0.125	0.5	1.0
200	0.15	0.10	3, 3, 1	**10, 1, 0.01**	6, 1, 0.06
		0.20	**35, 1, 0**	**34, 0, 0**	**36, 1, 0**
	0.50	0.10	**11, 0, 0**	**12, 2, 0.01**	**8, 1, 0.02**
		0.20	**46, 0, 0**	**55, 2, 0**	**66, 0, 0**
400	0.15	0.10	4, 1, 0.18	3, 1, 0.32	4, 1, 0.18
		0.20	**30, 1, 0**	**42, 6, 0**	**32, 2, 0**
	0.50	0.10	**7, 1, 0.03**	**7, 1, 0.03**	**12, 0, 0**
		0.20	**59, 0, 0**	**48, 2, 0**	**44, 0, 0**
800	0.15	0.10	1, 1, 1	4, 5, -0.74	2, 1, 0.56
		0.20	**31, 2, 0**	**30, 2, 0**	**35, 4, 0**
	0.50	0.10	**10, 2, 0.02**	**17, 5, 0.01**	**16, 2, 0**
		0.20	**14, 1, 0**	**10, 0, 0**	**10, 1, 0.01**
1600	0.15	0.10	**11, 2, 0.01**	**11, 2, 0.01**	**11, 2, 0.01**
		0.20	**18, 0, 0**	**18, 0, 0**	**18, 0, 0**
	0.50	0.10	**27, 3, 0**	**27, 3, 0**	**27, 3, 0**
		0.20	-, -, -	-, -, -	-, -, -

6.6 Simulation of Accuracy of Power and Sample Size Calculation Formulas

According to the power calculation formula (5.20) for normal mixture model (NMM), mixture intercept model (MIM) and mixture slope model (MSM), the power depends on the standard deviation σ, separation δ and mixture proportion p but is not sensitive to the regression parameters. This is confirmed by the simulations in the previous two sections. As shown in table 6.8, table 6.9, table 6.10, table 6.11, table 6.18 and table 6.19, the powers obtained by simulations are close to those estimated from formula (5.20) when $\delta \in \{2, 3\}$ for NMM and MIM and δ is .2 for MSM, especially when the significance level α is .05.

In order to document the accuracy of the sample size formula 5.22, we carry out three simulation studies for normal mixture model (NMM), mixture intercept model (MIM) and mixture slope model (MSM) according to the following procedure:

- For significance level $\alpha = .05$, power $= .80$ and $\sigma = 1$, calculate the necessary sample size for every combination of separation ($\delta \in \{1, 1.5, 2, 3, 4\}$ for NMM and MIM, $\delta \in \{.1, .15, .2, .3, .4\}$ for MSM) and mixture proportion ($p \in \{.1, .2, .3, .4, .5\}$) according to formula 5.22. The results are listed in table 6.24, table 6.25 and table 6.26.

- For each experiment setting in the three sample size tables (table 6.24, table 6.25 and table 6.26), run 1000 (200 for sample size over 1000) mixture detection simulations on randomly created samples with the corresponding sample size, and estimate the power from these simulation results. Then list all the simulated power results in table 6.27, table 6.28 and table 6.29. For MIM and MSM, we set all regression parameters as 1.0 except the separation δ.

In this study, we define the sample size estimation as successful if the simulated power

is between 75% and 85% (for target power 80%). From table 6.27, table 6.28 and table 6.29, we conclude that (under our design conditions):

- For normal mixture model (NMM) and mixture intercept model (MIM):
  - When the separation $\delta \geq 2$, the sample size formula 5.22 provides accurate sample size estimations for NMM and MIM in a broad range of mixture proportion p which includes $[0.1, 0.9]$.
  - When $1 \leq \delta \leq 2$, the sample size formula 5.22 provides accurate sample size estimations for NMM and MIM model only in two restricted ranges of mixture proportion p, which concentrate around 0.2 and 0.8, when the separation δ decreases the ranges shrink;
  - When the sample size formula 5.22 fails, 10 out of 11 sample size are underestimates of the correct sample size. The only overestimation happens in NMM when mixture proportion $p = 0.1$ and separation $\delta = 1$.
  - Our sample size estimations are close to the results from [29] when $\delta \in \{3, 4\}$, but are different when $\delta = 2$ and $p \in \{0.1, 0.2, 0.3, 0.4\}$. One possible reason for this is that we used $q_{1-\alpha}(n_1) \approx q_{1-\alpha}(n_2)$ in our derivation for sample size estimation formula 5.22. Another reason is that we might use different simulated null LRTS distributions. These findings are consistent with the approximate formula systematiccaly underestimating the correct n.
- For mixture slope model (MSM):
  - When the separation $\delta \geq 0.2$, the sample size formula 5.22 provides accurate sample size estimations for MSM in a broad range of mixture proportion p, which at least includes $[0.2, 0.8]$.

- When $0.15 \leq \delta \leq 0.2$, the sample size formula 5.22 provides accurate sample size estimations for MSM only in a restricted range of mixture proportion p, which includes $[0.4, 0.6]$.
- When the sample size formula 5.22 fails, it typically suggests a larger sample size than needed.

The above conclusions are only valid under the design condition examined. When we change the significance level α or other regression parameters, the results might change to some extents. This is especially true for the slope γ in MSM, which has some influence on simulated power in a few cases as shown in table 6.18 and table 6.19.

Table 6.24: Sample size table necessary to detect two-component normal mixture by using formula 5.22 under given design conditions (significance level $\alpha = .05$, target power = .80 and standard deviation $\sigma = 1$). The numbers in parentheses are the sample sizes published in [29].

Separation	Mixture Proportion p				
δ	.1	.2	.3	.4	.5
1	26915	9913	8934	9891	10421
1.5	1924	1212	1296	1591	1774
2	453 (578)	325 (459)	348 (446)	406 (452)	438 (456)
3	85 (100)	65 (76)	66 (73)	71 (73)	73 (73)
4	32 (38)	25 (28)	25 (26)	26 (26)	26 (26)

Table 6.25: Sample size table necessary to detect two-component mixture intercept by using formula 5.22 under given design conditions (significance level $\alpha = .05$, target power = .80 and standard deviation $\sigma = 1$).

Separation	Mixture Proportion p				
δ	.1	.2	.3	.4	.5
1	27133	10003	9019	9993	10535
1.5	1941	1223	1308	1606	1791
2	457	328	351	410	442
3	85	65	67	72	74
4	33	26	25	26	27

Table 6.26: Sample size table necessary to detect two-component mixture slope by using formula 5.22 under given design conditions (significance level $\alpha = .05$, target power = .80 and standard deviation $\sigma = 1$).

Separation	Mixture Proportion p				
δ	.1	.2	.3	.4	.5
.1	38577	13533	6886	7689	4720
.15	4515	1889	1188	1130	935
.2	1293	585	397	368	332
.3	273	136	101	93	89
.4	104	55	43	40	39

Table 6.27: Simulation results for power (in percent) for normal mixture detection using sample size specified in table 6.25 which has target power 80.

Separation	Mixture Proportion p				
δ	.1	.2	.3	.4	.5
1	98.0	**79.5**	57.0	40.5	25.0
1.5	**83.8**	**80.6**	**77.5**	68.1	67.8
2	**81.7**	**79.8**	**78.8**	**80.0**	**78.1**
3	**79.6**	**79.1**	**81.6**	**83.1**	**83.2**
4	**75.8**	**80.7**	**83.5**	**84.3**	**82.5**

Table 6.28: Simulation results for power (in percent) for mixture intercept detection using sample size specified in table 6.25 which has target power 80. (All the regression parameters are set as 1.0.)

Separation	Mixture Proportion p				
δ	.1	.2	.3	.4	.5
1	-	**76.0**	57.0	45.9	33.5
1.5	**79.0**	**84.0**	**79.0**	66.0	62.0
2	**81.7**	**82.2**	**80.3**	**79.8**	**75.3**
3	**81.9**	**81.4**	**81.6**	**84.4**	**83.0**
4	**76.9**	**81.6**	**83.2**	**84.4**	**84.6**

Table 6.29: Simulation results for power (in percent) for mixture slope detection using sample size specified in table 6.26 which has target power 80. (All other regression parameters are set as 1.0.)

Separation	Mixture Proportion p				
δ	.1	.2	.3	.4	.5
.1	-	-	-	97.0	88.5
.15	91.0	88.0	87.0	**84.0**	**85.0**
.2	85.7	**83.7**	**80.8**	**82.4**	**81.9**
.3	**81.5**	**81.9**	**80.4**	**82.4**	**81.8**
.4	**78.1**	**79.4**	**77.0**	**80.1**	**79.2**

Chapter 7

Application Study

In this chapter, we apply restricted mixture linear regression model (RMLRM) to two real data sets. Since the data sets are given, the main tasks here are parameter estimation and model selection. For model selection, in addition to using the LRTS based on our RMLRM, we also use the *Akaike's information criterion (AIC)* [1] and the *Bayesian information criterion (BIC)* [37].

7.1 Application to COGEND Data Set

Collaborative Genetic Study of Nicotine Dependence (COGEND) [3] is a retrospective study to find the biological mechanisms, genes and environmental features that determine nicotine consumption, and that predispose or protect individuals from the onset and persistence of the nicotine dependence. The current data provided to us from the COGEND group consist of 6429 individuals, with the following covariates:

- Gender;

- Years of education;
- Race (is black or not, defined as Black);
- Number of packs of cigarettes per day;
- Years between first puff and first regular smoking (defined as: YrOn);
- Age at first puff (defined as: Age0).

For variables YrOn and Age0, we defined the transformations as LogYrOn = log(YrOn+1.5) and LogAge0 = log(Age0).

For this data set, we try to find out whether age at first puff and race have some influence on how long it takes to become a regular smoker. If there is relationship between these variables, we also want to know whether this relationship changes among different groups of people and whether there is evidence of a mixture mechanism.

In [39], the ordinary linear regression model:

$$\text{LogYrOn} = \alpha + \gamma \cdot \text{LogAge0} + \beta \cdot \text{Black} + \varepsilon \tag{7.1}$$

was used to model this data set, and the OLS residuals were examined for a normal mixture. In addition to the OLS model (7.1), we use following four models to study this data set:

Mixture Intercept Model (MIM)

$$\text{LogYrOn} \sim \begin{cases} \alpha + \gamma \cdot \text{LogAge0} + \beta \cdot \text{Black} + \varepsilon & \text{with probability } p, \\ (\alpha + \delta) + \gamma \cdot \text{LogAge0} + \beta \cdot \text{Black} + \varepsilon & \text{with probability } 1 - p. \end{cases} \tag{7.2}$$

Mixture Slope Model (MSM)

$$\text{LogYrOn} \sim \begin{cases} \alpha + \gamma \cdot \text{LogAge0} + \beta \cdot \text{Black} + \varepsilon & \text{with probability } p, \\ \alpha + (\gamma + \delta) \cdot \text{LogAge0} + \beta \cdot \text{Black} + \varepsilon & \text{with probability } 1 - p. \end{cases} \tag{7.3}$$

Double Mixture Model (DMM)

$$\text{LogYrOn} \sim \begin{cases} \alpha_1 + \gamma \cdot \text{LogAge0} + \beta_1 \cdot \text{Black} + \varepsilon & \text{with probability } p, \\ \alpha_2 + \gamma \cdot \text{LogAge0} + \beta_2 \cdot \text{Black} + \varepsilon & \text{with probability } 1 - p. \end{cases} \tag{7.4}$$

Unrestricted Mixture Linear Regression Model (UMLRM)

$$\text{LogYrOn} \sim \begin{cases} \alpha_1 + \gamma_1 \cdot \text{LogAge0} + \beta_1 \cdot \text{Black} + \varepsilon & \text{with probability } p, \\ \alpha_2 + \gamma_2 \cdot \text{LogAge0} + \beta_2 \cdot \text{Black} + \varepsilon & \text{with probability } 1 - p. \end{cases} \tag{7.5}$$

Here, DMM is added by checking the results from UMLRM, and we assume equal variance in all those 4 models.

The model selection information is listed in table 7.1, and the model estimation results are listed in table 7.2. From these two table, we conclude:

1. From table 7.1, we can see that DMM is the best model from AIC and BIC (the smaller the AIC/BIC value, the better the model).

2. For the full model UMLRM, the estimate of mixing proportion p is 0.650 ± 0.0001 ($\hat{p} \pm s.e.$), which assure us that the LRTS should satisfy the regular conditions. Therefore, we can test UMLRM against DMM by the likelihood ratio test using χ^2_1, and get a p-value of 0.41 for LRTS value of 0.295. This means that there is no significant

difference on the slope of LogAge0 in two mixture components. Similarly, we also can test UMLRM against MSM, UMLRM against MIM, and DMM against MIM. Apparently, DMM is the best model among all mixture models from the likelihood ratio tests.

3. The LRTSs for testing MIM against OLS is 41.439, for MSM against OLS is 65.050 and for UMLRM against OLS is 107.095, which are probably highly significant when compared to the corresponding null LRTS distributions if all the corresponding null LRTS distributions diverge as slowly as the case for NMM.

4. DMM is the best model among the models in table 7.1 according to the LRTS.

5. Since $\hat{p}$ is far away from the boundary 0 or 1, the sample size is very large, and UMLRM is the full model, we can use *Wald tests* to do statistical inference for every single parameter in UMLRM.

Table 7.1: Model evaluation on regression models for COGEND data set.

Models	Log likelihood	LRTS	Degree of Freedom	AIC	BIC
OLS	-5574.620		4	11157.240	11184.302
MIM	-5553.900	41.439	6	11117.800	11151.627
MSM	-5542.095	65.050	6	11094.190	11128.020
DMM	-5521.220	106.800	7	11054.440	11086.267
UMLRM	-5521.073	107.095	8	11056.146	11103.504

From our best model, the double mixture model (DMM), we conclude that there are two groups of people regarding to their smoking characteristics. One group's transformed time to regular smoking is 4.314 ± 0.105, while the people from the second group need an additional transformed time 0.643 ± 0.019 to become regular smoker. The age effect seems the same in the two groups. In both components, subjects take less time to become regular

smoker when they become older. On the other hand, whether a person is black or not will not influence his time to become a regular smoker if he is in the more common component. But in the less common component, black people seem take longer time to become regular smoker.

Table 7.2: Regression models[a] for COGEND data set.

	Component 1 parameters (Std. Error)		Separation parameters (Std. Error)	
OLS				
	Intercept (α)	4.682 (0.104)		
	LogAge0 (γ)	-1.300 (0.039)		
	Black (β)	0.192 (0.027)		
MIM				
	Mixing proportion (p)	0.668 (0.0001)		
	Intercept (α)	4.265 (0.105)	(δ)	0.687 (0.018)
	LogAge0 (γ)	-1.228 (0.040)		
	Black (β)	0.172 (0.022)		
MSM				
	Mixing proportion (p)	0.818 (0.0001)		
	Intercept (α)	4.632 (0.108)		
	LogAge0 (γ)	-1.330 (0.041)	(δ)	0.274 (0.009)
	Black (β)	0.156 (0.023)		
DMM				
	Mixing proportion (p)	0.654 (0.0001)		
	Intercept (α)	4.314 (0.105)	($\alpha_2 - \alpha_1$)	0.643 (0.019)
	LogAge0 (γ)	-1.245 (0.040)		
	Black (β)	0.032 (0.033)	($\beta_2 - \beta_1$)	0.447 (0.050)
UMLRM				
	Mixing proportion (p)	0.650 (0.0001)		
	Intercept (α_1)	4.348 (0.145)	($\alpha_2 - \alpha_1$)	0.504 (0.243)
	LogAge0 (γ_1)	-1.260 (0.055)	($\gamma_2 - \gamma_1$)	0.055 (0.089)
	Black (β_1)	0.033 (0.033)	($\beta_2 - \beta_1$)	0.433 (0.051)

[a]OLS: ordinary least regression model, MIM: mixture intercept model, MSM: mixture slope model, DMM: double mixture model, UMLRM: unrestricted mixture linear regression model

7.2 Application to Pima Data Set

We use OLS to build a linear model on glucose concentration and other covariates for the Pima data set mentioned in Chapter One. We get the linear model as

$$\text{Glucose} = \alpha + \gamma \cdot \text{Insulin} + \beta \cdot \text{Age} + \varepsilon. \tag{7.6}$$

In order to answer the questions raised in Chapter One, we carry out several mixture linear regression model analyses. Based on the OLS model (7.6), we only include the covariates insulin and age in the mixture model study. The models used here are similar to those in the previous section.

The model selection information is listed in table 7.3, and the model estimation results are listed in table 7.4.

With the same procedures used in the COGEND data set, we find that the mixture intercept model is the best model for this Pima data set. Again, the LRTS results agree with the AIC and BIC model selection results.

From the mixture intercept model, we can conclude that:

- The Pima data set seems to come from two groups, with one component having about 84% of the total.
- The larger component has baseline glucose concentration (defined as insulin = 0) around 80.246, and the other component has mean baseline glucose concentration around 127.684.
- All subjects have the same glucose concentration increase rate of 0.136 per unit insulin increase.

Table 7.3: Model evaluation on regression models[a] for Pima data set

Models	Log likelihood	LRTS	Degree of Freedom	AIC	BIC
OLS	-1804.000		4	3616.000	3631.885
MIM	-1783.449	41.101	6	3576.898	3596.754
MSM	-1789.659	28.682	6	3589.318	3609.174
UMLRM	-1782.359	43.282	8	3578.718	3606.517

[a]OLS: ordinary least regression model, MIM: mixture intercept model, MSM: mixture slope model, UMLRM: unrestricted mixture linear regression model

Table 7.4: Regression models for Pima data set.

	Component 1 parameters (Std. Error)		Separation parameters (Std. Error)
OLS			
	Intercept (α)	79.771 (3.972)	
	Insulin (γ)	0.138 (0.010)	
	Age (β)	0.691 (0.122)	
MIM			
	Mixing proportion (p)	0.836 (0.022)	
	Intercept (α)	80.246 (3.138)	(δ) 47.438 (2.772)
	Insulin (γ)	0.136 (0.007)	
	Age (β)	0.434 (0.096)	
MSM			
	Mixing proportion (p)	0.783 (0.039)	
	Intercept (α)	77.835 (3.778)	
	Insulin (γ)	0.131 (0.009)	(δ) 0.190 (0.022)
	Age (β)	0.605 (0.114)	
UMLRM			
	Mixing proportion (p)	0.839 (0.022)	
	Intercept (α_1)	79.820 (10.529)	($\alpha_2 - \alpha_1$) 63.451 (10.512)
	Insulin (γ_1)	0.136 (0.033)	($\gamma_2 - \gamma_1$) -0.053 (0.034)
	Age (β_1)	0.453 (0.291)	($\beta_2 - \beta_1$) -0.188 (0.286)

- All subjects have the same glucose concentration increase rate of 0.434 per year older.

Chapter 8

Conclusions

This dissertation focused on the *restricted mixture linear regression models (RMLRM)*, especially two cases: the *mixture intercept model (MIM)* and the *mixture slope model (MSM)*.

We used the *Expectation-Maximization (EM)* algorithm to calculate the MLEs for the regression parameters and mixing proportions. We also provided the standard errors for the MLE. We provided two EM initialization procedures for MIM and MSM. Through pilot studies, we developed an EM implementation strategy by balancing the computation efficiency and high probability to obtain the global maximum of the LRTS. This implementation strategy has been validated by the follow-up simulation studies for finite sample sizes range from 100 to 1600.

We developed an approximation to decompose the distribution of the *likelihood ratio test statistic (LRTS)* for testing for a two-component mixture in *normal mixture model (NMM)*, MIM and MSM under the alternative. Using this decomposition method, we obtained two power and sample size estimation formulas for those mixture models. By a set of simulation studies, we verified that our power and sample size formulas gave usable estimations under a brand range of conditions. The results on NMM are close to a reported one [29] in many

circumstances. We also found that the accuracy of our power and sample size formulas drops when the separation between the two components decreases.

Through simulation studies, we investigated the null LRTS distributions of the test for two-component mixture in certain NMM, MIM, MSM and unrestricted mixture linear regression model (UMLRM) for sample sizes between 100 and 1600. We found that the empirical null LRTS distribution was relatively insensitive to the sample size. We also used simulation studies of the power to detect various mixture models by NMM, MIM, MSM and UMLRM. We found that among NMM, MIM and UMLRM, NMM was the more powerful method to detect mixture intercepts in many cases with sample size of 200. Among MSM and UMLRM, MSM was the more powerful method to detect mixture slopes.

We applied the RMLRM to two case studies with real data sets. In both cases, compared to the ordinary linear regression model (OLRM) and UMLRM, the RMLRM turned out to be the better model according three different model evaluation criteria: the LRTS, the *Akaike's information criterion (AIC)* and the *Bayesian information criterion (BIC)*. These two case studies demonstrated that RMLRM was a class of parsimonious mixture models that is very easy to use to explore and test different mixture mechanisms. We therefore believe that RMLRM will be a useful tool for searching for genetic effects and gene-environment interactions, which are the underlying causes for many mixture distributions in genetic studies.

There is still some room to improve our study results. One direction is to have a more advanced EM implementation strategy to make sure we have higher probability to obtain global maximum for LRTS. One idea is to have an adaptive process to monitor the convergence pattern(Section 6.2.1) from the multiple EM runs for every sample. For those samples converge with Patten 3, we can run EM multiple times once more from those points have highest LRTS results (as new starting points) and with more stringent stoping criteria

and number of maximum iterations.

When we derived the sample size formula (5.22), we used an approximation $q_{1-\alpha}(n_1) \approx q_{1-\alpha}(n_2)$ and have the final sample size formula as

$$n = \left\lceil \frac{2q_{1-\alpha}\overline{\lambda} + \sigma_\lambda^2 Z_\beta^2 + \sqrt{\sigma_\lambda^4 Z_\beta^4 + 4q_{1-\alpha}\overline{\lambda}\sigma_\lambda^2 Z_\beta^2}}{2\overline{\lambda}^2} \right\rceil, \tag{8.1}$$

where $q_{1-\alpha}$ was defined as the average of $q_{1-\alpha}(n)$ for sample sizes 100, 200, 400, 800 and 1600. In order to improve the accuracy of sample size formula, we might first obtain a rough sample size estimation n_r from formula 8.1, then use formula 8.1 again but with $q_{1-\alpha}(n_i)$ (n_i is the closet integer with n_r) replace $q_{1-\alpha}$. By repeating this process iteratively, sample size estimation might become more accurate.

Appendix A

Pilot Study Results

A.1 Figures for Normal Mixture Model Pilot Study Results

For every sample, we have 4 sets of figures. Sequentially, the first set of figures shows the simulation results based on 250 random starting points (RSPs) for p_0. The second set of figures shows the simulation results with 19 fixed starting points for p_0. The third set of figures shows only the simulation results with 250 RSPs for p_0 that have LRTS values greater than $(\text{LRTS}_{\max} + \text{LRTS}_{\min})/2$, where $\text{LRTS}_{\max}$ and $\text{LRTS}_{\min}$ are calculated based on 269 LRTS values. The fourth set of figures shows the simulation results started with 19 fixed starting p_0 that have LRTS values greater than $(\text{LRTS}_{\max} + \text{LRTS}_{\min})/2$. In all figures, each row corresponds to one simulation.

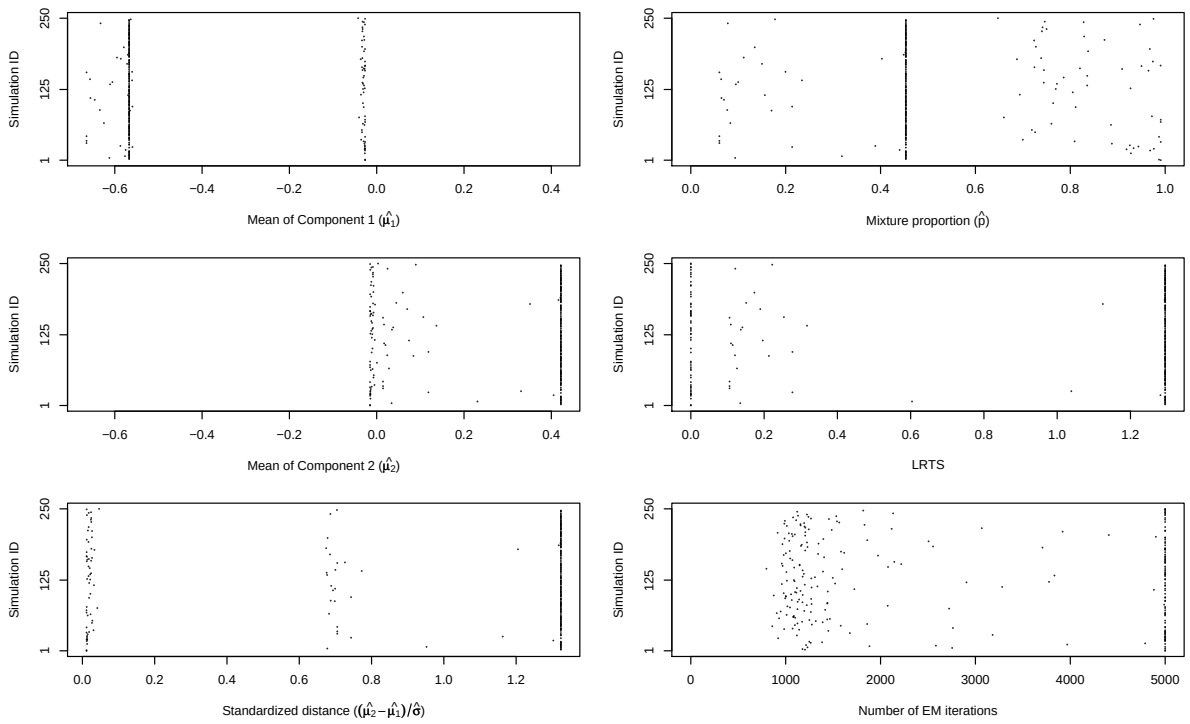

Figure A.1: Normal mixture model pilot study results for sample 2 ($n = 1600$) with 250 random starting p_0, stopping criteria $10e^{-12}$ and maximum number of iterations 5000.

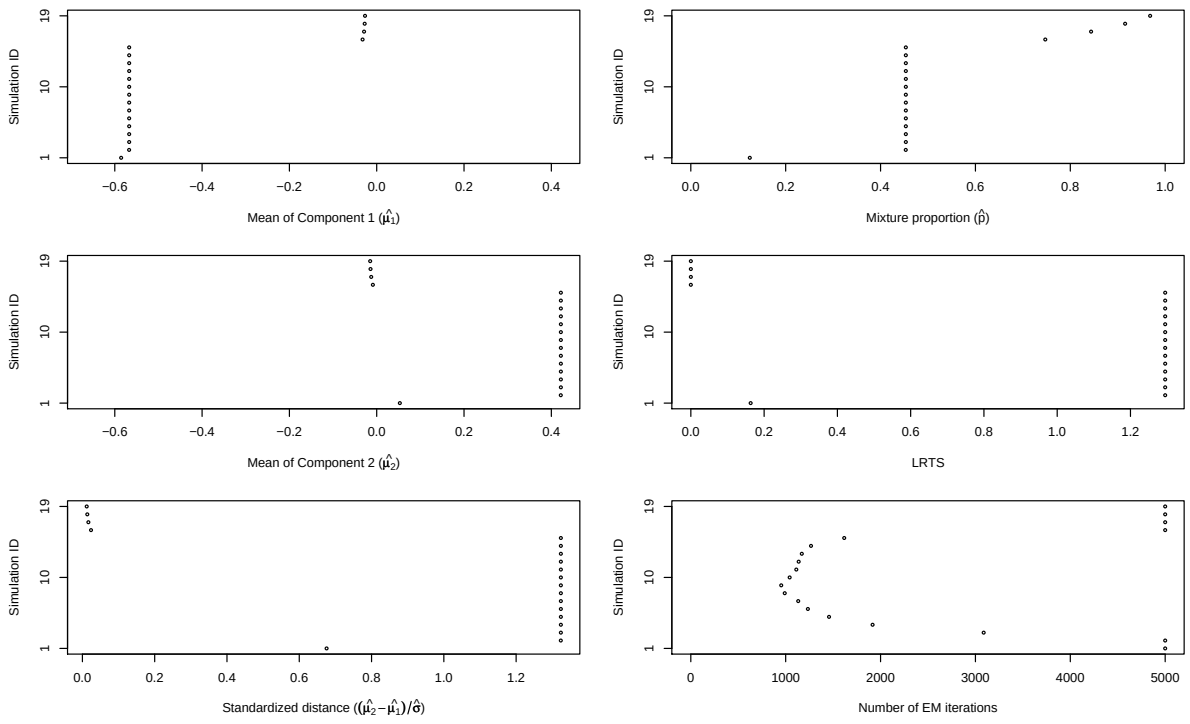

Figure A.2: Normal mixture model pilot study results for sample 2 ($n = 1600$) with 19 fixed starting p_0 $(0.05, 0.10, \ldots, 0.95)$, stopping criteria $10e^{-12}$ and maximum number of iterations 5000.

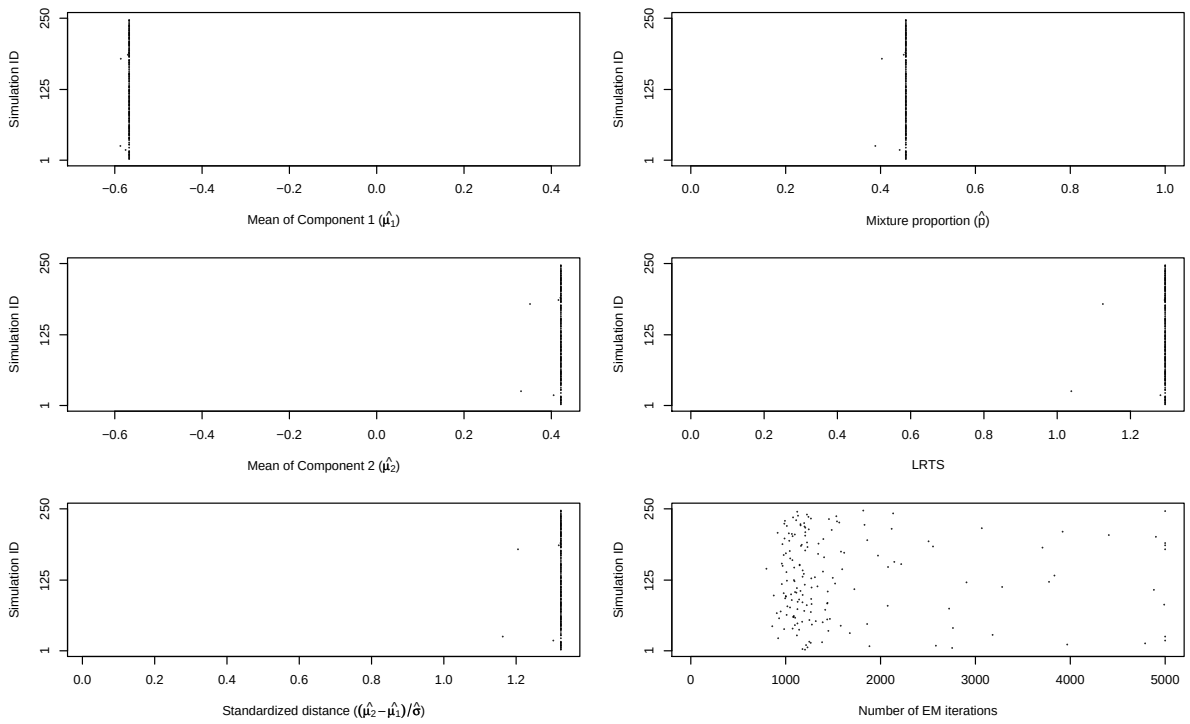

Figure A.3: Normal mixture model pilot study results for sample 2 ($n = 1600$) with 250 random starting p_0, stopping criteria $10e^{-12}$, maximum number of iterations 5000 and $\text{LRTS}(i) \geq (\text{LRTS}_{\max} + \text{LRTS}_{\min})/2$. Other starting point results are deleted.

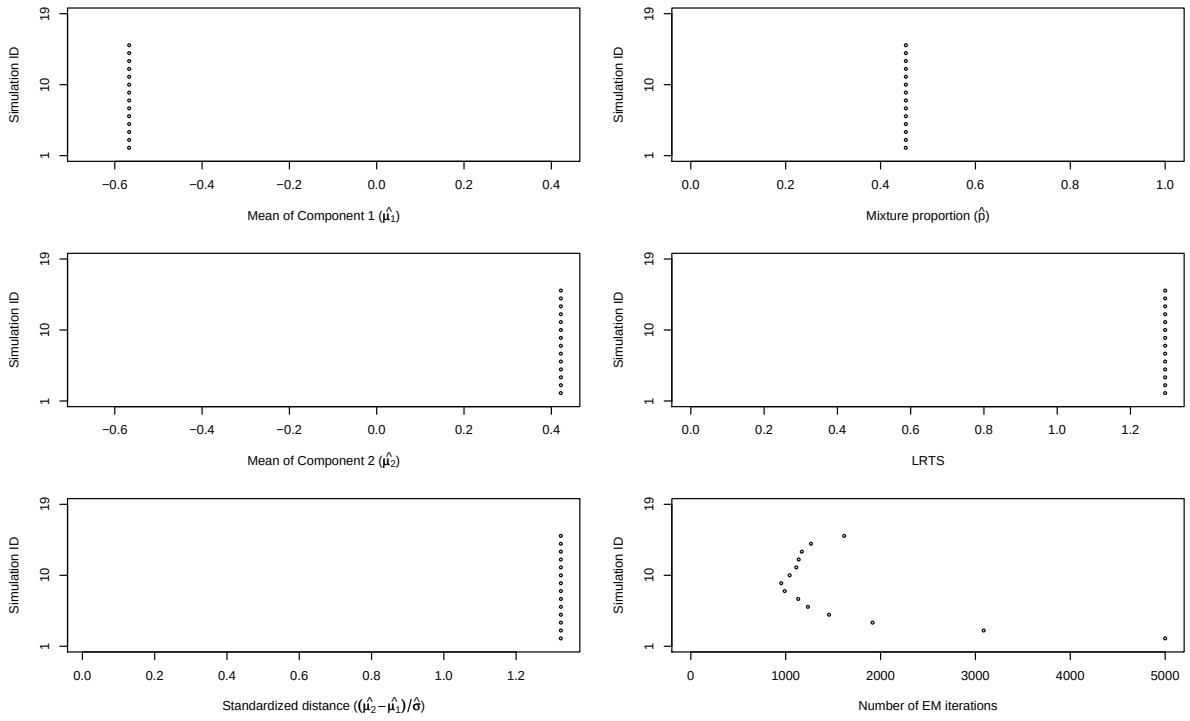

Figure A.4: Normal mixture model pilot study results for sample 2 ($n = 1600$) with 19 fixed starting p_0 $(0.05, 0.10, \ldots, 0.95)$, stopping criteria $10e^{-12}$, maximum number of iterations 5000 and $\text{LRTS}(i) \geq (\text{LRTS}_{\max} + \text{LRTS}_{\min})/2$. Other starting point results are deleted.

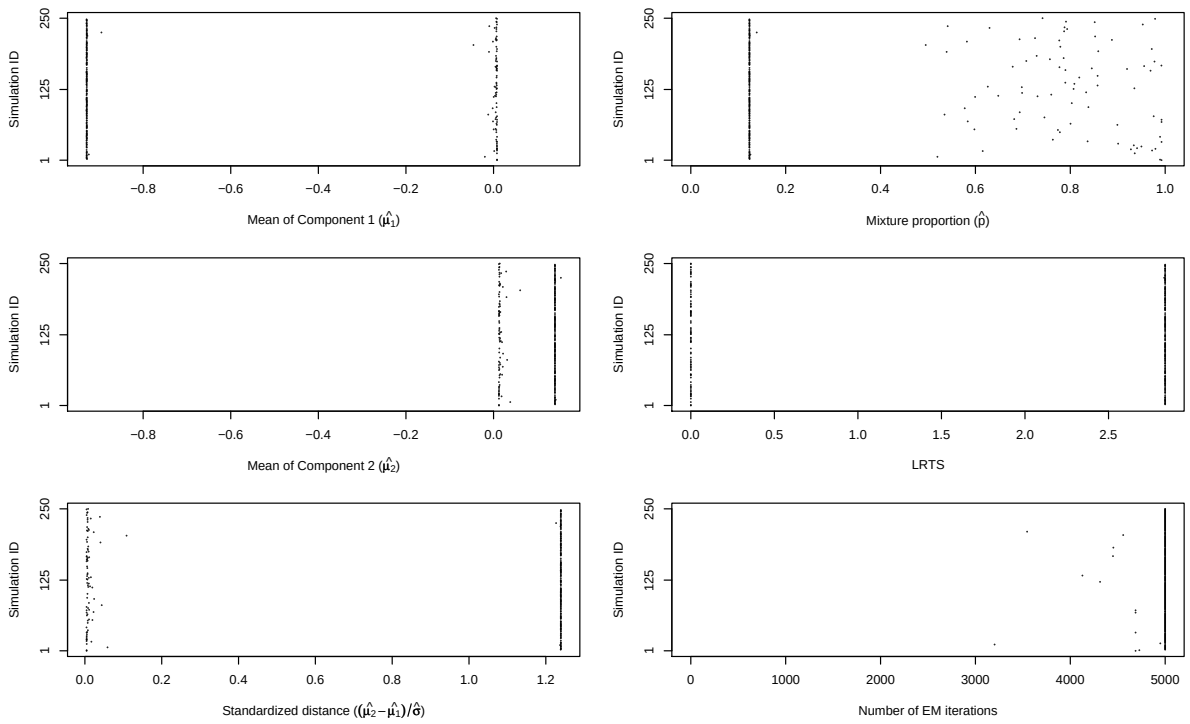

Figure A.5: Normal mixture model pilot study results for sample 5 ($n = 1600$) with 250 random starting p_0, stopping criteria $10e^{-12}$ and maximum number of iterations 5000.

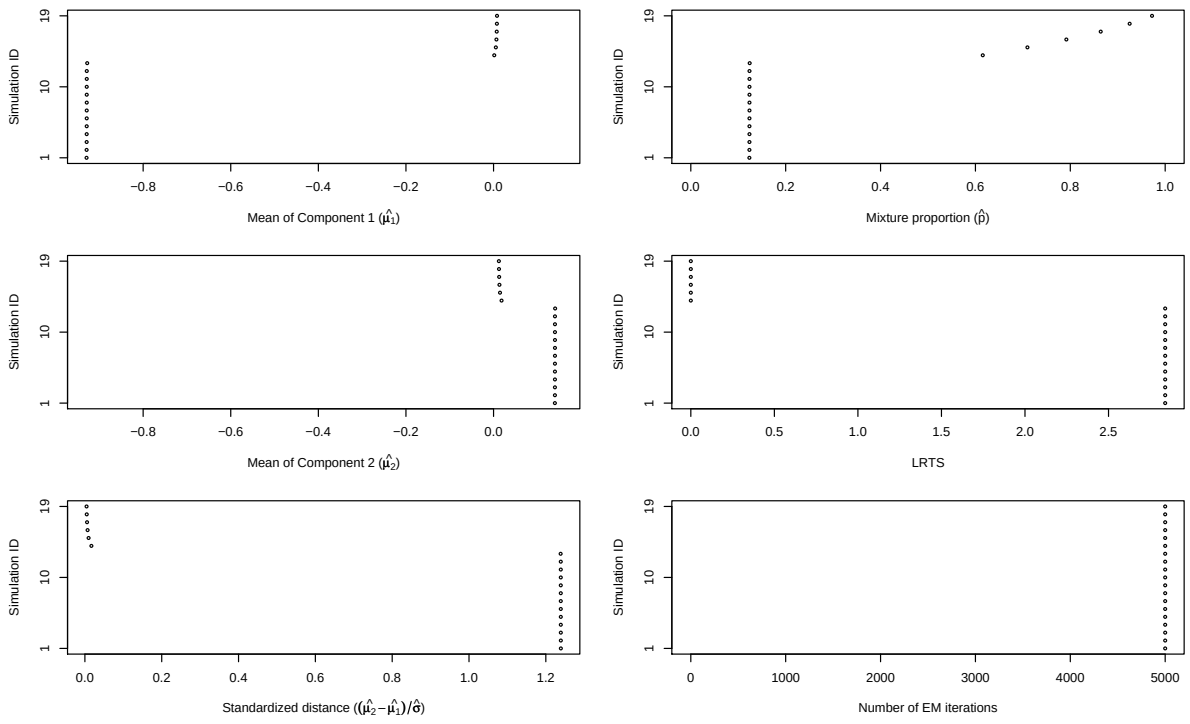

Figure A.6: Normal mixture model pilot study results for sample 5 ($n = 1600$) with 19 fixed starting p_0 $(0.05, 0.10, \ldots, 0.95)$, stopping criteria $10e^{-12}$ and maximum number of iterations 5000.

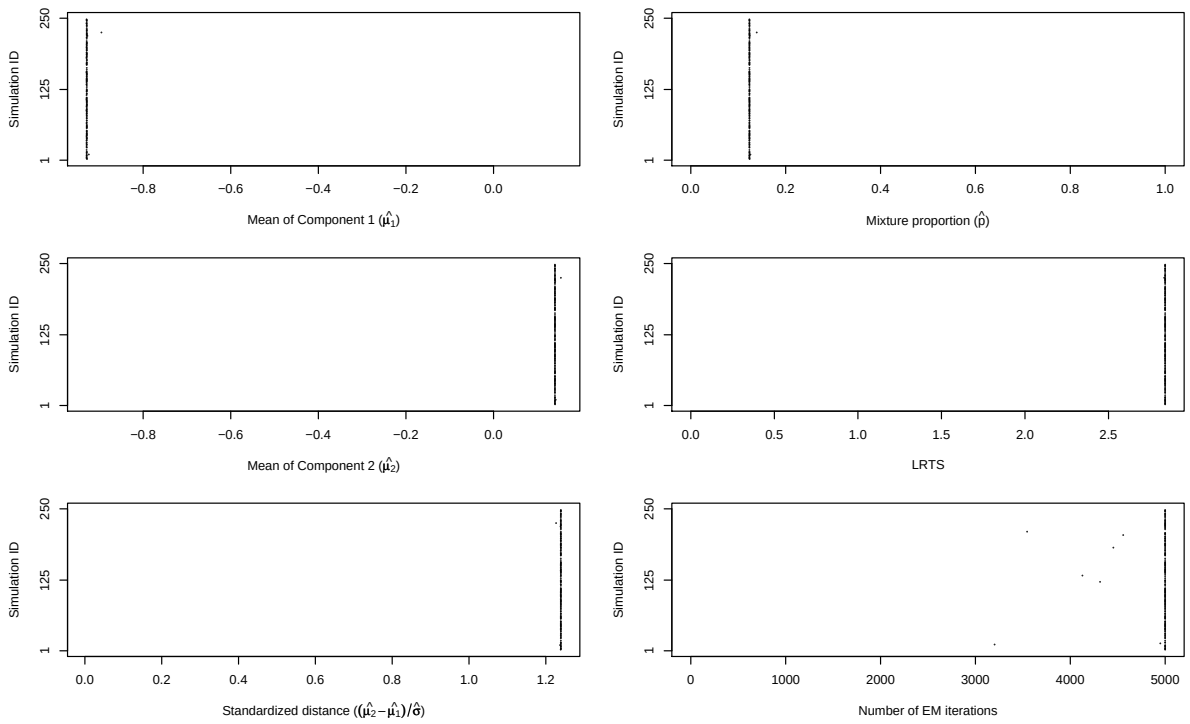

Figure A.7: Normal mixture model pilot study results for sample 5 ($n = 1600$) with 250 random starting p_0, stopping criteria $10e^{-12}$, maximum number of iterations 5000 and $\text{LRTS}(i) \geq (\text{LRTS}_{\max} + \text{LRTS}_{\min})/2$. Other starting point results are deleted.

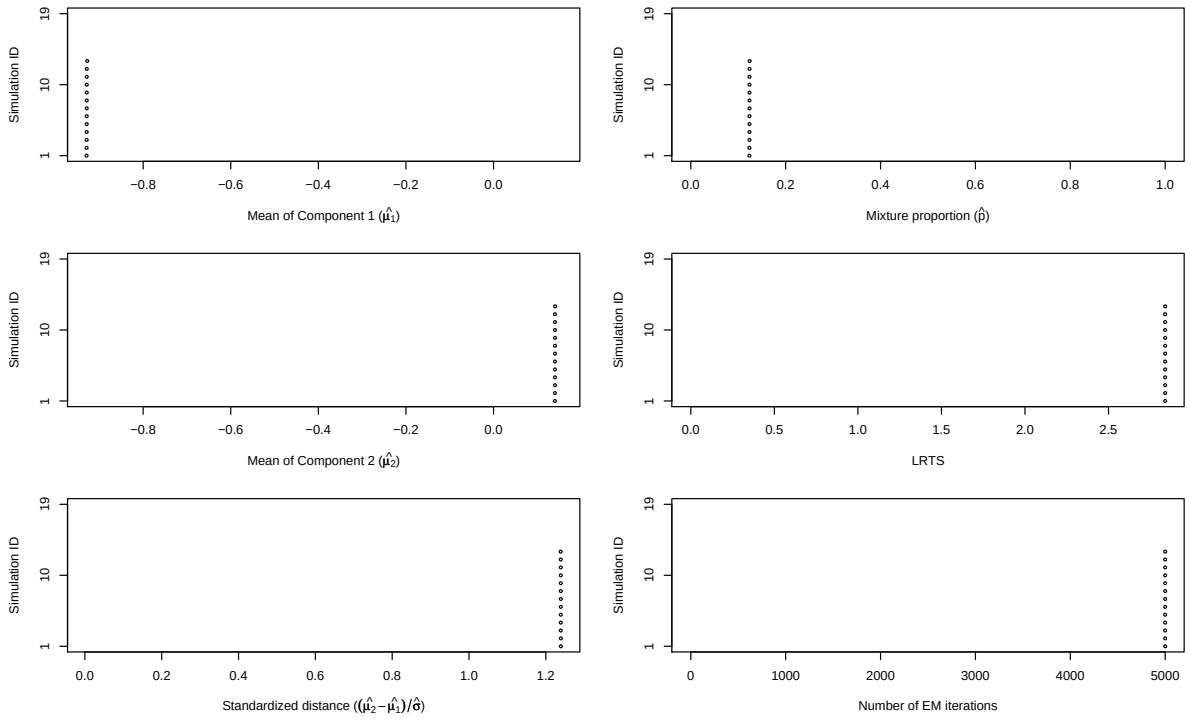

Figure A.8: Normal mixture model pilot study results for sample 5 ($n = 1600$) with 19 fixed starting p_0 $(0.05, 0.10, \ldots, 0.95)$, stopping criteria $10e^{-12}$, maximum number of iterations 5000 and $\text{LRTS}(i) \geq (\text{LRTS}_{\max} + \text{LRTS}_{\min})/2$. Other starting point results are deleted.

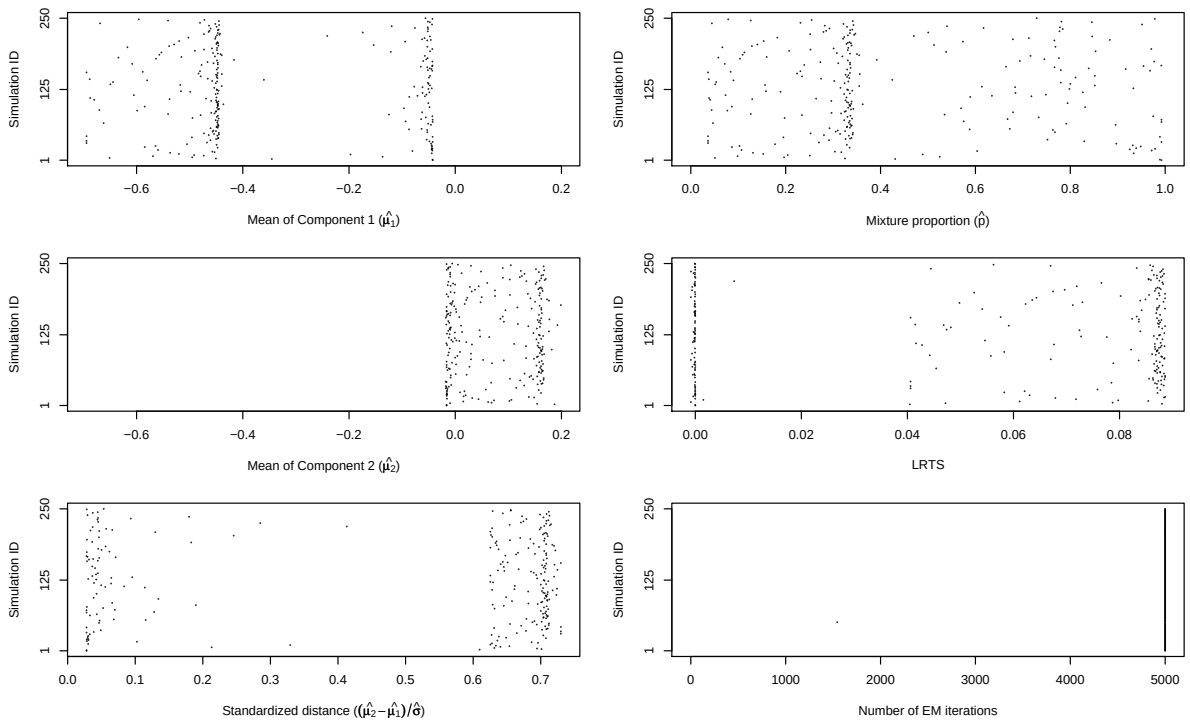

Figure A.9: Normal mixture model pilot study results for sample 6 ($n = 1600$) with 250 random starting p_0, stopping criteria $10e^{-12}$ and maximum number of iterations 5000.

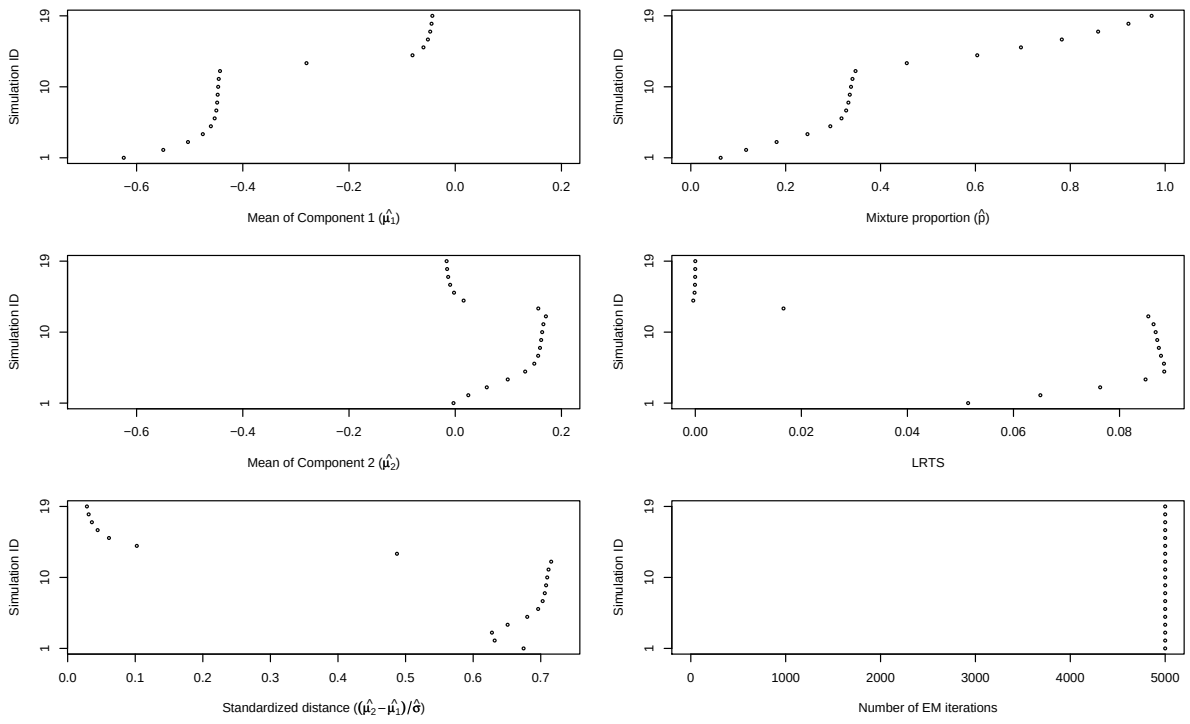

Figure A.10: Normal mixture model pilot study results for sample 6 ($n = 1600$) with 19 fixed starting p_0 $(0.05, 0.10, \ldots, 0.95)$, stopping criteria $10e^{-12}$ and maximum number of iterations 5000.

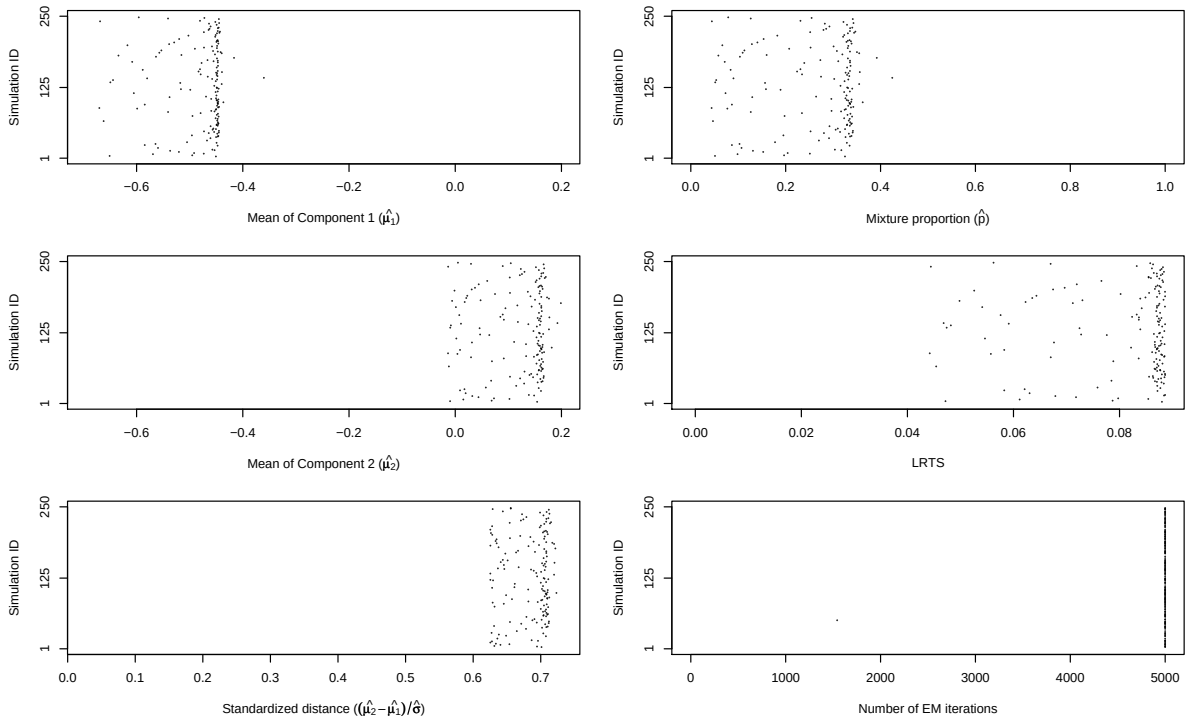

Figure A.11: Normal mixture model pilot study results for sample 6 ($n = 1600$) with 250 random starting p_0, stopping criteria $10e^{-12}$, maximum number of iterations 5000 and $\text{LRTS}(i) \geq (\text{LRTS}_{\max} + \text{LRTS}_{\min})/2$. Other starting point results are deleted.

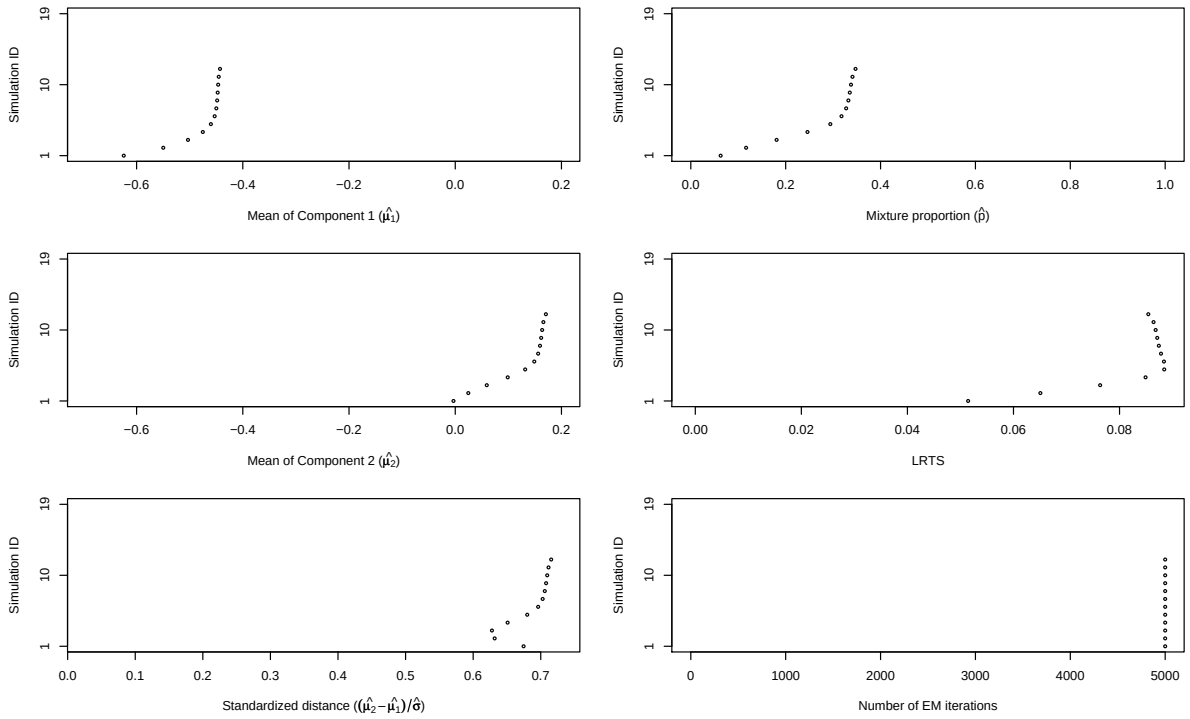

Figure A.12: Normal mixture model pilot study results for sample 6 ($n = 1600$) with 19 fixed starting p_0 $(0.05, 0.10, \ldots, 0.95)$, stopping criteria $10e^{-12}$, maximum number of iterations 5000 and $\text{LRTS}(i) \geq (\text{LRTS}_{\max} + \text{LRTS}_{\min})/2$. Other starting point results are deleted.

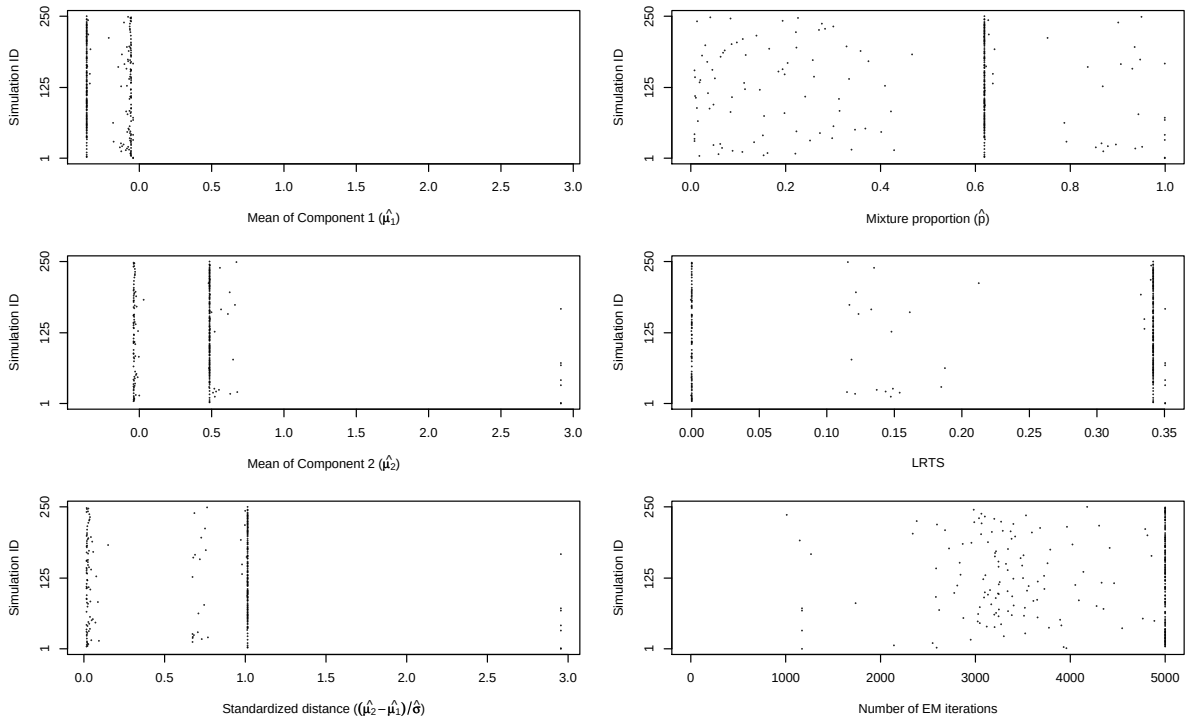

Figure A.13: Normal mixture model pilot study results for sample 17 ($n = 1600$) with 250 random starting p_0, stopping criteria $10e^{-12}$ and maximum number of iterations 5000.

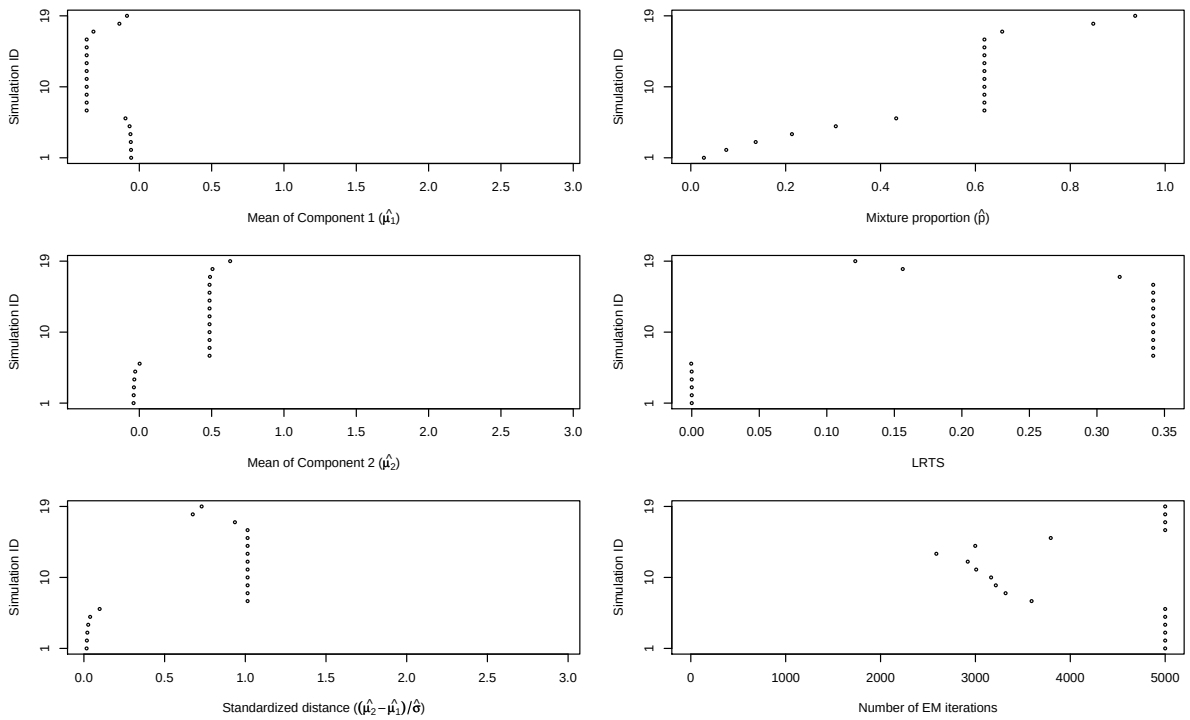

Figure A.14: Normal mixture model pilot study results for sample 17 ($n = 1600$) with 19 fixed starting p_0 $(0.05, 0.10, \ldots, 0.95)$, stopping criteria $10e^{-12}$ and maximum number of iterations 5000.

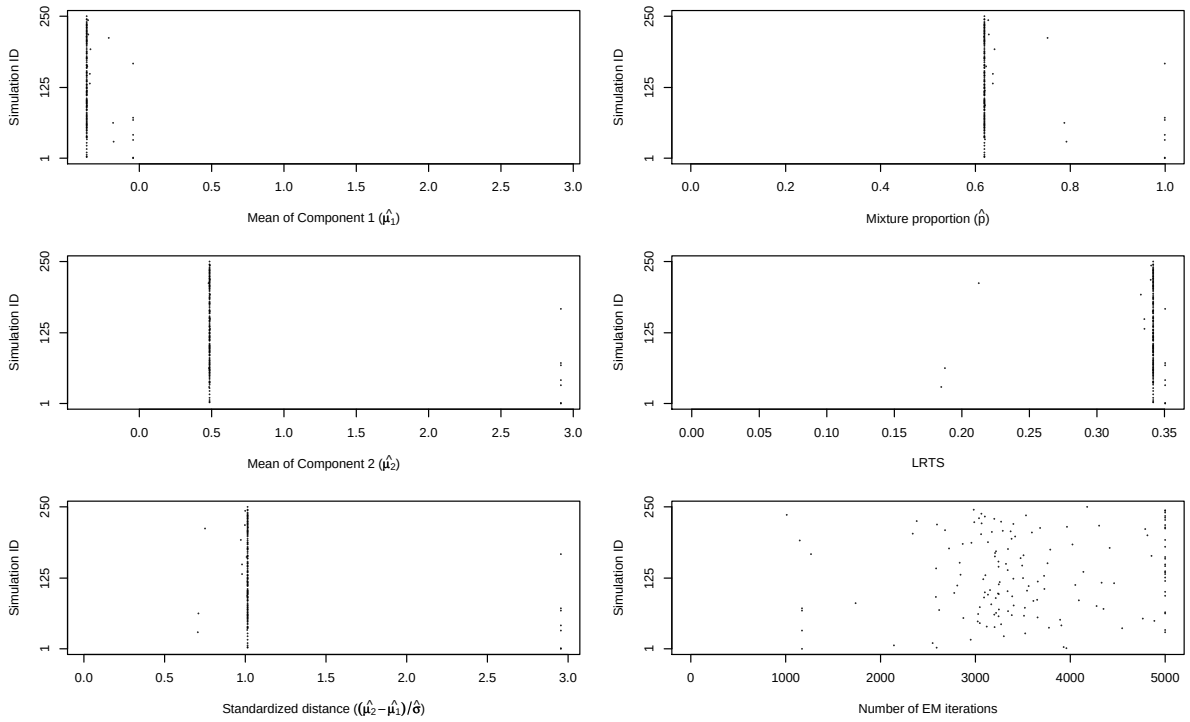

Figure A.15: Normal mixture model pilot study results for sample 17 ($n = 1600$) with 250 random starting p_0, stopping criteria $10e^{-12}$, maximum number of iterations 5000 and $\text{LRTS}(i) \geq (\text{LRTS}_{\max} + \text{LRTS}_{\min})/2$. Other starting point results are deleted.

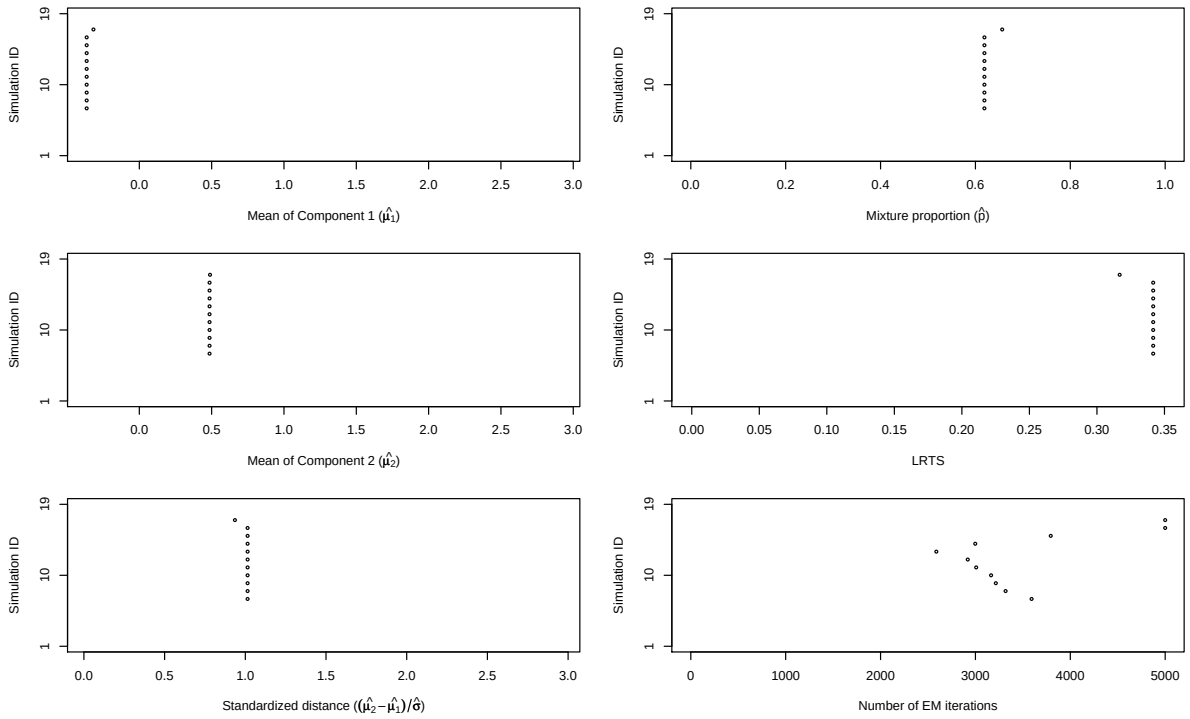

Figure A.16: Normal mixture model pilot study results for sample 17 ($n = 1600$) with 19 fixed starting p_0 $(0.05, 0.10, \ldots, 0.95)$, stopping criteria $10e^{-12}$, maximum number of iterations 5000 and $\text{LRTS}(i) \geq (\text{LRTS}_{\max} + \text{LRTS}_{\min})/2$. Other starting point results are deleted.

A.2 Lists for Highest LRTS Results from Normal Mixture Model Pilot Study

List five highest LRTS values for the 25 samples of sample size 1600 from 250 runs of EM with random starting points. The stopping criterion and the maximum number of iterations are set as $10e^{-12}$ and 5000 respectively.

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7
1	0.5635536	1.294745	1.072115	1.924701	2.838997	0.08860385	3.242292
2	0.5635532	1.294745	1.072115	1.924701	2.838997	0.08859925	3.242292
3	0.5635531	1.294745	1.072115	1.924701	2.838997	0.08859559	3.242292
4	0.5635480	1.294745	1.072115	1.924701	2.838997	0.08859039	3.242292
5	0.5635464	1.294745	1.072115	1.924701	2.838997	0.08857682	3.242292

	Sample 8	Sample 9	Sample 10	Sample 11	Sample 12	Sample 13	Sample 14
1	1.015727	1.328578	1.924701	0.5635536	2.199168	1.213649	0.2696057
2	1.015727	1.328578	1.924701	0.5635532	2.199168	1.213649	0.2696057
3	1.015727	1.328578	1.924701	0.5635531	2.199168	1.213649	0.2696057
4	1.015727	1.328578	1.924701	0.5635480	2.199168	1.213649	0.2696057
5	1.015727	1.328578	1.924701	0.5635464	2.199168	1.213649	0.2696057

	Sample 15	Sample 16	Sample 17	Sample 18	Sample 19	Sample 20	Sample 21
1	1.568665	2.069853	0.3505202	4.848534	3.811048	0.9022956	1.292325
2	1.568665	2.069853	0.3505202	4.848534	3.811048	0.9022942	1.292325
3	1.568665	2.069853	0.3505202	4.848534	3.811048	0.9022932	1.292325
4	1.568665	2.069853	0.3505202	4.848534	3.811048	0.9022930	1.292325
5	1.568665	2.069853	0.3505202	4.848534	3.811048	0.9022765	1.292325

	Sample 22	Sample 23	Sample 24	Sample 25
1	0.1671775	0.1616927	1.371054	0.929378
2	0.1671775	0.1616927	1.371054	0.929378
3	0.1671775	0.1616927	1.371054	0.929378
4	0.1671775	0.1616927	1.371054	0.929378
5	0.1671775	0.1616927	1.371054	0.929378

List five highest LRTS values for the 25 samples of sample size 1600 from 19 fixed starting p_0 $(0.05, 0.10, \ldots, 0.95)$. The stopping criterion and the maximum number of iterations are set as $10e^{-12}$ and 5000 respectively.

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7
1	0.5635075	1.294745	1.072115	1.924701	2.838997	0.08842816	3.242292
2	0.5634571	1.294745	1.072115	1.924701	2.838997	0.08837543	3.242292
3	0.5633672	1.294745	1.072115	1.924701	2.838997	0.08781697	3.242292
4	0.5632727	1.294745	1.072115	1.924701	2.838997	0.08741121	3.242292
5	0.5632126	1.294745	1.072115	1.924701	2.838997	0.08710567	3.242292

	Sample 8	Sample 9	Sample 10	Sample 11	Sample 12	Sample 13	Sample 14
1	1.015727	1.328578	1.924701	0.5635075	2.199168	1.21364935	0.2696057
2	1.015727	1.328578	1.924701	0.5634571	2.199168	0.62734374	0.2695738
3	1.015727	1.328578	1.924701	0.5633672	2.199168	0.02973261	0.2570691
4	1.015727	1.328578	1.924701	0.5632727	2.199168	0.02450725	0.1806021
5	1.015727	1.328578	1.924701	0.5632126	2.199168	0.01761707	0.1277288

	Sample 15	Sample 16	Sample 17	Sample 18	Sample 19	Sample 20	Sample 21
1	1.568665	2.069853	0.3416523	4.848534	3.811048	0.9022675	1.292325
2	1.568665	2.069852	0.3416523	4.848534	3.811048	0.9019160	1.292325
3	1.568665	2.069852	0.3416523	4.848534	3.811048	0.9011290	1.292325
4	1.568665	2.069850	0.3416523	4.848534	3.811048	0.9009077	1.292325
5	1.568665	2.069850	0.3416523	4.848534	3.811048	0.9005665	1.292325

	Sample 22	Sample 23	Sample 24	Sample 25
1	0.16717748	0.16169274	1.371054	0.929378
2	0.16717503	0.08021941	1.371054	0.929378
3	0.07599023	0.07620611	1.371054	0.929378
4	0.02548698	0.01928422	1.371054	0.929378
5	0.01126419	0.01847642	1.371054	0.929378

List five highest LRTS values for the 25 samples of sample size 1600 from 19 fixed starting p_0 $(0.05, 0.10, \ldots, 0.95)$. The stopping criterion and the maximum number of iterations are set as $10e^{-6}$ and 1000 respectively.

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7
1	0.5615610	1.294636	1.072078	1.924677	2.838050	0.08783886	3.242141
2	0.5605051	1.294635	1.072039	1.924610	2.837066	0.08685329	3.242141
3	0.5537783	1.294634	1.072039	1.924610	2.832685	0.08152506	3.242141
4	0.5454167	1.294634	1.072039	1.924610	2.827528	0.08148298	3.242140
5	0.5431101	1.294634	1.072039	1.924610	2.823774	0.07440544	3.242140

	Sample 8	Sample 9	Sample 10	Sample 11	Sample 12	Sample 13	Sample 14
1	1.015264	1.3283909	1.924677	0.5615610	2.199091	1.21364	0.2645797
2	1.015213	1.2722012	1.924610	0.5605051	2.199090	0.02919	0.2075024
3	1.014947	1.0483872	1.924610	0.5537783	2.199090	0.02285	0.1498254
4	1.014525	0.8947805	1.924610	0.5454167	2.199090	0.01550	0.1096358
5	1.014203	0.7867136	1.924610	0.5431101	2.199090	0.00777	0.0787382

	Sample 15	Sample 16	Sample 17	Sample 18	Sample 19	Sample 20	Sample 21
1	1.568262	2.068951	0.3413153	4.848457	3.811007	0.9017311	1.2922088
2	1.568261	2.066360	0.3412958	4.848437	3.811004	0.8967667	1.2902800
3	1.568219	2.061688	0.3412953	4.848437	3.811003	0.8882521	1.0731705
4	1.568055	2.052841	0.3412951	4.848437	3.811003	0.8801589	0.7342403
5	1.567873	2.045081	0.3412488	4.848436	3.811003	0.8625504	0.5118575

	Sample 22	Sample 23	Sample 24	Sample 25
1	0.1617292	0.0934758	1.370987	0.9286565
2	0.0780836	0.0512582	1.370970	0.9285343
3	0.0355069	0.0313076	1.370970	0.9266783
4	0.0175219	0.0109998	1.370970	0.9248821
5	0.0082942	0.0040486	1.370969	0.9233363

List five highest LRTS values for the 25 samples of sample size 1600 from 19 fixed starting p_0 $(0.05, 0.10, \ldots, 0.95)$. The stopping criterion and the maximum number of iterations are set as $10e^{-4}$ and 1000 respectively.

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7
1	0.5564750	1.289480	1.068653	1.922323	2.822416	0.077091	3.239656
2	0.5551346	1.289283	1.066952	1.915778	2.812049	0.074372	3.228673
3	0.5298504	1.284576	1.066665	1.915732	2.782185	0.071994	3.228660
4	0.5258008	1.284521	1.064940	1.915544	2.782139	0.063209	3.228651
5	0.4871876	1.284451	1.064895	1.915499	2.782112	0.062198	3.228646

	Sample 8	Sample 9	Sample 10	Sample 11	Sample 12	Sample 13	Sample 14
1	1.006335	1.289733	1.922323	0.5564750	2.196285	1.2127920	0.2440496
2	0.999849	1.272201	1.915778	0.5551346	2.191783	0.0239441	0.1782463
3	0.982570	0.875795	1.915732	0.5298504	2.191736	0.0144483	0.1291768
4	0.982556	0.744960	1.915544	0.5258008	2.191707	0.0031206	0.0919265
5	0.982516	0.704730	1.915499	0.4871876	2.191689	-0.0106960	0.0602772

	Sample 15	Sample 16	Sample 17	Sample 18	Sample 19	Sample 20	Sample 21
1	1.560292	2.061380	0.3333734	4.847100	3.809186	0.8996562	1.2760285
2	1.549542	2.045702	0.3297904	4.839785	3.806914	0.8886076	1.2760011
3	1.539345	2.000308	0.3247375	4.839763	3.806637	0.8761912	1.0731705
4	1.539343	2.000306	0.3176581	4.839755	3.806630	0.8437032	0.7342403
5	1.539291	2.000302	0.3163081	4.839733	3.806625	0.7925276	0.5118575

	Sample 22	Sample 23	Sample 24	Sample 25
1	0.1234545	0.0628374	1.367787	0.9217783
2	0.0561300	0.0270301	1.365906	0.9165858
3	0.0225779	0.0195531	1.363151	0.8837018
4	0.0031636	-0.0028043	1.362997	0.8821973
5	-0.0105910	-0.0084720	1.362942	0.8821918

A.3 Lists for Highest LRTS Results from Mixture Slope Model Pilot Study

List five highest LRTS values for the 25 samples of sample size 100 from 250 runs of EM with random starting points. The stopping criterion and the maximum number of iterations are set as $10e^{-10}$ and 5000 respectively.

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7
1	2.77118	0.16154	0	0.135	4.11833	2.93271	0
2	2.77118	0.16154	0	0.135	4.11833	2.93271	0
3	2.77118	0.16154	0	0.135	4.11833	2.93271	0
4	2.77118	0.16154	0	0.135	4.11833	2.93271	0
5	2.77118	0.16154	0	0.135	4.11833	2.93271	0

	Sample 8	Sample 9	Sample 10	Sample 11	Sample 12	Sample 13	Sample 14
1	0	0.07544	0.135	2.77118	0	0.15893	0.09069
2	0	0.07544	0.135	2.77118	0	0.15893	0.09069
3	0	0.07544	0.135	2.77118	0	0.15893	0.09069
4	0	0.07544	0.135	2.77118	0	0.15893	0.09069
5	0	0.07544	0.135	2.77118	0	0.15893	0.09069

	Sample 15	Sample 16	Sample 17	Sample 18	Sample 19	Sample 20	Sample 21
1	0.90692	4.07236	0.05258	2.46485	0	0.98577	0.11876
2	0.90692	4.07236	0.05258	2.46485	0	0.98577	0.11875
3	0.90692	4.07236	0.05258	2.46485	0	0.00000	0.11875
4	0.90692	4.07236	0.05258	2.46485	0	0.00000	0.11875
5	0.90692	4.07236	0.05258	2.46485	0	0.00000	0.11875

	Sample 22	Sample 23	Sample 24	Sample 25
1	1.33397	1.19183	4.58051	0.00485
2	1.33397	1.19183	4.58051	0.00485
3	1.33397	1.19183	4.58051	0.00485
4	1.33397	0.00003	4.58051	0.00485
5	1.33397	0.00000	4.58051	0.00485

List five highest LRTS values for the 25 samples of sample size 100 from 19 runs of EM with fixed starting points p_0 $(0.05, 0.10, \ldots, 0.95)$. The stopping criterion and the maximum number of iterations are set as $10e^{-10}$ and 5000 respectively.

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7
1	2.77118	0.16154	0	0.135	4.11833	2.93271	0
2	2.77118	0.16154	0	0.135	4.11833	2.93271	0
3	2.77118	0.16154	0	0.135	4.11833	2.93271	0
4	2.77118	0.16154	0	0.135	4.11833	2.93271	0
5	2.77118	0.16154	0	0.135	4.11833	2.93271	0

	Sample 8	Sample 9	Sample 10	Sample 11	Sample 12	Sample 13	Sample 14
1	0	0.07544	0.135	2.77118	0	0.15893	0.09069
2	0	0.07544	0.135	2.77118	0	0.15893	0.09069
3	0	0.07544	0.135	2.77118	0	0.15893	0.09069
4	0	0.07544	0.135	2.77118	0	0.15893	0.09069
5	0	0.07544	0.135	2.77118	0	0.15893	0.09069

	Sample 15	Sample 16	Sample 17	Sample 18	Sample 19	Sample 20	Sample 21
1	0.90692	4.07236	0.05258	2.46485	0	0.98577	0.11875
2	0.90692	4.07236	0.00000	2.46485	0	0.00000	0.11875
3	0.90692	4.07236	0.00000	2.46485	0	0.00000	0.11875
4	0.90692	4.07236	0.00000	2.46485	0	0.00000	0.11875
5	0.90692	4.07236	0.00000	2.46485	0	0.00000	0.11875

	Sample 22	Sample 23	Sample 24	Sample 25
1	1.33397	1.19183	4.58051	0.00485
2	1.33397	1.19183	4.58051	0.00485
3	1.33397	0.00000	4.58051	0.00485
4	1.33397	0.00000	4.58051	0.00485
5	1.33397	0.00000	4.58051	0.00485

List five highest LRTS values for the 25 samples of sample size 100 from 19 runs of EM with fixed starting points p_0 $(0.05, 0.10, \ldots, 0.95)$. The stopping criterion and the maximum number of iterations are set as $10e^{-4}$ and 1000 respectively.

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7
1	2.77021	0.15889	-0.00035	0.13344	4.11804	2.93246	-0.00038
2	2.77014	0.15883	-0.00035	0.13288	4.11804	2.93221	-0.00040
3	2.77014	0.15882	-0.00036	0.13205	4.11804	2.93220	-0.00043
4	2.77010	0.15880	-0.00036	0.13021	4.11803	2.93218	-0.00043
5	2.77010	0.15875	-0.00036	0.12734	4.11803	2.93218	-0.00046

	Sample 8	Sample 9	Sample 10	Sample 11	Sample 12	Sample 13	Sample 14
1	-0.00142	0.07329	0.13344	2.77021	-0.00183	0.15738	0.09014
2	-0.00177	0.07296	0.13288	2.77014	-0.00191	0.15491	0.07891
3	-0.00198	0.07167	0.13205	2.77014	-0.00202	0.15388	0.07883
4	-0.00206	0.07013	0.13021	2.77010	-0.00207	0.14823	0.07882
5	-0.00230	0.06928	0.12734	2.77010	-0.00211	0.14282	0.07872

	Sample 15	Sample 16	Sample 17	Sample 18	Sample 19	Sample 20	Sample 21
1	0.90604	4.07179	0.04982	2.46457	-0.00051	0.98568	0.11747
2	0.90499	4.07176	-0.00056	2.46456	-0.00063	-0.00053	0.11696
3	0.90230	4.07176	-0.00057	2.46456	-0.00063	-0.00053	0.11546
4	0.90228	4.07175	-0.00064	2.46454	-0.00064	-0.00054	0.11363
5	0.90226	4.07175	-0.00065	2.46454	-0.00064	-0.00054	0.10835

	Sample 22	Sample 23	Sample 24	Sample 25
1	1.33351	1.19176	4.57973	0.00312
2	1.33224	1.19175	4.57972	0.00183
3	1.33186	-0.00074	4.57971	-0.00100
4	1.33184	-0.00075	4.57970	-0.00142
5	1.33183	-0.00076	4.57970	-0.00174

Bibliography

[1] Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*, 6 (1974), 716–723.

[2] Atiqullah, M. Estimation of residual variance in quadratically balanced least-squares problems and robustness of F-test. *Biometrika 49*, 1-2 (1962), 83–91.

[3] Bierut, L. The collaborative genetic study of nicotine dependence (http://cancercontrol.cancer.gov/grants/abstract.asp?applid=6947236). Tech. rep., National Cancer Institue.

[4] Casella, G., and Berger, R. *Statistical inference (2 edition)*. Duxbury Press, 2001.

[5] Caspi, A., McClay, J., Moffitt, T., Mill, J., Martin, J., Craig, I., Taylor, A., and Poulton, R. Role of genotype in the cycle of violence in maltreated children. *Science 297*, 5582 (Aug. 2002), 851–854.

[6] Ceppellini, R., Siniscalco, M., and Smith, C. The estimation of gene frequencies in a random-mating population. *Annals of Human Genetics 20*, 2 (1955), 97–115.

[7] Christensen, R. *Plane answers to complex questions: the theory of linear models (3 edition)*. Springer, 2002.

[8] Cohen, A. Estimation in mixtures of two normal distributions. *Technometrics 9*, 1 (Feb. 1967), 15–28.

[9] Davison, A. *Statistical models*. Cambridge University Press, 2003.

[10] Dempster, A., Laird, N., and Rubin, D. Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society Series B-Methodological 39*, 1 (1977), 1–38.

[11] Everitt, B. Maximum-likelihood estimation of the parameters in a mixture of 2 univariate normal-distributions - a comparison of different algorithms. *Statistician 33*, 2 (1984), 205–215.

[12] Feng, Z., and McCulloch, C. Using bootstrap likelihood ratios in finite mixture models. *Journal of the Royal Statistical Society Series B-Methodological 58*, 3 (1996), 609–617.

[13] Finch, S., Mendell, N., and Thode, H. Probabilistic measures of adequacy of a numerical search for a global maximum. *Journal of the American Statistical Association 84*, 408 (Dec. 1989), 1020–1023.

[14] Garel, B. Likelihood ratio test for univariate gaussian mixture. *Journal of Statistical Planning And Inference 96*, 2 (July 2001), 325–350.

[15] Ghosh, J., and Sen, P. On the asymptotic performance of the log-likelihood ratio statistic for the mixture model and related results. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, vol. II* (Monterey, CA, 1985), L. Le Cam and R. Olshen, Eds., Wadsworth, pp. 789–806.

[16] Hall, P., and Stewart, M. Theoretical analysis of power in a two-component normal mixture model. *Journal of Statistical Planning And Inference 134*, 1 (Sept. 2005), 158–179.

[17] Hartigan, J. A failure of likelihood asymptotics for normal mixtures. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, vol. II* (Monterey, CA, 1985), L. Le Cam and R. Olshen, Eds., Wadsworth, pp. 807–810.

[18] Liu, X., and Shao, Y. Asymptotics for the likelihood ratio test in a two-component normal mixture model. *Journal of Statistical Planning And Inference 123*, 1 (June 2004), 61–81.

[19] Lo, Y. Likelihood ratio tests of the number of components in a normal mixture with unequal variances. *Statistics & Probability Letters 71*, 3 (Mar. 2005), 225–235.

[20] Lo, Y., Mendell, N., and Rubin, D. Testing the number of components in a normal mixture. *Biometrika 88*, 3 (Sept. 2001), 767–778.

[21] Louis, T. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society Series B-Methodological 44*, 2 (1982), 226–233.

[22] Luenberger, D. *Linear and nonlinear programming (2 edition).* Springer, 2003.

[23] Maclean, C., Morton, N., Elston, R., and Yee, S. Skewness in commingled distributions. *Biometrics 32*, 3 (1976), 695–699.

[24] MacQueen, J. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability* (1967), University of California Press, pp. 281–297.

[25] McLachlan, G. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics-Journal of the Royal Statistical Society Series C 36*, 3 (1987), 318–324.

[26] McLachlan, G., and Peel, D. *Finite mixture models.* Wiley, New York, 2000.

[27] Meilijson, I. A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society Series B-Methodological 51*, 1 (1989), 127–138.

[28] Mendell, N., Finch, S., and Thode, H. Where is the likelihood ratio test powerful for detecting 2-component normal mixtures. *Biometrics 49*, 3 (Sept. 1993), 907–915.

[29] Mendell, N., Thode, H., and Finch, S. The likelihood ratio test for the 2-component normal mixture problem - power and sample-size analysis. *Biometrics 47*, 3 (Sept. 1991), 1143–1148.

[30] Newman, D., Hettich, S., Blake, C., and Merz, C. UCI repository of machine learning databases, 1998.

[31] Ning, Y., and Finch, S. The null distribution of the likelihood ratio test for a mixture of two normals after a restricted box-cox transformation. *Communications in Statistics-simulation And Computation 29*, 2 (2000), 449–461.

[32] Ning, Y., and Finch, S. The likelihood ratio test with the box-cox transformation for the normal mixture problem: Power and sample size study. *Communications in Statistics-simulation And Computation 33*, 3 (2004), 553–565.

[33] Pearson, K. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London Series A 185*, 1 (Jan. 1894), 71–110.

[34] Quandt, R., and Ramsey, J. Estimating mixtures of normal-distributions and switching regressions. *Journal of the American Statistical Association 73*, 364 (1978), 730–738.

[35] Rao, C. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society Series B-statistical Methodology 10*, 2 (1948), 159–203.

[36] Schork, N., Allison, D., and Thiel, B. Mixture distributions in human genetics research. *Statistical Methods in Medical Research 5*, 1 (Feb. 1996), 155–178.

[37] Schwarz, G. Estimating the dimension of a model. *Annals of Statistics 6*, 2 (1978), 461–464.

[38] Tan, W., and Chang, W. Some comparisons of method of moments and method of maximum likelihood in estimating parameters of a mixture of 2 normal densities. *Journal of the American Statistical Association 67*, 339 (1972), 702–708.

[39] Tashman, A., Gordon, D., Breslau, N., Yang, Z., Lee, J., Mendell, N., Johnson, E., Chase, G., Bierut, L., and Finch, S. Application of mixture modeling to test homogeneity in cross-sectional smoking data suggests evidence for mixture of faster and slower progression to regular smoking (manuscripts), 2006.

[40] Thode, H., Finch, S., and Mendell, N. Simulated percentage points for the null distribution of the likelihood ratio test for a mixture of 2 normals. *Biometrics 44*, 4 (Dec. 1988), 1195–1201.

[41] Turner, T. Estimating the propagation rate of a viral infection of potato plants via mixtures of regressions. *Journal of the Royal Statistical Society Series C-Applied Statistics 49* (2000), 371–384.

[42] Vuong, Q. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica 57*, 2 (Mar. 1989), 307–333.

[43] Wang, P., and Puterman, M. Mixed logistic regression models. *Journal of Agricultural, Biological, and Environmental Statistics 3*, 2 (June 1998), 175–200.

[44] Wang, P., Puterman, M., Cockburn, I., and Le, N. Mixed poisson regression models with covariate dependent rates. *Biometrics 52*, 2 (June 1996), 381–400.

[45] Young, G. *Essentials of statistical inference*. Cambridge University Press, 2005.