# Stony Brook University

# The Bivariate Normal Mixture Distribution:

# A Power Study of Bootstrap Test

A Dissertation Presented

by

**TINGTING HE**

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

December 2011

**Stony Brook University**

The Graduate School

**Tingting He**

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

**Nancy R. Mendell - Dissertation Advisor**
**Professor, Department of Applied Mathematics and Statistics**

**Stephen J. Finch - Chairperson of Defense**
**Professor, Department of Applied Mathematics and Statistics**

**Wei Zhu**
**Professor, Department of Applied Mathematics and Statistics**

**Yungtai Lo**
**Associate Professor, Department of Epidemiology and Population Health,**
**Albert Einstein College of Medicine, Yeshiva University**

This dissertation is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

**The bivariate normal mixture distribution: A power study of bootstrap test**

by

**Tingting He**

**Doctor of Philosophy**

in

**Applied Mathematics and Statistics**

Stony Brook University

**2011**

Univariate analysis has been commonly used in the studies of disease-related phenotypes. The need for multivariate analysis on linkage studies of complex disease/traits has grown with the increasing use of multiple phenotypes. This research extends the model for testing a single bivariate normal distribution versus a two component bivariate normal mixture distribution. Previous research restricted the two variables to have equal means and variance. Our study considers the more general case with no restrictions on these parameter values. Simulations are used to conduct a power study of bootstrap test under different combinations of parameter values. We note that samples of sample size n = 200 or more and an average mixture effect size of 2.5 or more is needed with mixing proportions between 0.1 and 0.9 to achieve reasonable power. Regression models of LRT statistic values are also fitted to calculate the type I error rate and power. Finally the bootstrap method is shown to be a reliable approach for evaluating the LRT statistics.

# Table of Contents

List of Figures……vi
List of Tables……vii
Acknowledgements……ix
**1 Introduction and Literature Review……1**
1.1 Introduction…….....1
1.2 Literature Review……..3
1.2.1 Estimation of Parameters……....3
1.2.2 Bootstrap Method…….....6
1.2.3 Testing the Number of Components of Mixture Model ……7
1.3 Summary of Our Study…… 9
**2 Maximum Likelihood Estimation for the Parameters of a Bivariate Normal Mixture……...……...11**
2.1 The Problem Statement……11
2.2 Bivariate Normal Mixture Distribution…….....13
2.2.1 The Definition.....……...13
2.2.2 Marginal Correlation between Two Random Variable…….....13
2.3 Maximum Likelihood Estimators ……...15
2.4 The EM Algorithm to Obtain the MLE…….....16
2.4.1 Description of Equations…….....16
2.4.2 Initial Values for the EM Algorithm…….20
2.5 Evaluation of the EM Algorithm…….22
2.5.1 Choosing the Number of Random Starting Points……22
2.5.2 Distribution of MLE of Mixing Proportion……..25
**3 Statistical Methods……27**
3.1 Likelihood Ratio Test……..27
3.2 Parametric Bootstrap Method of Inference……..29
**4 Power Study and Simulation Results……..…….. .31**
4.1 Estimating the Power of the Bootstrap Test……31

4.1.1 Data Simulation……31
4.1.2 Power Study of the Parametric Bootstrap Test…….33
4.2 Precision of the MLEs under the Alternative Hypothesis……...40
**5 Modeling LRT Statistics.....44**
5.1 Empirical Null Distribution.....44
5.1.1 Data Simulation.....44
5.1.2 The Null Distribution of LRT.....45
5.1.3 Modeling the LRT under the Null Hypothesis.....47
5.2 Distribution of LRT under the Alternative Hypothesis.....54
**6 Discussion and Conclusion……...63**

**Bibliography…….67**
**Appendix...……72**

# List of Figures

Figure 2.1 Scatter plot of two-component bivariate normal mixture distribution with parameter values $n$=400, $\mu_{11} = \mu_{12} = 0, \sigma_1 = \sigma_2 = 1, D_1 = 3, D_2 = 5, \rho = 0.5, \pi = 0.5$ .....13

Figure 2.2 Histogram of the MLE of the mixing proportion under the null hypothesis……26

Figure 3.1 Flow chart of p-vale calculation for bootstrap test……...30

Figure 4.1 Variable generation from a bivariate normal distribution…….32

Figure 4.2 Power results of the parametric bootstrap test for sample size $n = 50$ ……36

Figure 4.3 Power results of the parametric bootstrap test for sample size $n = 100$ ……..37

Figure 4.4 Power results of the parametric bootstrap test for sample size $n = 200$ ……..38

Figure 4.5 Power results of the parametric bootstrap test for sample size $n = 500$ ……..39

Figure 5.1 LRT calculation chart under the null hypothesis.....45

Figure 5.2 Normal Q-Q plot of $\sqrt[3]{LRT}$ with sample size n=500.....48

Figure 5.3 Comparison of Type I error rate using regressing model and bootstrap method ($\alpha = 0.05$).....53

Figure 5.4 Comparison of Type I error rate using regressing model and bootstrap method ($\alpha = 0.01$).....53

Figure 5.5 Selected Q-Q plot of $\log LRT$ with sample size n=100.....56

Figure 5.6 Comparison of fitted and observed power at significance level $\alpha = 0.05$ .....61

Figure 5.7 Comparison of fitted and observed power at significance level $\alpha = 0.01$.....61

Figure 5.8 Power difference between fitted and observed at $\alpha = 0.05$ .....62

Figure D.1 Power results of the parametric bootstrap test for sample size $n = 200$ ($\alpha = 0.01$)…… ……78

# List of Tables

Table 2.1 $\pi, \overline{D}, \rho$ and corresponding correlation for bivariate normal mixture $\rho_{X_1X_2}$ ……14

Table 2.2 Change in observed maximum log likelihood with increasing numbers of random starting points……24

Table 2.3 Skewness and median of the MLE of the mixing proportion……25

Table 4.1 The empirical power of the bootstrap test for sample size $n = 200$ ……..35

Table 4.2 Mean MLEs with the standard error and MSE when $\pi = 0.5$ in two component bivariate normal mixture model……...42

Table 5.1 Simulation result of LRT under the null hypothesis when correlation coefficient $\rho_0 = 0.0$ .....46

Table 5.2 Simulation result of LRT under the null hypothesis when correlation coefficient $\rho_0 = 0.5$ .....46

Table 5.3 Simulation result of LRT under the null hypothesis when correlation coefficient $\rho_0 = 0.9$ .....47

Table 5.4 ANOVA table for comparing the means of $\sqrt[3]{LRT}$ .....49

Table 5.5 Linear regression results for the mean of $\sqrt[3]{LRT}$ .....50

Table 5.6 Summary statistics of $\sqrt[3]{LRT}$ and Type I error rate.....50

Table 5.7 Critical values of $\sqrt[3]{LRT}$ based on regression model.....51

Table 5.8 Simulation results of LRT under the alternative hypothesis when $\pi = 0.5$, $\overline{D} = 2 \sim 2.5$ .....54

Table 5.9 ANOVA table for the mean of $\log LRT$ .....56

Table 5.10 Regression results for mean of $\log LRT$ under alternative hypothesis.....58

Table 5.11 Summary statistics of $\log LRT$ and fitted power with sample size $n$=200.....59

Table D.1 The empirical power of the bootstrap test for sample size $n = 50$ ……...75

Table D.2 The empirical power of the bootstrap test for sample size $n = 100$ ……....76

Table D.3 The empirical power of the bootstrap test for sample size $n = 500$ ……77

Table E.1 The model numbers at sample size $n$=50……...79

# ACKNOWLEDGMENT

I would like to express my sincere appreciations to all of you who assisted me during my Ph.D. study. First and Foremost I wish to thank my thesis advisor, Dr. Nancy R. Mendell, for continuously encouraging me and offering generous guidance on this thesis. Professor Mendell has been so supportive, kindness and share her knowledge and thoughtful ideas generously, which I will be benefit from for a long time to come.

I would also like to thank Dr. Stephen Finch for his advice and constant support. In addition, he provided important references to this thesis, which have been invaluable to my work. Thanks also go to Dr. Wei Zhu and Dr. Yungtai Lo. Their thoughtful comments gave me full support to my thesis.

I have very much enjoyed collaborating with my group members, Chengrui Huang, Minyoung Li, Donghyung Lee and Xiawei Tu. Their wonderful friendship support me during my study. I wish them all the best in their future careers.

To my dear husband, Yaqiu, thank you for everything you did for me and our family during my Ph.D. study. Finally, I must give special thanks to my little son, parents, brother, and parents-in-law for their constant love, support, encouragement and blessings. This thesis would not have been possible finished without all of their trust.

# Chapter 1

# Introduction and Review of Literature

## 1.1 Introduction

During the last decades finite mixture distributions have been a boost to the statistical modeling of heterogeneity. Numerous studies have been conducted on various practical and theoretical aspects. A finite mixture distribution is a weighted sum of several distributions. There are monographs written by Everitt and Hand (1981) [18], Titterington et al. (1985) [51], McLachlan and Basford (1988) [36], and Peel (2000) [38] on finite mixture model problems. Nowadays, normal mixture models are most commonly applied in many areas, such as economics (Arcidiacono and Jones (2003) [4]), human genetics (Schork et al. (1996) [47]), and astronomy (Kriessler and Beers (1997) [31]).

Pearson (1894) [39] first introduced mixture models in the late 1800's. He fitted a two component normal mixture distribution with unequal means and variances to Naples crabs body-length data provided by Weldon (1893) [52]. After that many researchers made their contributions to the univariate normal mixture model. In the genetic statistics field, the need for multivariate analysis in linkage studies of complex disease/traits has grown with the increasing number of multiple phenotypes recorded. He et al. (2006) [28]

used a multivariate normal mixture model to detect differential gene expression on microarray data. Consequently, bivariate models have been extensively studied as well. Chatterjee and Shih (2001) [9] proposed a bivariate model for the lifetime risk of two individuals and the age at onset between two susceptible individuals. Motivated by the work of Chatterjee and Shih, Wienke et al. (2003) [53] suggested a cure-mixture model to analyze bivariate time-to-event data obtained in twins. They used a one-stage estimation procedure to estimate the size of the susceptible fraction and the correlation between the frailties of the twin partners. Kang et al. (2004) [29] applied a bivariate t mixture distribution to Single-Nucleotide Polymorphism data and compared the proposed approach to the K-means algorithm.

Chuang and Mendell (1997) [8] developed a mixture likelihood approach to perform maximum likelihood estimation for the parameters of two component bivariate normal mixture distribution. They considered the special case in which the mixture parameter values for both variables were equal. Their hypotheses were stated as

$$H_0 : g(\underline{X};\underline{\theta_0}) = N(\underline{X};\begin{pmatrix}\mu_0\\ \mu_0\end{pmatrix},\begin{pmatrix}\sigma_0^{\,2} & \rho_0\sigma_0^{\,2}\\ \rho_0\sigma_0^{\,2} & \sigma_0^{\,2}\end{pmatrix}$$

Versus

$$H_1 : g(\underline{X};\underline{\theta_1}) = pN(\underline{X};\begin{pmatrix}\mu_1\\ \mu_1\end{pmatrix},\begin{pmatrix}\sigma_1^{\,2} & \rho_1\sigma_1^{\,2}\\ \rho_1\sigma_1^{\,2} & \sigma_1^{\,2}\end{pmatrix}) + (1-p)N(\underline{X};\begin{pmatrix}\mu_1 + D_1\sigma_1\\ \mu_1 + D_1\sigma_1\end{pmatrix},\begin{pmatrix}\sigma_1^{\,2} & \rho_1\sigma_1^{\,2}\\ \rho_1\sigma_1^{\,2} & \sigma_1^{\,2}\end{pmatrix}.$$

They concluded that the null distribution of the likelihood ratio test statistic converges to a chi-squared distribution with 2 degrees of freedom very slowly. In other words, one needs a sample size of 5000 or more for the $\chi_2^2$ critical values to be valid. They also suggested a "Pseudo chi-squared" distribution with $2 + 2.8/\sqrt{n}$ degrees of freedom for

the null distribution of the likelihood ratio test statistic of their special single bivariate normal distribution versus a two component bivariate normal mixture distribution.

## 1.2 Literature Review

### 1.2.1 Estimation of Parameters

The estimation of parameters for finite mixture distributions has been a research topic for a long time. Cohen (1967) [10] proposed to estimate the parameters of a mixture of two normal distributions with different means and standard deviations based on Pearson's (1894) [42] moment method. Day (1969) [11] discussed moment, minimum chi-square and Bayesian methods and found that all of them were inferior to maximum likelihood estimation in the multivariate normal mixture distribution case. Since the 1960s, maximum likelihood has been popular in the fitting of finite mixture models. Wolfe (1967,1970) [54] [55] used the computations of maximum likelihood to estimate the parameters of multivariate normal mixture distributions. Computational difficulties existed when he calculated maximum likelihood estimates. Some general iterative methods, such as Newton's method and quasi-Newton methods, were considered for computing an approximate solution for the likelihood equations. Dempster et al.(1977) [12] published the Expectation-Maximization (EM) algorithm for computing maximum likelihood estimates from incomplete data. Their algorithm led to simplification of the estimation of parameters. They noted that the finite mixture distribution is an example for which the EM directly applied. This is because the data from a mixture distribution can be treated as incomplete data. In this dissertation we focus on developing the EM

algorithm for our mixture case. Quandt and Ramsey (1978) [43] introduced a moment generating function estimator for finite mixtures of normal distributions. Render and Walker (1984) [45] discussed the theoretical and practical properties of the EM algorithm. The EM algorithm was found to be competitive as a general iterative method to find optimum maximum likelihood estimates. Titterington (1984a) [50] introduced a recursive procedure to estimate the parameters with incomplete data. Aitkin and Rubin (1985) [2] applied the EM algorithm to estimate the unknown parameters of the mixture distribution based on the prior distribution of the mixing proportion. McLachlan and Basford (1988) [36] demonstrated the use of the EM algorithm, especially for the univariate normal mixture case. Lindsay and Basak (1993) [32] proposed a moment method to obtain estimates that used fewer iterations to converge. Another approach was taken from Diebolt and Robert (1994) [13], who considered Bayesian sampling to solve the problem of estimation for finite mixture distributions. At the same time, software programs were developed. McLachlan et al. (1999) [37] developed EMMIX software to fit general multivariate normal mixture model without constraints on the means and covariance matrix and a t-distribution mixture model. Chen and Tan (2009) [7] applied the EM algorithm to get penalized maximum likelihood estimations (PMLE) and proved that PMLE is more accurate than regular MLE by a simulation study.

The issue of choosing a set of good initial parameter values plays a major role in the use of the EM algorithm. Fowlkes (1979) [23] proposed a Q-Q plot method for setting the initial values of parameters in a normal mixture distribution. Following the initializations suggested by Engelman and Hartigan (1969) [17], Thode et al. (1988) [49] set the initial values for the mixing proportion equal to 1/n, (n-1)/n, 1/4, and 3/4.

McLachlan and Basford (1988) [36] used posterior probabilities to cluster scatter plot of the observed values in the multivariate mixture case and obtained the initial values of parameters based on the clusters.

Finch et al. (1989) [22] proposed the following idea to get the initial values of a two component normal mixture distribution with common variance. They first obtained a random starting value for the mixing proportion, $\hat{\pi}^{(0)}$, from a uniform distribution with interval 0 to 1. Then they ordered the observed values of $X_i$, i.e. $x_{(1)} \le x_{(2)},...,\le x_{(n)}$ and divided the values into two groups (consisting of $x_{(1)}$ to $x_{([n\hat{\pi}^{(0)}])}$ in one group and $x_{([n\hat{\pi}^{(0)}]+1)}$ to $x_{(n)}$ in the other). Finally the initial values of the remaining parameters were

$$\hat{\mu}_1^{(0)} = \frac{\sum_{i=1}^{[n\hat{\pi}^{(0)}]} x_{(i)}}{[n\hat{\pi}^{(0)}]}, \quad \hat{\mu}_2^{(0)} = \frac{\sum_{i=[n\hat{\pi}^{(0)}]+1}^{n} x_{(i)}}{n-[n\hat{\pi}^{(0)}]}$$

$$\sigma^{2(0)} = \frac{[\sum_{i=1}^{[n\hat{\pi}^{(0)}]} (x_{(i)} - \hat{\mu}_1^{(0)})^2 + \sum_{i=[n\hat{\pi}^{(0)}]+1}^{n} (x_{(i)} - \hat{\mu}_2^{(0)})^2}{n-2}$$

Karlis and Xekalaki (2003) [30] examined the performance of eight methods which were defined in the literature on the choice of initial values of parameters for EM algorithm by conducting simulation study in a two component and a three component normal mixture with common variance case. They concluded that the moment estimates method was the superior method but it required more computation to converge.

## 1.2.2 Bootstrap Method

The bootstrap method is a modern re-sampling approach for making statistical inferences, which is also computer-intensive. It was first suggested by Efron (1979a, b, 1982) [14] [15] [16] in the late 1970s. Aitkin et al.(1981) [1] were the first to apply this resampling method to assess the null distribution of the likelihood ratio test statistic. The simplest form of the bootstrap is random resampling with replacement from the original sample observations $x_1, \ldots, x_n$ and then constructing a number of resample sets. Then the statistics of interest can be calculated from the new sample set. This is sometimes referred to as the nonparametric bootstrap. Alternatively one can generate parametric bootstrap sample sets from the parametric model under the null hypothesis. If the distribution of the parametric model is known, this approach is equivalent to the Monte Carlo procedure with the parameters replaced by MLEs of the parameters under the null hypothesis. McLachlan (1987) [35] used a parametric bootstrap to access the null distribution of log likelihood ratio test statistics for single univariate normal distribution versus two-component univariate normal mixture distribution with common variance. He found that the simulated power increased as the distance of component mean $\Delta = |\mu_1 - \mu_2| / \sigma$ and the number of bootstrap samples increase. Even when the number of bootstrap samples became indefinitely large, reasonable power could not be generated for a small distance between component means. Feng and McCulloch (1996) [20] noted that the bootstrap confidence procedure was better than the confidence region based on theoretical distribution of likelihood ratio test $\chi_1^2$ for solving the problem of the number of components in a univariate normal mixture with known mean, variance and unknown mixing proportion. To obtain a bootstrap confidence procedure was recommended from a

justified bootstrap likelihood ratio method. Schlattmann (2005) [46] compared parametric bootstrap and non parametric bootstrap methods to test the number of components of Poisson mixture regression model.

### 1.2.3 Testing the Number of Components of Mixture Model

The likelihood ratio test (LRT) is a general statistical approach to access the test of a hypothesis and decide the number of components in a finite mixture model problem.

Wolfe (1970, 1971) [55] [56] considered the theoretical asymptotic distribution of a revised likelihood ratio test statistic. That is, he suggested that the quantity $\lambda = -(\frac{2}{n})(n-k-2)\ln\frac{L_1}{L_2}$ has an approximate $\chi^2_{2k}$ distribution, where k is equal to the number of parameters. Based on Wolfe's result, Everitt (1981) [19] performed a simulation study to test the null hypothesis of the single multivariate normal distribution against the alternative of two component multivariate normal mixture distribution. According to his simulation results, the argument that Wolfe's $\lambda$ was asymptotically distributed as $\chi^2_{2k}$ appeared valid for those cases when the sample size n was at least ten times as large as the number of parameter values. Based on the power curves, he noted that the power of the test was pretty low when the generalized distance between the two components $\Delta = (\mu_1 - \mu_2)' \sum^{-1} (\mu_1 - \mu_2)$ was smaller than *2.5*. Hartigan (1985a,b) [26] [27] investigated the LRT of single normal distribution under the null hypothesis against two component normal mixture distribution under the alternative hypothesis with unknown mixing proportion, known common variance and mean. He found the asymptotic null distribution of -2 log λ was zero with probability 0.5 and, with the same

probability, was distributed as chi-squared with one degree of freedom. This could be expressed -2 log λ ~ $\frac{1}{2}\chi_0^2+\frac{1}{2}\chi_1^2$. Thode et al. (1988) [49] gave the simulated percentage points for the null distribution of the likelihood ratio test of the null hypothesis of single univariate normal distribution against the alternative of two component univariate normal mixture distribution with different component means and a common variance. They conducted 2500 simulation samples for each of 13 different sample sizes and found no obvious convergence to the asymptotic distribution $\chi_2^2$. Their simulation result supported Hartigan's (1977) [25] conjecture which the asymptotic null distribution of $-2\ln L_1/L_2$ was between $\chi_1^2$ and $\chi_2^2$. Quinn et al. (1987) [44] proved the asymptotical distribution of LRT was not a chi-square distribution. Feng and McCulloch (1994) [20] performed a simulation study on univariate normal mixture distributions with unequal variance. They concluded the reference distribution of LRT could be different when they chose $10^{-6}$,$10^{-10}$,$10^{-20}$ as the convergence criteria when estimating the variance, i.e. the distribution of $-2\log\lambda$ was between $\chi_4^2$ and $\chi_5^2$ when the criteria was $10^{-6}$; the distribution of $-2\log\lambda$ was between $\chi_5^2$ and $\chi_6^2$ when the criteria was $10^{-10}$. Berdai and Garel (1996) [6] found the asymptotic distribution of the LRT in testing a single versus two component univariate normal approached the $\chi_2^2$ distribution when the distance between component means was restricted. Lo et al. (2001) [33] documented that the asymptotic null distribution for the likelihood ratio statistic when testing whether a random sample was drawn from a $k_0$-component normal mixture distribution or from a $k_1$-component normal mixture distribution was a weighted sum of independent chi-squared random variables with one degree of freedom. They suggested the Kullback-

Leibler information criterion for estimation of these weights. More recently, Lo (2005) [34] performed simulation studies in the case of a single univariate normal versus a two component normal mixture distribution with unequal variances. He reported the power results at each parameter setting based on the null distribution of a modified likelihood ratio statistic that followed a weighted sum of 7 independent $\chi_1^2$ distribution. He compared it with bootstrap test and posterior predictive check test. There was no compelling evidence that showed any of these methods as being better than the other two. Mendell et al. (1991) [41] demonstrated the power and sample size requirements for LRT. Mendell et al. (1993) [40] compared the power of the likelihood ratio test and other tests to detect a two component mixture univariate normal distribution with different means but equal variance. They pointed out that the likelihood ratio test had the best power at most mixing proportion values.

Another approach based on clustering analysis to determine the number of components of mixture model was proposed by Fraley and Raftery (1998) [24]. They defined the model types by covariance structure and used the BIC value to decide the number of components.

## 1.3 Summary of our study

In genetics, the major gene model for a single quantitative trait has been a three component mixture distribution with each genotype having a separate mean. A pleiotropic locus, that is, a major gene that determines two traits would result in a bivariate normal mixture. In the special case where the same allele is dominant for

both of these quantitative traits, we would have a bivariate two component normal mixture.

This dissertation extends the work of Chuang and Mendell's (1997) [8] for the case of a single bivariate normal distribution versus two component bivariate normal mixture distribution. Instead of restricting the mixture for the two variables to have equal means, variance and shift parameters applicable to two correlated traits with a common cause, the variables have different scales. Hence, the mean and variance resulting in distance between the two components may not be equal. We will apply the EM algorithm to get maximum likelihood estimates for all unknown parameters and use the bootstrap technique to investigate whether the data is taken from a single bivariate normal distribution or from a two component bivariate normal mixture distribution. We will conduct a power study for this technique by using simulation and compare the power for different combinations of parameters. Then we will model the values of the LRT statistics and obtain some fitted model results.

In summary, we define our problem and present an EM algorithm to compute maximum likelihood estimates in Chapter 2. Evaluation of the algorithm is also discussed. In Chapter 3, the likelihood ratio test is discussed, and the parametric bootstrap technique is proposed for our problem. In Chapter 4, a simulation study is conducted to determine the power of bootstrap test. The precision of the maximum likelihood estimates is discussed as well. In Chapter 5, we attempt to model the null and alternative distribution of the LRT based on the empirical findings of our simulations. The dissertation is concluded with Chapter 6 with summary remarks and discussions on our study.

# Chapter 2

# Maximum Likelihood Estimation for the Parameters of A Bivariate Normal Mixture

## 2.1 The Problem Statement

Let $\underline{X_1}, \underline{X_2}, ..., \underline{X_n}$ be 2-dimensional random variables having probability density function (pdf) which we denote as $g(\underline{X})$. And $\underline{x_1}, \underline{x_2}, ..., \underline{x_n}$ denote the observed values of random variables. We commonly want to identify whether these variables on two traits come from an underlying two component normal mixture distribution or a single component normal distribution. Our null hypothesis is

$$H_0 : g(\underline{X}; \underline{\theta_0}) = N(\underline{X}; \underline{\mu_0}, \Sigma_0) \tag{2.1.1}$$

and the alternative hypothesis is

$$H_1 : g(\underline{X}; \underline{\theta_1}) = \pi N_1(\underline{X}; \underline{\mu_1}, \Sigma) + (1-\pi) N_2(\underline{X}; \underline{\mu_2}, \Sigma) \tag{2.1.2}$$

where $\underline{X} \in \mathbb{R}^2$, $\underline{X}$ denotes a 2×1 vector random variable with entries $X_j (j = 1,2)$, $\underline{\mu_0}$ denotes a 2×1 vector with entries $\mu_{0j} (j = 1,2)$, $\underline{\mu_1}$ denotes a 2×1 vector with entries $\mu_{1j} (j = 1,2)$, $\underline{\mu_2}$ denotes a 2×1 vector with entries $\mu_{2j} (j = 1,2)$, $\Sigma_0$ denotes a covariance matrix with entries $\sigma_{0jk} = \sigma_{0j}{}^2 (j = k)$, and $\sigma_{0jk} = \rho_0 \sigma_{0j} \sigma_{0k}$ $(j \neq k), j, k = 1,2$. In the

mixture distribution, $\Sigma$ denotes a common covariance matrix with entries $\sigma_{jk} = \sigma_j^{\,2}(j=k)$, $\sigma_{jk} = \rho\sigma_j\sigma_k(j \neq k), j,k = 1,2$, and $\pi$ is the mixing proportion or probability that $\underline{X}$ has the distribution $N_1(\underline{X};\underline{\mu_1},\Sigma)$ (as opposed to $N_2(\underline{X};\underline{\mu_2},\Sigma)$). $\underline{\theta_0}$ is the vector of unknown parameters under the null hypothesis; and $\underline{\theta_1}$ is the vector of unknown parameters under the alternative hypothesis. All of the unknown parameters are estimated from a random sample of size *n* with observed values $\underline{x_1}, \underline{x_2}, ..., \underline{x_n}$.

Without loss of generality, we assume $\mu_{21} \geq \mu_{11}$ and reparameterize with $D_1 = \frac{\mu_{21} - \mu_{11}}{\sigma_1}, D_2 = \frac{\mu_{22} - \mu_{12}}{\sigma_2}, D_1 > 0$. The problem then is to test

$$H_0 : g(\underline{X};\underline{\theta_0}) = N(\underline{X};\begin{pmatrix}\mu_{01}\\ \mu_{02}\end{pmatrix},\begin{pmatrix}\sigma_{01}^2 & \rho_0\sigma_{01}\sigma_{02}\\ \rho_0\sigma_{01}\sigma_{02} & \sigma_{02}^2\end{pmatrix}) \tag{2.1.3}$$

versus the unique alternative

$$H_1 : g(\underline{X};\underline{\theta_1}) = \pi N_1(\underline{X};\begin{pmatrix}\mu_{11}\\ \mu_{12}\end{pmatrix},\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix})$$

$$+(1-\pi)N_2(\underline{X};\begin{pmatrix}\mu_{11}+D_1\sigma_1\\ \mu_{12}+D_2\sigma_2\end{pmatrix},\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}) \tag{2.1.4}$$

In particular, there are five unknown parameters under the null hypothesis $H_0$; i.e. $\underline{\theta_0} = (\mu_{01}, \mu_{02}, \sigma_{01}, \sigma_{02}, \rho_0)$. There are eight unknown parameters under the alternative hypothesis $H_1$; i.e. $\underline{\theta_1} = (\pi, \mu_{11}, \mu_{12}, D_1, D_2, \sigma_1, \sigma_2, \rho)$, where $0 < \pi < 1$, $D_1 \geqslant 0$, $-\infty < D_2 < \infty$, $\sigma_1 > 0$, $\sigma_2 > 0$, $-1 < \rho < 1$.

## 2.2 Bivariate Normal Mixture Distribution

### 2.2.1 Definition

Under the alternative hypothesis, we allow unequal values for the means and variances of the two variables. Therefore, the standardized differences between the two component means for the two variables might not be equal. A standardized example of one of our alternative hypotheses could be

$$\mu_{11} = \mu_{12} = 0,\ \sigma_1 = \sigma_2 = 1,$$

$$D_1 = 3,\ D_2 = 5,\ \rho = 0.5,\ \pi = 0.5$$

This bivariate normal mixture distribution is shown in Figure 2.1.

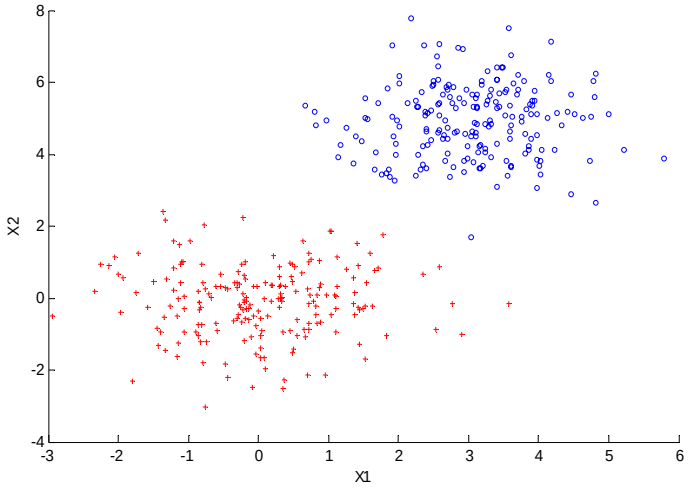



Figure 2.1 Scatter plot of two-component bivariate normal mixture distribution with parameter values $n = 400, \mu_{11} = \mu_{12} = 0, \sigma_1 = \sigma_2 = 1, D_1 = 3, D_2 = 5, \rho = 0.5, \pi = 0.5$.

### 2.2.2 Correlation Between two Random Variables

To investigate the relationship between two random variables taken from a two-

component bivariate normal mixture distribution, we calculate the correlation based on the methods demonstrated in Appendix A.

Under different average values of mixture effect size $D = \frac{D_1 + D_2}{2}$, where $D_i (i = 1,2)$ are defined in equation (2.1.4), we consider combinations of mixing proportion $\pi = 0.5$ or 0.9 with two different values of within component correlation, $\rho = 0.0$ or 0.9. Table 2.1 illustrates the values of the correlation of different combinations of within component $\rho$ and average mixture effect size $\overline{D}$. As shown in the table, for each mixing proportion value, the correlation between the two variables increases as the average effect size increases when the within component correlation $\rho = 0$. However, the correlation decreases as the average mixture effect size increases when the within component correlation $\rho$ is very large. We also note that in the case of even a small average mixture effect size $\overline{D} = 2$, the correlation between two random variables $\rho_{X_1 X_2}$ is greater than 0.25.

Table 2.1 – $\pi, \overline{D}, \rho$ and corresponding correlation for bivariate normal mixture $\rho_{X_1 X_2}$

| Mixing proportion $\pi$ | Within component correlation $\rho$ | Average effect size $\overline{D}$ | Correlation $\rho_{X_1 X_2}$ |
|---|---|---|---|
| 0.5 | 0.0 | 2 | 0.5 |
| | | 2.5 | 0.59 |
| | | 3 | 0.63 |
| | 0.9 | 2 | 0.95 |
| | | 2.5 | 0.94 |
| | | 3 | 0.92 |
| 0.9 (0.1) | 0.0 | 2 | 0.26 |
| | | 2.5 | 0.34 |
| | | 3 | 0.40 |
| | 0.9 | 2 | 0.93 |
| | | 2.5 | 0.92 |
| | | 3 | 0.89 |

## 2.3 Maximum Likelihood Estimators

Under the null hypothesis $H_0$, if a 2-dimensional data set of $n$ values $\underline{x_1}, \underline{x_2}, ..., \underline{x_n}$ is a random sample from a population with single bivariate normal distribution function $g(\underline{X}; \underline{\theta_0})$, the likelihood function is defined as

$$L(\underline{\theta_0}; \underline{X}) = L(\underline{\theta_0}; \underline{x_1}, \underline{x_2}, ..., \underline{x_i}) = \prod_{i=1}^{n} g(\underline{x_i}; \underline{\theta_0})$$
$$= \prod_{i=1}^{n} \frac{1}{2\pi\sigma_{01}\sigma_{02}\sqrt{1-\rho_0^2}} \exp\{-\frac{1}{2(1-\rho_0^2)}[(\frac{x_{i1}-\mu_{01}}{\sigma_{01}})^2 - 2\rho(\frac{x_{i1}-\mu_{01}}{\sigma_{01}})(\frac{x_{i2}-\mu_{02}}{\sigma_{02}}) + (\frac{x_{i2}-\mu_{02}}{\sigma_{02}})^2]\}$$

(2.3.1)

The maximized likelihood function can be written as $L(\hat{\underline{\theta_0}}; \underline{X}) = \prod_{i=1}^{n} g(\underline{x_i}; \hat{\underline{\theta_0}})$, and the maximum likelihood estimator of the unknown parameters $\hat{\underline{\theta_0}}$ can be easily computed based on the above equation. The detailed calculation is provided in Appendix B.

Under the alternative hypothesis $H_1$, if a 2-dimensional data set of $n$ values $\underline{x_1}, \underline{x_2}, ..., \underline{x_n}$ is a random sample from a population with two-component bivariate normal mixture distribution function $g(\underline{X}; \underline{\theta_1})$, the likelihood function is

$$L(\underline{\theta_1}; \underline{X}) = L(\underline{\theta_1}; \underline{x_1}, \underline{x_2}, ..., \underline{x_i}) = \prod_{i=1}^{n} g(\underline{x_i}; \underline{\theta_1}) \qquad (2.3.2)$$

So the maximized likelihood function is $L(\hat{\underline{\theta_1}}; \underline{X}) = \prod_{i=1}^{n} g(\underline{x_i}; \hat{\underline{\theta_1}})$ and the maximum likelihood estimator of the unknown parameters $\hat{\underline{\theta_1}}$ can be computed by the EM

algorithm method, which is described in section 2.4.

## 2.4 The EM Algorithm to Obtain the MLE

### 2.4.1 Description of Equations

Now we obtain the MLE of the parameters $\underline{\theta} = (\pi, \underline{\mu_1}, \underline{\mu_2}, \Sigma)$ for the bivariate mixture defined in section 2.3. We first convert the problem into an incomplete-data problem by introducing a new indicator variable $\underline{z_i}$ as a "missing variable" and set $\underline{Y} = (\underline{Z}, \underline{X})$ as the sample observe values $\underline{y_1} = (\underline{z_1}, \underline{x_1}), \underline{y_2} = (\underline{z_2}, \underline{x_2}), ..., \underline{y_n} = (\underline{z_n}, \underline{x_n})$, where $\underline{z_i}$ is a 2-dimensional vector with $z_{mi} = (\underline{z_i})_m = 1$ or 0. The indicator $z_{mi} = 1$ if the $i\,th$ observation comes from the $m\,th$ component of the mixture normal model while $z_{mi} = 0$ if the $i\,th$ observation does not come from the $m\,th$ component of the mixture normal model $(i = 1,2,...,n)$. Then $\underline{z_i} = (z_{1i}, z_{2i})^T$, $\underline{z_1}, ..., \underline{z_n}$ is a sample from a multinomial distribution consisting of one draw from two categories with the respective probabilities $\pi$ and $1 - \pi$; i.e. $z_1, ..., z_n \overset{iid}{\sim} Mult_2(1, \pi_m)$, $\pi_m = (\pi, 1-\pi)^T$. We are interested in maximizing the incomplete-data likelihood function (2.3.2) with the observed data generated given the model parameters $\underline{\theta}$. This can be solved by Expectation Maximization (EM) algorithm.

The log function of the complete-data likelihood for $\underline{\theta}$ is given by:

$$\log L_c(\underline{\theta}) = \log p(\underline{x}, \underline{z}; \underline{\theta})$$
$$= \log \prod_{i=1}^{n} p(\underline{x_i}, \underline{z_i}; \underline{\theta})$$
$$= \log \prod_{i=1}^{n} \{[p(\underline{x_i} \mid z_{1i} = 1; \underline{\theta}) p(z_{1i} = 1)]^{z_{1i}} [1\,1 p(\underline{x_i} \mid z_{2i} = 1; \underline{\theta}) p(z_{2i} = 1)]^{z_{2i}}\}$$
$$= \sum_{i=1}^{n} \{z_{1i}[\log \pi + \log N_1(\underline{x_i}; \underline{\mu_1}, \Sigma)] + z_{2i}[\log(1-\pi) + \log N_2(\underline{x_i}; \underline{\mu_2}, \Sigma)]\}$$

(2.4.1)

The EM algorithm starts from an initial value $\underline{\theta}^{(0)}$ and performs two-step iterations until convergence to a maximum. During E-step, assume $\underline{\theta}^{(k)}$ is the value of $\underline{\theta}$ which is calculated after $k$ iterations of the EM algorithm. Then on the $(k+1)\,th$ iteration, given the observed data $\underline{x}$ and the current value $\underline{\theta}$, the conditional expectation of $\log L_c(\underline{\theta})$ can be written as

$$Q(\underline{\theta}; \underline{\theta}^{(k)}) = E_{\underline{\theta}^{(k)}}[\log L_c(\underline{\theta}) \mid \underline{x}] \tag{2.4.2}$$

The current conditional expectation of $Z_{mi}$ given the observed data $\underline{x}$, where $Z_{mi}$ is the random variable corresponding to $z_{mi}$, is calculated as:

$$E_{\underline{\theta}^{(k)}}(z_{mi} \mid \underline{x}) = p_{\underline{\theta}^{(k)}}(z_{mi} = 1 \mid \underline{x_i}; \underline{\theta})$$
$$= \frac{p_{\underline{\theta}^{(k)}}(\underline{x_i} \mid z_{mi} = 1; \underline{\theta}) p(z_{mi} = 1)}{p_{\underline{\theta}^{(k)}}(\underline{x_i} \mid z_{1i} = 1; \underline{\theta}) p(z_{1i} = 1) + p_{\underline{\theta}^{(k)}}(\underline{x_i} \mid z_{2i} = 1; \underline{\theta}) p(z_{2i} = 1)} \tag{2.4.3}$$

We define

$$\tau_1(\underline{x_i}; \underline{\theta}^{(k)}) = E_{\underline{\theta}^{(k)}}(z_{1i} \mid \underline{x}) = \frac{\pi N_1(\underline{x_i}; \underline{\mu_1}, \Sigma)}{\pi N_1(\underline{x_i}; \underline{\mu_1}, \Sigma) + (1-\pi) N_2(\underline{x_i}; \underline{\mu_2}, \Sigma)} \tag{2.4.4}$$

$$\tau_2(\underline{x}_i;\underline{\theta}^{(k)}) = E_{\underline{\theta}^{(k)}}(z_{2i} \mid \underline{x}) = \frac{(1-\pi)N_2(\underline{x}_i;\underline{\mu}_2,\Sigma)}{\pi N_1(\underline{x}_i;\underline{\mu}_1,\Sigma)+(1-\pi)N_2(\underline{x}_i;\underline{\mu}_2,\Sigma)} \tag{2.4.5}$$

Since the unobservable data $z_{mi}(m=1,2)$ is a linear term in the function $\log L_c(\underline{\theta})$, the conditional expectation of $\log L_c(\underline{\theta})$ can be further computed as

$$Q(\underline{\theta};\underline{\theta}^{(k)}) = \sum_{i=1}^{n}\{\tau_1(\underline{x}_i;\underline{\theta}^{(k)})[\log\pi + \log N_1(\underline{x}_i;\underline{\mu}_1,\Sigma)] + \tau_2(\underline{x}_i;\underline{\theta}^{(k)})[\log(1-\pi) + \log N_2(\underline{x}_i;\underline{\mu}_2,\Sigma)]\}$$

(2.4.6)

Then we come to the "M-step". The M-step on the $(k+1)\,th$ iteration requires the global maximization of $Q(\underline{\theta};\underline{\theta}^{(k)})$ with respect to $\underline{\theta}$ over the whole parameter space to yield the updated estimate $\underline{\theta}^{(k+1)}$. In our problem we maximize it with respect to $\pi, \underline{\mu}_m, \Sigma$. Differentiating equation (2.4.1) with respect to $\underline{\mu}_1$ we get

$$\begin{aligned}
\frac{\partial Q(\underline{\theta};\underline{\theta}^{(k)})}{\partial \underline{\mu}_1} &= \sum_{i=1}^{n}\tau_1(\underline{x}_i;\underline{\theta}^{(k)})\frac{\partial}{\partial \underline{\mu}_1}\log N_1(\underline{x}_i;\underline{\mu}_1,\Sigma) \\
&= \sum_{i=1}^{n}\tau_1(\underline{x}_i;\underline{\theta}^{(k)})\frac{\partial}{\partial \underline{\mu}_1}\log\{\frac{1}{2\pi|\Sigma|^{1/2}}\exp[-\frac{1}{2}(\underline{x}_i-\underline{\mu}_1)^T\Sigma^{-1}(\underline{x}_i-\underline{\mu}_1)]\} \\
&= \sum_{i=1}^{n}\tau_1(\underline{x}_i;\underline{\theta}^{(k)})\{-\frac{1}{2}\frac{\partial}{\partial \underline{\mu}_1}[(\underline{x}_i-\underline{\mu}_1)^T\Sigma^{-1}(\underline{x}_i-\underline{\mu}_1)]\} \\
&= \sum_{i=1}^{n}\tau_1(\underline{x}_i;\underline{\theta}^{(k)})\{\frac{1}{2}(\underline{x}_i-\underline{\mu}_1)^T[(\Sigma^{-1}+(\Sigma^{-1})^T]\} \\
&= \sum_{i=1}^{n}\tau_1(\underline{x}_i;\underline{\theta}^{(k)})(\underline{x}_i-\underline{\mu}_1)^T\Sigma^{-1}
\end{aligned} \tag{2.4.7}$$

We set equation (2.4.7) to 0. Then the updated value of the estimate of the first

component means $\underline{\mu}_1$ is

$$\underline{\mu}_1^{(k+1)} = \frac{\sum_{i=1}^{n} \tau_1(\underline{x}_i; \underline{\theta}^{(k)}) \underline{x}_i}{\sum_{i=1}^{n} \tau_1(\underline{x}_i; \underline{\theta}^{(k)})} \tag{2.4.8}$$

Similarly, the updated value of the estimate of the second component means $\underline{\mu}_2$ is

$$\underline{\mu}_2^{(k+1)} = \frac{\sum_{i=1}^{n} \tau_2(\underline{x}_i; \underline{\theta}^{(k)}) \underline{x}_i}{\sum_{i=1}^{n} \tau_2(\underline{x}_i; \underline{\theta}^{(k)})} \tag{2.4.9}$$

Differentiating equation (2.4.1) with respect to $\sum^{-1}$ we get

$$\begin{aligned}
&\frac{\partial Q(\underline{\theta}; \underline{\theta}^{(k)})}{\partial \Sigma^{-1}} = \sum_{i=1}^{n} \{\tau_1(\underline{x}_i; \underline{\theta}^{(k)}) \frac{\partial}{\partial \Sigma^{-1}} \log N_1(\underline{x}_i; \underline{\mu}_1, \Sigma) + \tau_2(\underline{x}_i; \underline{\theta}^{(k)}) \frac{\partial}{\partial \Sigma^{-1}} \log N_2(\underline{x}_i; \underline{\mu}_2, \Sigma)\} \\
&= \sum_{i=1}^{n} \tau_1(\underline{x}_i; \underline{\theta}^{(k)}) \frac{\partial}{\partial \Sigma^{-1}} \log \{\frac{1}{2\pi |\Sigma|^{1/2}} \exp[-\frac{1}{2}(\underline{x}_i - \underline{\mu}_1)^T \Sigma^{-1} (\underline{x}_i - \underline{\mu}_1)]\} \\
&+ \sum_{i=1}^{n} \tau_2(\underline{x}_i; \underline{\theta}^{(k)}) \frac{\partial}{\partial \Sigma^{-1}} \log \{\frac{1}{2\pi |\Sigma|^{1/2}} \exp[-\frac{1}{2}(\underline{x}_i - \underline{\mu}_2)^T \Sigma^{-1} (\underline{x}_i - \underline{\mu}_2)]\} \\
&= \sum_{i=1}^{n} \tau_1(\underline{x}_i; \underline{\theta}^{(k)}) [\frac{1}{2}\Sigma - \frac{1}{2}(\underline{x}_i - \underline{\mu}_1)(\underline{x}_i - \underline{\mu}_1)^T] + \sum_{i=1}^{n} \tau_2(\underline{x}_i; \underline{\theta}^{(k)}) [\frac{1}{2}\Sigma - \frac{1}{2}(\underline{x}_i - \underline{\mu}_2)(\underline{x}_i - \underline{\mu}_2)^T]
\end{aligned} \tag{2.4.10}$$

Similarly we set equation (2.4.10) to 0. The updated value of the estimate of the component common covariance matrix $\sum$ is

$$\Sigma^{(k+1)} = \frac{\sum_{i=1}^{n} \tau_1(\underline{x}_i; \underline{\theta}^{(k)})(\underline{x}_i - \underline{\mu}_1^{(k)})(\underline{x}_i - \underline{\mu}_1^{(k)})^T + \sum_{i=1}^{n} \tau_2(\underline{x}_i; \underline{\theta}^{(k)})(\underline{x}_i - \underline{\mu}_2^{(k)})(\underline{x}_i - \underline{\mu}_2^{(k)})^T}{\sum_{i=1}^{n} \tau_1(\underline{x}_i; \underline{\theta}^{(k)}) + \sum_{i=1}^{n} \tau_2(\underline{x}_i; \underline{\theta}^{(k)})} \tag{2.4.11}$$

Finally, by differentiating equation (2.4.6) with respect to $\pi$ we get

$$\frac{\partial Q(\underline{\theta};\underline{\theta}^{(k)})}{\partial \pi} = \frac{1}{\pi}\sum_{i=1}^{n}\tau_1(\underline{x}_i;\underline{\theta}^{(k)}) + \frac{1}{\pi - 1}\sum_{i=1}^{n}\tau_2(\underline{x}_i;\underline{\theta}^{(k)}) \tag{2.4.12}$$

We also set equation (2.4.12) to 0. The updated value of the estimate of the mixing proportion $\pi$ is

$$\pi^{(k+1)} = \frac{\sum_{i=1}^{n}\tau_1(\underline{x}_i;\underline{\theta}^{(k)})}{n} \tag{2.4.13}$$

We perform the E-step and the M-step iteratively until the difference $l(\underline{\theta}^{(k+1)}) - l(\underline{\theta}^{(k)})$ reduces below an arbitrarily small amount (here $10^{-5}$).

## 2.4.2 Initial Values for the EM Algorithm

To ensure convergence to the global maximum, it is critical to choose a set of good initial values. Usually we first choose an initial value for the mixing proportion $\pi^{(0)}$, and then use it to decide the values for the other unknown parameters. Since our sample observed values are two dimensional, we may not simply order the observed values as suggested for the univariate case. In this dissertation, we propose the following procedure to compute the initial parameter values. First we select a random value from a uniform distribution with interval 0 to 1 and set this value as initial value for the mixing proportion $\pi$, denoted by $\hat{\pi}^{(0)}$. Then we select $n$ random values $U_i (i = 1,2,...,n)$ from a uniform distribution with interval 0 to 1. If $U_i \le \hat{\pi}^{(0)}$, the value of $\underline{X}_i$ is assigned to the first component of the mixture distribution, $\underline{A}$; otherwise the first variable $\underline{X}_i$ is

assigned to the second component of the mixture distribution, $\underline{B}$. Finally we have two groups of $\underline{X_i}$ as

$$\underline{A} = (\underline{A_1}, \underline{A_2}, ..., \underline{A_{n_1}})$$
$$\underline{B} = (\underline{B_1}, \underline{B_2}, ..., \underline{B_{n-n_1}})$$

Using equations given in Appendix A, the initial estimates of the remaining parameters are computed as follows:

$$\hat{\mu}_{11}^{(0)} = \frac{\sum_{i=1}^{n_1} A_{i1}}{n_1}, \hat{\mu}_{12}^{(0)} = \frac{\sum_{i=1}^{[n_1]} A_{i2}}{n_1}$$

$$\hat{\mu}_{21}^{(0)} = \frac{\sum_{i=1}^{n-n_1} B_{i1}}{n-n_1}, \hat{\mu}_{22}^{(0)} = \frac{\sum_{i=1}^{n-n_1} B_{i2}}{n-n_1}$$

$$\hat{D}_1^{(0)} = \hat{\mu}_{21}^{(0)} - \hat{\mu}_{11}^{(0)}, \hat{D}_2^{(0)} = \hat{\mu}_{22}^{(0)} - \hat{\mu}_{12}^{(0)} \qquad (2.4.14)$$

and

$$\hat{\sigma}_1^{2(0)} = \frac{\sum_{i=1}^{n_1} (A_{i1} - \hat{\mu}_{11}^{(0)})^2 + \sum_{i=n}^{n-n_1} (B_{i1} - \hat{\mu}_{21}^{(0)})^2}{n-2},$$

$$\hat{\sigma}_2^{2(0)} = \frac{\sum_{i=1}^{n_1} (A_{i2} - \hat{\mu}_{12}^{(0)})^2 + \sum_{i=1}^{n-n_1} (B_{i1} - \hat{\mu}_{22}^{(0)})^2}{n-2}$$

$$\hat{\rho}^{(0)} = \rho_{X_1X_2} \sqrt{[1+\pi(1-\pi)\hat{D}_1^{(0)}][1+\pi(1-\pi)\hat{D}_2^{(0)}]} - \pi(1-\pi)\hat{D}_1^{(0)}\hat{D}_2^{(0)}$$

(2.4.15)

In order to get our unique mixture distribution, we restrict $\hat{D}_1^{(0)}$ to be positive, i.e. if $\hat{D}_1^{(0)} < 0$ , then $\hat{\pi}_0 = 1 - \hat{\pi}_0$ , $\hat{\mu}_{11}^{(0)} = \hat{\mu}_{21}^{(0)}$ , $\hat{\mu}_{12}^{(0)} = \hat{\mu}_{22}^{(0)}$ , $\hat{D}_1^{(0)} = -\hat{D}_1^{(0)}$ , $\hat{D}_2^{(0)} = -\hat{D}_2^{(0)}$ .

## 2.5 Evaluation of the EM Algorithm

### 2.5.1 Choosing the Number of Random Starting Points

One of the difficulties is that we can only obtain local maxima, but we cannot guarantee that we have found the global maximum in EM algorithm even if we had good initial values. One solution is to generate several initial values and choose the "best" one which most likely to be global maximum. We also need to determine the optimal number of sets of initial values needed, since there is a trade-off between confidence that the global maximum has been found and computation resources. For a given sample, we consider a number of sets of random starting points and compute the maximum log likelihood value for each set of random starting points. We then calculate the difference between maximum log likelihood values among these sets of random starting points. The maximum log likelihood will be considered as the 'global' maximum once the differences between a pair of converge to an arbitrary small amount, i.e. $10^{-5}$ . This can be an effective approach to ensure convergence to global maximum. In order to find the optimal number of random starting points in this dissertation, we generate 20 replications each with sample size $n = 200$ from bivariate normal mixture distribution. The parameter

values of the generating bivariate normal mixture are $\pi = 0.5, \mu_{11} = \mu_{12} = 0, \rho = 0,$ $D_1 = D_2 = 2, \sigma_1 = \sigma_2 = 1$. For each replicate, we consider 5, 20, 45, 60, 120 and 240 sets of random starting points. We then (1) obtain a total of 240 values of maximum log likelihood. We define $l_{\max(i)}$ to denote the maximum value over the values from 1st to $i^{th}$ set of random starting points. For instance, $l_{\max(30)}$ is the maximum value over 30 values of maximum log likelihood calculated from 1st to $30^{th}$ set of random starting points. (2) Calculate the difference between maximum value of a set size and the next smallest set size (ie. $l_{\max(30)} - l_{\max(20)}$ ). We consider $l_{\max(i)}$ being the global maximum log likelihood if $l_{\max(j)} - l_{\max(i)} \leq 10^{-5} (j > i)$ for all replicates.

Table 2.2 shows the change in the differences of maximum log likelihoods for specified numbers of random starting points, where the specified numbers are 5, 10, 20, 30, 45, 60, 120 and 240. When the number sets of random points is 60 and 120, the differences in the maximum log likelihoods appear to be consistently a very small number (less than $10^{-5}$) in all replicates. Hence, we are successful getting the global maximum with 60 or above random starting points. For this reason we choose 60 as the optimal number sets of random starting points.

Table 2.2 Change in observed maximum log likelihood with increasing numbers of random starting points

| replicates | Difference between random starting points | | | | | | |
|---|---|---|---|---|---|---|---|
| | $l_{max(10)} - l_{max(5)}$ | $l_{max(20)} - l_{max(10)}$ | $l_{max(30)} - l_{max(20)}$ | $l_{max(45)} - l_{max(30)}$ | $l_{max(60)} - l_{max(45)}$ | $l_{max(120)} - l_{max(60)}$ | $l_{max(240)} - l_{max(120)}$ |
| 1 | 0 | 1.4E-06 | 0 | 7.6E-06 | 0 | 0 | 0 |
| 2 | 6E-07 | 7.5E-05 | 0 | 0 | 0 | 0 | 0 |
| 3 | 2.4E-06 | 2.4E-06 | 0 | 0 | 0 | 5E-07 | 3.56E-05 |
| 4 | 5E-07 | 0 | 3.87E-05 | 0 | 0 | 0 | 1E-07 |
| 5 | 2.5E-06 | 0 | 0 | 1E-07 | 0 | 1E-07 | 1.94E-05 |
| 6 | 9.56E-05 | 0.709207 | 0 | 0 | 0 | 0 | 4.6E-06 |
| 7 | 0.099067 | 2.718332 | 3E-07 | 1E-07 | 0 | 5E-07 | 0 |
| 8 | 3.6E-06 | 3.27E-05 | 0 | 0 | 0 | 0 | 0 |
| 9 | 0 | 0.000133 | 0 | 0 | 0.000107 | 0 | 0 |
| 10 | 0 | 0.000248 | 0 | 0 | 0 | 0 | 0 |
| 11 | 1.746202 | 0 | 0 | 0 | 0 | 0 | 1E-06 |
| 12 | 0 | 0 | 0 | 0 | 1E-06 | 2E-06 | 0 |
| 13 | 2.2E-06 | 0 | 0 | 0.000124 | 0 | 2.01E-05 | 0 |
| 14 | 2.2E-06 | 0.862634 | 0 | 1E-07 | 0 | 0 | 0 |
| 15 | 1.2E-06 | 0.996454 | 0 | 0 | 0 | 0 | 1.5E-05 |
| 16 | 0.937037 | 0 | 0 | 0 | 0 | 0 | 0 |
| 17 | 0 | 4E-07 | 0 | 0 | 0 | 1.26E-05 | 0 |
| 18 | 0 | 0 | 0 | 0 | 0 | 0 | 7.09E-05 |
| 19 | 4.225813 | 0 | 0 | 0 | 0 | 4.31E-05 | 0 |
| 20 | 0.79643 | 0 | 4.3E-06 | 2.4E-06 | 0 | 7.2E-06 | 8.3E-06 |

## 2.5.2 Distribution of MLE of Mixing Proportion

Another way to evaluate the EM algorithm is to investigate the null distribution of the MLEs. In particular we would expect the null distribution of mixing proportion to be symmetric about $\pi = 0.5$, possibly with this modes at $\varepsilon$ and $1-\varepsilon$.

We generate 1000 samples under the null hypothesis at various sample size ($n$=50, 100, 200, 500, 1000). In this evaluation procedure, three different values for correlation coefficient $\rho_0$ (0, 0.5 and 0.9) are considered with $\mu_{01}= \mu_{02} = 0$, $\sigma_{01} = \sigma_{02} = 1$ in our bivariate normal distribution case. Table 2.3 shows our findings on these MLEs. The skewness is close to 0 and the median of sample is very close to 0.5 for each sample size. The standard error of skewness is 0.08. Thus we conclude that the distribution of the MLE of the mixing proportion is symmetric. Figure 2.2 shows the histogram of the MLE of the mixing proportion with sample size $n$=500 and $n$=1000. Of interest in the distribution of the MLE of the mixing proportion for n=1000 and $\rho_0 = 0$. It appears to be symmetrical about 0.5 which is expected ( $\chi_5^2 = 1.14, p - value = 0.951$).

Table2.3 Skewness and median of the MLE of the mixing proportion

| Sample size | $\rho_0 = 0$ | | $\rho_0 = 0.5$ | | $\rho_0 = 0.9$ | |
|---|---|---|---|---|---|---|
| | skewness | median | skewness | median | skewness | median |
| $n = 50$ | -0.0339 | 0.5074 | -0.0159 | 0.5097 | 0.0542 | 0.4857 |
| $n = 100$ | -0.0150 | 0.5088 | 0.0070 | 0.5021 | -0.0241 | 0.5045 |
| $n = 200$ | 0.0415 | 0.4824 | -0.0434 | 0.5129 | -0.0152 | 0.4998 |
| $n = 500$ | 0.0649 | 0.4775 | -0.0543 | 0.5228 | 0.0082 | 0.4934 |
| $n = 1000$ | -0.0203 | 0.5015 | 0.0103 | 0.4984 | -0.0350 | 0.5264 |

Note: The MLE of the mixing proportion are based on 1000 replicates for each sample size with parameter value.

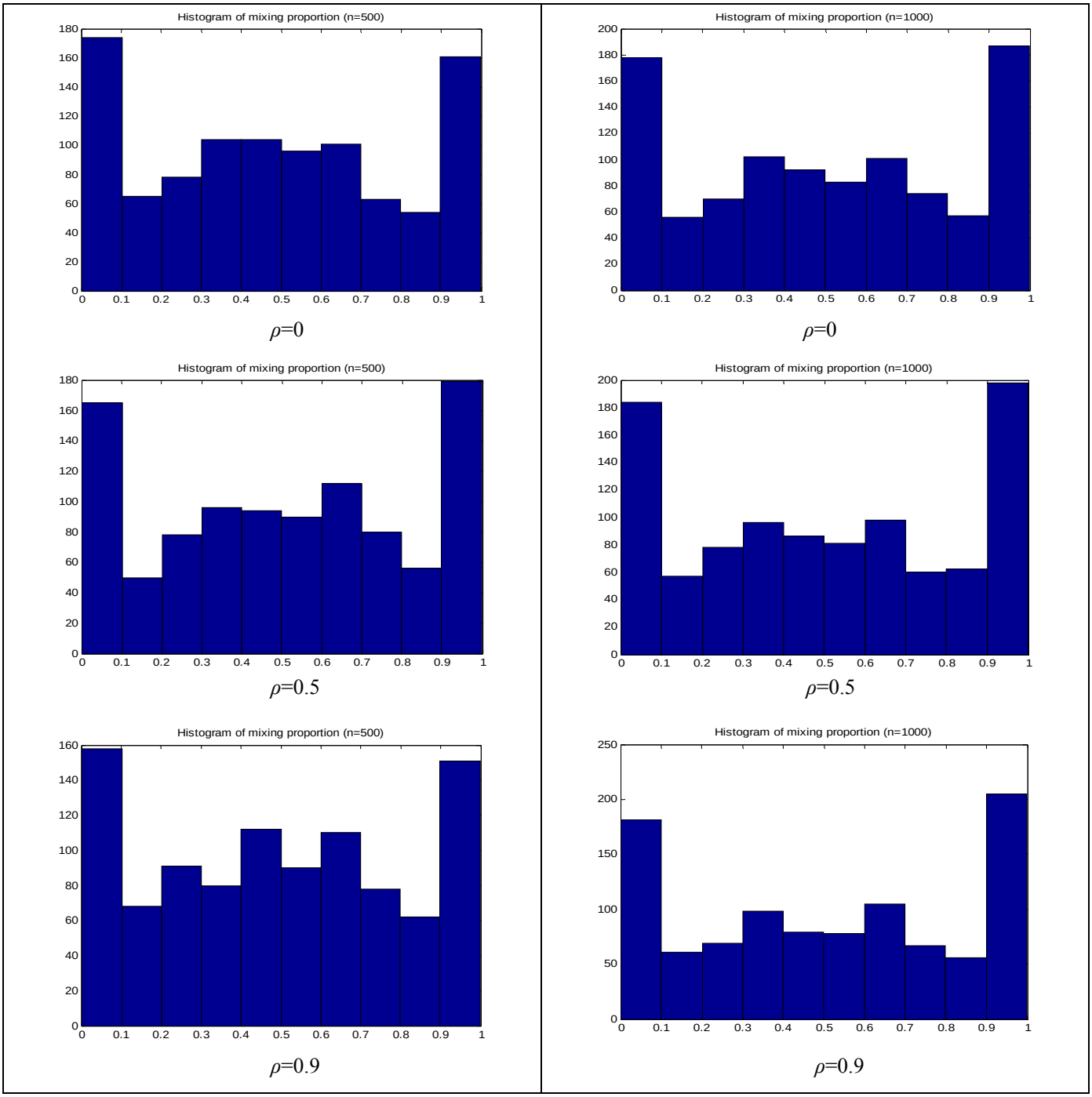



Figure 2.2 Histogram of the MLE of the mixing proportion under the null hypotheses

# Chapter 3

# Statistical Methods

## 3.1 Likelihood Ratio Test

The likelihood ratio test (LRT) is a general method which is designed to test model assumptions and is related to the maximum likelihood estimators. The LRT has been widely applied to statistics based on the maximum likelihood estimation.

The likelihood ratio test compares the likelihood of the data under the alternative hypothesis to the likelihood of the data under the null hypothesis. First we consider the likelihood ratio (LR). Let $\Theta$ denote the entire parameter space and define $\sup_{\Theta} L(\underline{\theta};\underline{X})$ and $\sup_{\Theta_0} L(\underline{\theta};\underline{X})$ as the maxima of the likelihood for each of the hypotheses. The likelihood ratio $\lambda$ for testing $H_0:\ \underline{\theta} \in \Theta_0$ versus $H_1:\ \underline{\theta} \in \Theta_0^c$ is

$$\lambda(\underline{X}) = \frac{\sup_{\Theta_0} L(\underline{\theta};\underline{X})}{\sup_{\Theta} L(\underline{\theta};\underline{X})} = \frac{L(\hat{\underline{\theta}}_0;\underline{X})}{L(\hat{\underline{\theta}};\underline{X})} \tag{3.1.1}$$

where $\hat{\underline{\theta}}$ is an MLE of $\underline{\theta}$ obtained by performing an unrestricted maximization of $L(\underline{\theta};\underline{X})$; and $\hat{\underline{\theta}}_0$ is the MLE of $\underline{\theta}$ obtained by performing a restricted maximization of

$L(\underline{\theta};\underline{X})$. The rejection region of the likelihood ratio test is $\{\underline{X}:\lambda(\underline{X})\le c\}$, where c is any number that satisfies $0\le c\le 1$. Since $\lambda(\underline{X})$ might be a complicated function of $\underline{X}$, it may be difficult to express $c$ in the inequality.

In this dissertation, the issue is to distinguish between the null hypothesis (data taken from single component bivariate normal distribution) and the alternative hypothesis (data taken from two component bivariate normal mixture distribution). Instead of the above $\lambda$ statistic, we may equivalently consider the following likelihood ratio test (LRT) statistic:

$$\begin{aligned} -2\log\lambda &= -2\log\frac{L_0(\hat{\underline{\theta}}_0;\underline{X})}{L_1(\hat{\underline{\theta}}_1;\underline{X})} \\ &= 2[\log L_1(\hat{\underline{\theta}}_1;\underline{X}) - \log L_0(\hat{\underline{\theta}}_0;\underline{X})] \\ &= 2[l(\hat{\underline{\theta}}_1) - l(\hat{\underline{\theta}}_0)] \end{aligned} \qquad (3.1.2)$$

where $\hat{\underline{\theta}}_1$ and $\hat{\underline{\theta}}_0$ denote the MLE of $\underline{\theta}$ calculated under the alternative hypothesis and the null hypothesis respectively.

Under some regularity conditions, the asymptotic null distribution of the likelihood ratio test has a chi-squared distribution whose degrees of freedom is equal to the difference of the number of parameters being estimated under the alternative hypothesis and the null hypothesis. In our case, this difference is 3. That is if the regularity conditions held the null distribution of -2logλ would be a chi-squared distribution with 3 degrees of freedom. However, since both the mixing proportion ($\pi$) and effect size values ($D_1, D_2$) are on the boundary of the parameter space under the null

hypothesis, the regularity conditions do not hold in our mixture model case. Hence the chi-squared distribution may not be a reasonable approximation to the asymptotic distribution of the likelihood ratio test.

## 3.2 Parametric Bootstrap Method of Inference

The parametric bootstrap method is employed to get the critical values of the observed LRT statistic. For a simulated data set, we can calculate the estimated values of the unknown parameters and obtain a large number of bootstrap samples based on them. Consider a data set of size $n$, the basic steps in the parametric bootstrap procedure are:

Step 1: Simulate a data set $\underline{x_1}, \underline{x_2}, ..., \underline{x_n}$;

Step 2: Fit a single component bivariate normal and a two component bivariate normal mixture model to the data $\underline{x_1}, \underline{x_2}, ..., \underline{x_n}$, and use the EM algorithm to get MLE $\underline{\hat{\theta}_0}$ and $\underline{\hat{\theta}_1}$ respectively;

Step 3: Calculate the statistic of interest, the log likelihood ratio statistic $G^2 = -2\log\lambda$, denoted as $G^2_{obs}$;

Step 4: From the maximum likelihood estimates of parameters computed under the null hypothesis, $\underline{\hat{\theta}_0}$, draw a data set of size *n*. This is a single bootstrap data set;

Step 5: Fit a single component bivariate normal and a two component bivariate normal mixture model to the bootstrap data set, and then compute the corresponding log likelihood ratio statistic $G^2 = -2\log\lambda$, for this bootstrap data sample, yielding $G_1^{2*}$;

Step 6: Repeat steps 4and 5 independently B times, where B is a large number, in order to create B bootstrap samples and obtain their corresponding values of $G_i^{2*}$, $i = 1,2,...,B$;

Step 7: Count the number of $G_i^{2*} \geq G_{obs}^2$, for i=1, 2, …, B. Then we can calculate the p-value using the bootstrap as

$$p_B = \frac{1}{B}\sum_{i=1}^{B} I\{G_i^{2*} \geq G_{obs}^2\} \tag{3.1.3}$$

where $I\{A\}$=1, if {A} occurs; $I\{A\}$=0, otherwise

If $p_B$ is less than the given significance level of $\alpha$, we reject the null hypothesis at the $\alpha$ level of significance.

The above steps for bootstrap procedure can be simply expressed by Figure 3.1.

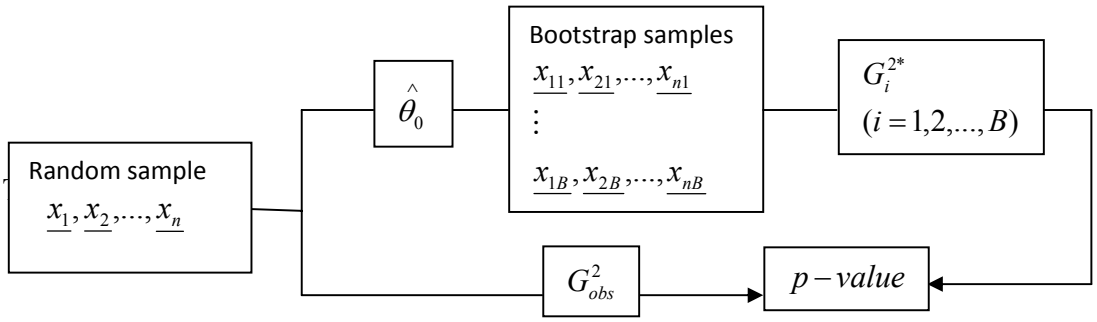



Figure 3.1: Flow chart of p-value calculation for bootstrap test

# Chapter 4

# Power Study and Simulation Results

## 4.1 Estimating the Power of the Bootstrap Test

In this section we examine the power of the parametric bootstrap test (defined in Section 3.2) for detecting whether the data is a random sample from a single bivariate normal distribution or from a two component bivariate normal mixture distribution. We investigate how well the bootstrap test performs when the simulated data are generated from the alternative hypothesis. For a given significance level $\alpha$, the power of the test is a function of the sample size, the mixing proportion, the within component correlation and the distance between component means.

### 4.1.1 Data Simulation

In our power study, samples are simulated from the two component bivariate normal mixture distribution. We considered 36 combinations of parameter values along with 4 different sample size values ($n$ = 50, 100, 200 and 500). This yields 144 combinations in total. For each setting, we ran $N_s (N_s = 1000)$ simulations. The parameter values used to simulate the data are listed below:

$$\mu_{11} = \mu_{12} = 0,\ \sigma_1 = \sigma_2 = 1,$$

$$D = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \end{pmatrix}, \begin{pmatrix} 3 \\ 4 \end{pmatrix},$$

$\pi = 0.5$, 0.7, 0.9, and

$\rho = 0$, 0.5, 0.9.

The procedure to generate a random sample from the two component bivariate normal mixture with joint pdf

$$g(\underline{X};\underline{\theta_1}) = \pi N_1(\underline{X};\underline{\mu_1},\Sigma) + (1-\pi)N_2(\underline{X};\underline{\mu_2},\Sigma) \qquad (4.1.1)$$

Step 1: Generate a uniform random number $U_i (i = 1,2,...,n)$ between 0 and 1. If $U_i \le \pi$, proceed to Step 2; otherwise, execute Step 3.

Step 2: Generate variable $\underline{X_i}$ from bivariate normal distribution $N_1(\underline{X};\underline{\mu_1},\Sigma)$.

Step 3: Generate variable $\underline{X_i}$ from bivariate normal distribution $N_2(\underline{X};\underline{\mu_1},\Sigma)$.

The steps to generate the variable $\underline{X_i}$ from a single bivariate normal distribution are summarized in Figure 4.1. Simulations of data can also be obtained by using a subroutine in the GNU scientific library (GSL) and other math software.

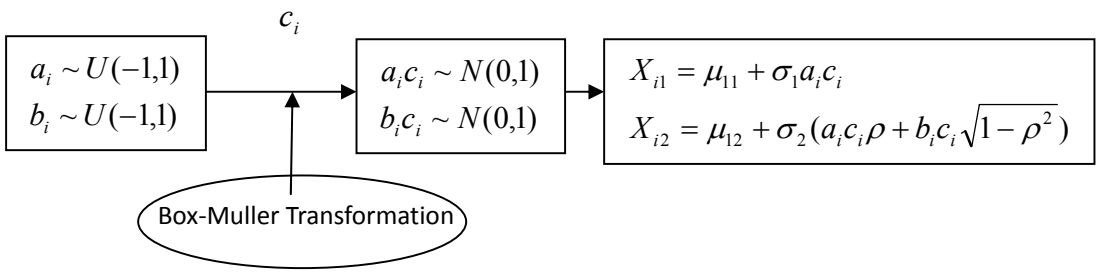



Figure 4.1: Variable generation from a bivariate normal distribution

## 4.1.2 Power Study of the Parametric Bootstrap Test

First, we randomly draw 1000 samples under different parameter combination settings. For each sample, we calculate the LRT, which are defined as $LRT_i$, $i = 1,2,...,1000$. Furthermore, the maximum likelihood estimates for unknown parameters $\underline{\hat{\theta}_{0i}} = (\hat{\mu}_{01i}, \hat{\mu}_{02i}, \hat{\sigma}_{01i}, \hat{\sigma}_{02i}, \hat{\rho}_{0i})$ under the null hypothesis are obtained as well. Second, we perform the parametric bootstrap study, which is described in Section 3.2, under the parameter values $\underline{\hat{\theta}_{0i}}$. The values $LRT_{ij}, j = 1,2,...,N_B$, are obtained for each bootstrap sample, where $N_B$ also equals to 1000. For each simulated sample, the *p*-value can be computed by counting the number of times that $LRT_{ij} \geq LRT_i$:

$$p_i = \frac{\#LRT_{ij} \geq LRT_i}{N_B} (i = 1,...,1000) \tag{4.1.2}$$

With a chosen significance level $\alpha$ (e.g. 0.05, 0.01), we can decide whether to reject or accept the null hypothesis $H_0$ according to the bootstrap *p*-value. Then we count the number of rejections of the null hypothesis $H_0$ among $N_s (N_s = 1000)$ samples and use the following expression to compute the power:

$$Power(\alpha) = P(Reject\ H_0 \mid H_1\ is\ true)$$

$$= \frac{\sum_{i=1}^{N_S} I(p_i < \alpha)}{N_S} \tag{4.1.3}$$

Table 4.1 contains the values of power results of the bootstrap test based on the simulated data with 36 parameter settings for sample size $n = 200$. The power results of the bootstrap test based on the simulated data for 36 parameter settings with sample size

n = 50, 100, 500 can be found in Appendix D. Figure 4.2, 4.3, 4.4, and 4.5 illustrate the power results of the parametric bootstrap test for each sample size for comparison purpose. We can see that the power of bootstrap test increases as the sample size $n$ increases and the average of the two mixture effect sizes, $\overline{D}$ , increases. The power of bootstrap test is slightly sensitive to $\pi$ in the range $0.5 \le \pi \le 0.9$ (which is equivalent to $0.1 \le \pi \le 0.9$ ). The power is low when the sample size $n \le 100$ and the average of mixture effect size $\overline{D} = 2$ . Additionally, sample size $n = 100$ or above is needed to obtain a reasonable power ( $\ge 0.80$ ) for those mixtures with $\overline{D} \ge 2.5$ and $0 \le \rho \le 0.5$ . When the sample size $n = 50$ , a larger average standardized distance between the two component means, i.e. " mixture effect size = 3.0" is needed to obtain a reasonable power. When the average of the two mixture effect sizes, $\overline{D} = 2$ , reasonable power cannot be achieved even with a sample size of 500. When $\overline{D}$ is greater or equal to 3, the power of bootstrap test is essentially 1.0 for all samples of 50 or more regardless mixing proportion $\pi$ and within component correlation $\rho$ . The results are similar for significance level 0.01 (See Appendix D).

Table 4.1 The empirical power of the bootstrap test for sample size $n = 200$

| Mixing proportion | Mean of effect size | within component correlation | Power( $\alpha = 0.05$ ) | Power( $\alpha = 0.01$ ) |
|---|---|---|---|---|
| 0.5 | 2 | 0 | 0.946 | 0.844 |
| | | 0.5 | 0.496 | 0.314 |
| | | 0.9 | 0.257 | 0.114 |
| | 2.5 | 0 | 0.995 | 0.995 |
| | | 0.5 | 0.975 | 0.941 |
| | | 0.9 | 0.995 | 0.995 |
| | 3 | 0 | 0.995 | 0.995 |
| | | 0.5 | 0.990 | 0.990 |
| | | 0.9 | 0.995 | 0.995 |
| | 3.5 | 0 | 1.000 | 1.000 |
| | | 0.5 | 0.995 | 0.995 |
| | | 0.9 | 0.993 | 0.993 |
| 0.7 | 2 | 0 | 0.968 | 0.906 |
| | | 0.5 | 0.605 | 0.368 |
| | | 0.9 | 0.351 | 0.193 |
| | 2.5 | 0 | 1.000 | 1.000 |
| | | 0.5 | 1.000 | 0.990 |
| | | 0.9 | 1.000 | 1.000 |
| | 3 | 0 | 1.000 | 1.000 |
| | | 0.5 | 1.000 | 1.000 |
| | | 0.9 | 1.000 | 1.000 |
| | 3.5 | 0 | 1.000 | 1.000 |
| | | 0.5 | 1.000 | 1.000 |
| | | 0.9 | 1.000 | 1.000 |
| 0.9 | 2 | 0 | 0.906 | 0.820 |
| | | 0.5 | 0.580 | 0.319 |
| | | 0.9 | 0.336 | 0.178 |
| | 2.5 | 0 | 1.000 | 1.000 |
| | | 0.5 | 0.988 | 0.948 |
| | | 0.9 | 0.998 | 0.985 |
| | 3 | 0 | 1.000 | 1.000 |
| | | 0.5 | 1.000 | 1.000 |
| | | 0.9 | 1.000 | 1.000 |
| | 3.5 | 0 | 1.000 | 1.000 |
| | | 0.5 | 1.000 | 1.000 |
| | | 0.9 | 1.000 | 1.000 |

Note: 1. The power results are based on 1000 bootstrap samples and 1000 replicates.

2. Significance level of $\alpha$ =0.05, 0.01.

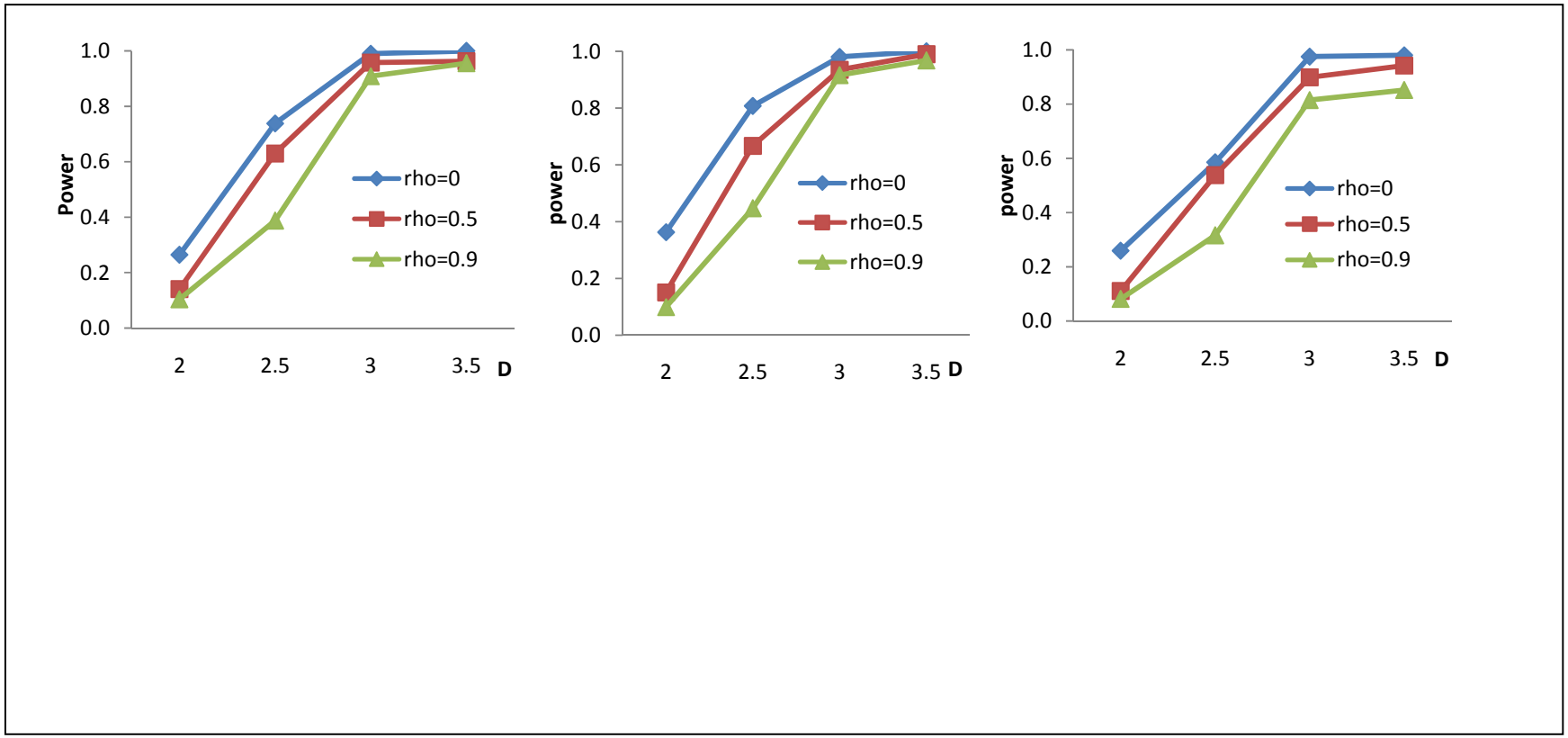



Figure 4.2: Power results of the parametric bootstrap test for sample size $n = 50$

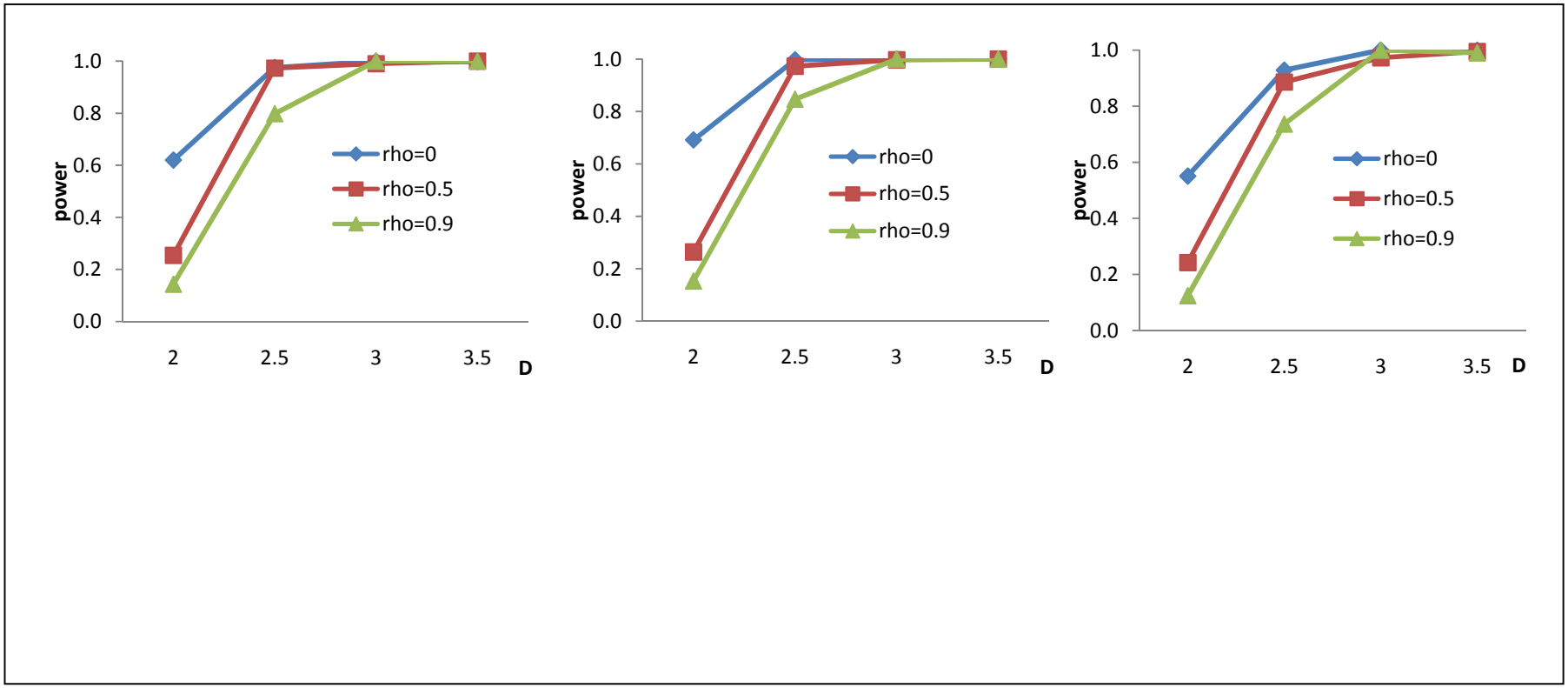



Figure 4.3: Power results of the parametric bootstrap test for sample size $n = 100$

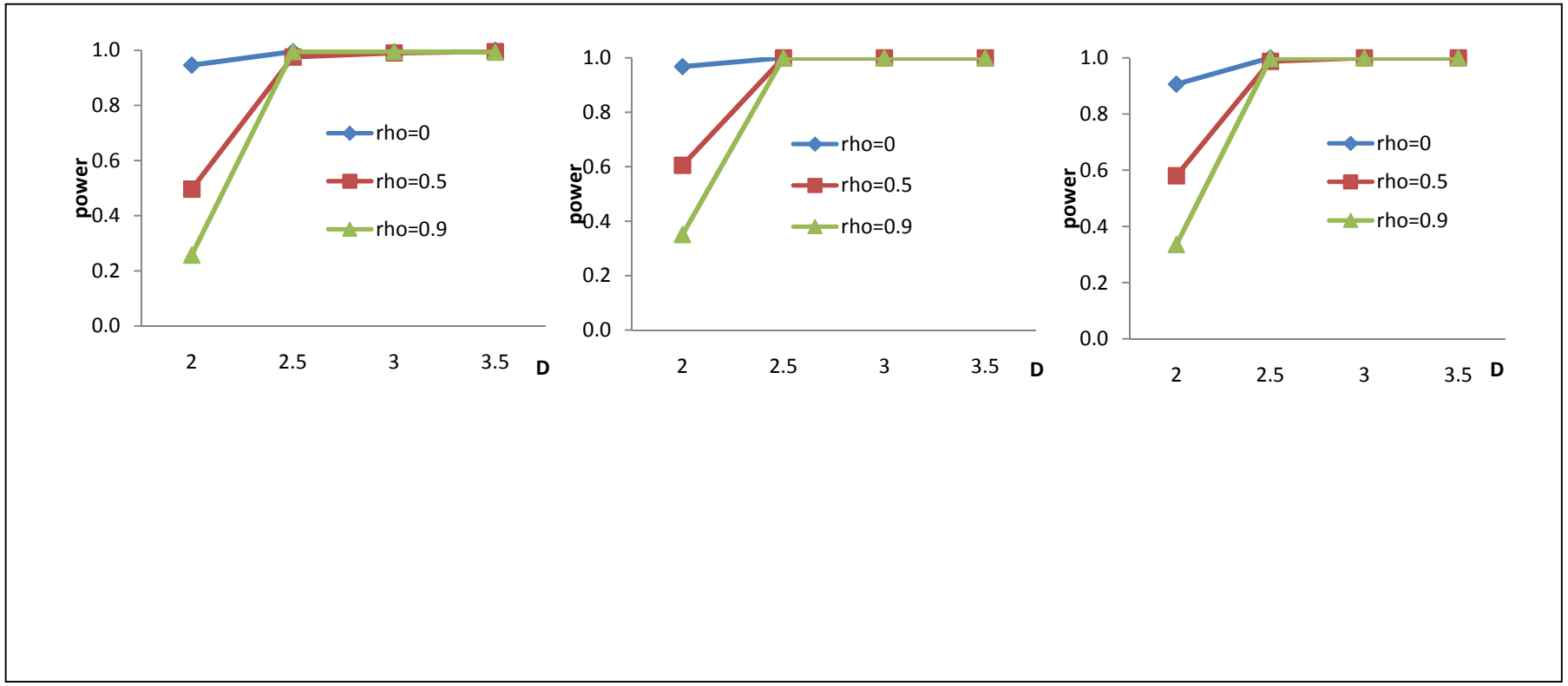



Figure 4.4: Power results of the parametric bootstrap test for sample size $n = 200$

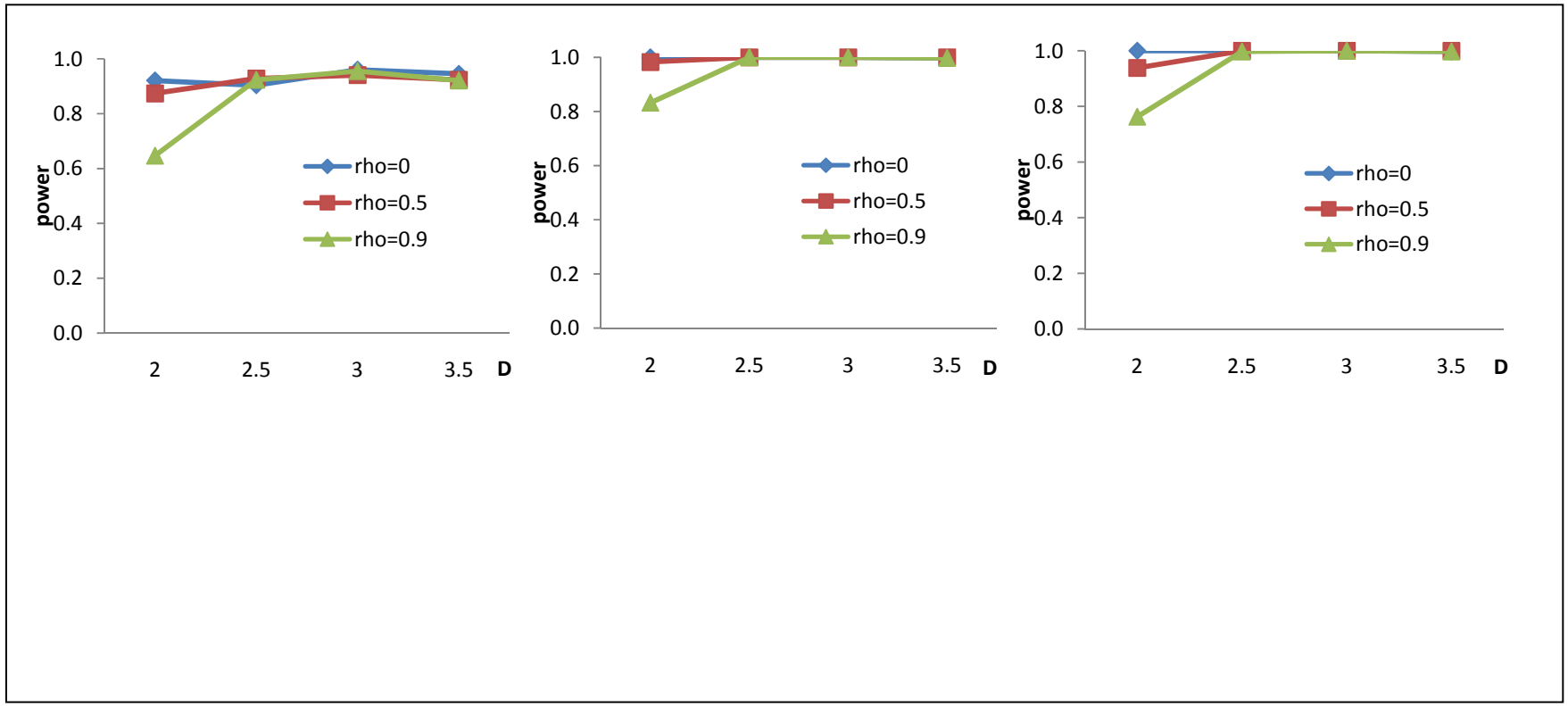



Figure 4.5: Power results of the parametric bootstrap test for sample size $n = 500$

## 4.2 Precision of the MLEs under the Alternative Hypothesis

In order to understand how well our algorithm program works, we examine the summary statistics of the MLEs. Parameter values defined in section 4.1.1 for various sample size ( $n = 50, 100, 200, 500$) are considered.

For each given sample size, we generate data from a two component bivariate normal mixture distribution with specified parameter values. Then the MLEs of parameters are obtained using the EM algorithm described in Section 2.3. This is repeated for 500 simulated samples generated under the alternative hypothesis. Table 4.2 contains the mean and standard error of the MLEs of the parameters under the alternative hypothesis when mixing proportion $\pi = 0.5$. The mean MLEs of $\pi$ are close to 0.5, and the mean of the MLEs of others parameters are close to their parameter settings. Meanwhile, the MLEs seem to converge to their true values with decreasing standard errors as sample size increases. Another way to evaluate the difference between estimated value and the true value is to calculate the mean square error (MSE) of the estimator. Clearly, it will be preferable to have very small values of MSE, i.e. 0.0. For a given estimate $\hat{\theta}$ of the parameter $\theta$, the MSE is calculated as

$$\begin{aligned} MSE(\hat{\theta}) &= E((\hat{\theta}-\theta)^2) \\ &= Var(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2 \end{aligned} \tag{4.2.1}$$

The results of MSE of MLEs are also reported in Table 4.2. As expected, the MSE of the MLEs decrease as sample size and the distance between component means increases. In

addition, higher within component correlation is associated with larger MSE of the MLEs in most cases. The MSE values are associated with the standard error. For example, one can see in table that for the MLE of mixing proportion with an item within component correlation of 0.0 and a sample of 100, the MSE decreases from 0.011 to 0.003 when $D_2$ increases from 2.0 to 3.0. The respective standard error for this MLE also decreases from 0.005 to 0.003. In the extreme scenario, the MSEs of $\hat{\pi}$ and $\hat{\rho}$ are very close to zero which implied that $\hat{\pi}$ and $\hat{\rho}$ are good estimators of their true values. Similar results hold under other mixing proportion settings (results not shown in the table). Armed with this information, our estimation procedure seems to work very well in estimating the MLEs of all parameters at all parameter settings.

Table 4.2 Mean MLEs with the standard error and MSE when $\pi = 0.5$ in two component bivariate normal mixture model

| $D_1$ | $D_2$ | $\rho$ | $n$ | mean MLEs(SE) | | | | MSE | MSE | MSE | MSE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | $\hat{\pi}$ | $\hat{D}_1$ | $\hat{D}_2$ | $\hat{\rho}$ | ($\hat{\pi}$) | ($\hat{D}_1$) | ($\hat{D}_2$) | ($\hat{\rho}$) |
| 2 | 2 | 0.0 | 50 | 0.488(0.007) | 1.843(0.039) | 1.805(0.039) | 0.118(0.015) | 0.025 | 0.767 | 0.806 | 0.122 |
| | | | 100 | 0.503(0.005) | 1.946(0.026) | 1.941(0.026) | 0.060(0.009) | 0.011 | 0.340 | 0.349 | 0.048 |
| | | | 200 | 0.498(0.003) | 2.006(0.011) | 1.974(0.014) | 0.008(0.005) | 0.004 | 0.063 | 0.092 | 0.014 |
| | | | 500 | 0.501(0.003) | 1.962(0.013) | 1.941(0.017) | 0.018(0.005) | 0.004 | 0.091 | 0.155 | 0.014 |
| | | 0.5 | 50 | 0.488(0.009) | 1.741(0.040) | 1.745(0.041) | 0.613(0.010) | 0.041 | 0.846 | 0.907 | 0.061 |
| | | | 100 | 0.504(0.008) | 1.854(0.035) | 1.827(0.033) | 0.561(0.008) | 0.030 | 0.626 | 0.589 | 0.037 |
| | | | 200 | 0.504(0.005) | 1.930(0.021) | 1.930(0.022) | 0.531(0.005) | 0.012 | 0.233 | 0.239 | 0.016 |
| | | | 500 | 0.504(0.004) | 1.909(0.020) | 1.923(0.020) | 0.520(0.004) | 0.010 | 0.204 | 0.212 | 0.008 |
| | | 0.9 | 50 | 0.502(0.010) | 1.734(0.038) | 1.763(0.037) | 0.922(0.003) | 0.047 | 0.786 | 0.734 | 0.004 |
| | | | 100 | 0.497(0.009) | 1.777(0.033) | 1.744(0.034) | 0.920(0.002) | 0.040 | 0.590 | 0.631 | 0.002 |
| | | | 200 | 0.498(0.008) | 1.818(0.028) | 1.818(0.029) | 0.914(0.002) | 0.028 | 0.411 | 0.457 | 0.001 |
| | | | 500 | 0.496(0.006) | 1.874(0.021) | 1.875(0.021) | 0.908(0.001) | 0.017 | 0.246 | 0.245 | 0.001 |
| 2 | 3 | 0.0 | 50 | 0.503(0.004) | 1.956(0.022) | 2.928(0.027) | 0.033(0.010) | 0.008 | 0.242 | 0.359 | 0.055 |
| | | | 100 | 0.502(0.003) | 1.989(0.010) | 2.999(0.012) | 0.007(0.005) | 0.003 | 0.049 | 0.076 | 0.015 |
| | | | 200 | 0.499(0.002) | 1.991(0.007) | 3.001(0.007) | 0.008(0.004) | 0.002 | 0.022 | 0.023 | 0.006 |
| | | | 500 | 0.500(0.003) | 1.960(0.015) | 2.938(0.020) | 0.015(0.005) | 0.004 | 0.116 | 0.200 | 0.013 |
| | | 0.5 | 50 | 0.501(0.006) | 1.907(0.028) | 2.767(0.036) | 0.545(0.009) | 0.017 | 0.389 | 0.705 | 0.040 |
| | | | 100 | 0.495(0.003) | 2.003(0.014) | 2.976(0.015) | 0.497(0.005) | 0.006 | 0.103 | 0.106 | 0.013 |
| | | | 200 | 0.499(0.002) | 2.013(0.009) | 3.009(0.009) | 0.497(0.003) | 0.002 | 0.041 | 0.037 | 0.005 |
| | | | 500 | 0.505(0.004) | 1.948(0.015) | 2.907(0.022) | 0.511(0.004) | 0.006 | 0.113 | 0.250 | 0.006 |
| | | 0.9 | 50 | 0.488(0.005) | 1.974(0.020) | 2.919(0.025) | 0.896(0.002) | 0.012 | 0.205 | 0.324 | 0.003 |
| | | | 100 | 0.501(0.003) | 2.005(0.012) | 3.004(0.010) | 0.897(0.001) | 0.004 | 0.067 | 0.054 | 0.001 |
| | | | 200 | 0.499(0.002) | 2.000(0.008) | 3.003(0.010) | 0.899(0.001) | 0.002 | 0.028 | 0.045 | 0.000 |
| | | | 500 | 0.505(0.003) | 1.984(0.008) | 2.954(0.016) | 0.901(0.001) | 0.004 | 0.034 | 0.125 | 0.000 |

Table 4.2 Mean MLEs with the standard error and MSE when $\pi = 0.5$ in two component bivariate normal mixture model (continued)

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 4 | 0.0 | 50 | 0.497(0.003) | 1.985(0.014) | 4.005(0.013) | 0.008(0.008) | 0.006 | 0.092 | 0.086 | 0.028 |
| | | | 100 | 0.497(0.002) | 1.979(0.009) | 3.989(0.009) | 0.000(0.005) | 0.003 | 0.040 | 0.039 | 0.011 |
| | | | 200 | 0.500(0.002) | 2.011(0.006) | 4.003(0.006) | 0.001(0.003) | 0.001 | 0.020 | 0.020 | 0.006 |
| | | | 500 | 0.502(0.003) | 1.975(0.014) | 3.926(0.023) | 0.017(0.005) | 0.004 | 0.093 | 0.273 | 0.012 |
| | | 0.5 | 50 | 0.505(0.004) | 1.988(0.016) | 3.977(0.016) | 0.494(0.006) | 0.006 | 0.128 | 0.126 | 0.020 |
| | | | 100 | 0.499(0.002) | 1.978(0.010) | 3.987(0.010) | 0.500(0.003) | 0.003 | 0.052 | 0.047 | 0.009 |
| | | | 200 | 0.499(0.002) | 2.014(0.009) | 4.000(0.013) | 0.499(0.003) | 0.002 | 0.037 | 0.085 | 0.004 |
| | | | 500 | 0.503(0.003) | 1.957(0.015) | 3.916(0.025) | 0.505(0.003) | 0.004 | 0.107 | 0.315 | 0.004 |
| | | 0.9 | 50 | 0.498(0.003) | 1.994(0.013) | 4.000(0.013) | 0.898(0.002) | 0.005 | 0.088 | 0.085 | 0.001 |
| | | | 100 | 0.499(0.002) | 2.009(0.009) | 4.011(0.009) | 0.899(0.001) | 0.002 | 0.039 | 0.042 | 0.000 |
| | | | 200 | 0.502(0.002) | 2.000(0.006) | 3.997(0.006) | 0.900(0.001) | 0.001 | 0.020 | 0.020 | 0.000 |
| | | | 500 | 0.499(0.003) | 1.982(0.010) | 3.929(0.020) | 0.900(0.000) | 0.004 | 0.047 | 0.214 | 0/001 |
| 3 | 4 | 0.0 | 50 | 0.500(0.003) | 2.991(0.013) | 3.967(0.012) | -0.003(0.007) | 0.005 | 0.078 | 0.078 | 0.023 |
| | | | 100 | 0.500(0.002) | 2.995(0.009) | 4.010(0.009) | 0.000(0.005) | 0.003 | 0.043 | 0.042 | 0.011 |
| | | | 200 | 0.499(0.002) | 2.996(0.006) | 4.004(0.007) | 0.005(0.003) | 0.001 | 0.019 | 0/021 | 0.005 |
| | | | 500 | 0.500(0.002) | 2.963(0.014) | 3.947(0.022) | 0.009(0.005) | 0.003 | 0.093 | 0.243 | 0.011 |
| | | 0.5 | 50 | 0.501(0.003) | 2.977(0.015) | 3.982(0.017) | 0.503(0.006) | 0.006 | 0.117 | 0.139 | 0.016 |
| | | | 100 | 0.503(0.002) | 2.993(0.009) | 3.997(0.010) | 0.494(0.004) | 0.003 | 0.044 | 0.047 | 0.008 |
| | | | 200 | 0.497(0.002) | 3.001(0.007) | 3.987(0.011) | 0.502(0.003) | 0.002 | 0.026 | 0.061 | 0.003 |
| | | | 500 | 0.503(0.002) | 2.959(0.015) | 3.944(0.020) | 0.505(0.003) | 0.002 | 0.120 | 0.201 | 0.004 |
| | | 0.9 | 50 | 0.500(0.003) | 2.982(0.016) | 3.977(0.017) | 0.899(0.002) | 0.006 | 0.135 | 0.139 | 0.001 |
| | | | 100 | 0.498(0.002) | 3.020(0.008) | 4.019(0.008) | 0.898(0.001) | 0.003 | 0.035 | 0.036 | 0.000 |
| | | | 200 | 0.501(0.002) | 3.003(0.007) | 4.002(0.006) | 0.900(0.001) | 0.001 | 0.022 | 0.021 | 0/000 |
| | | | 500 | 0.501(0.002) | 2.975(0.010) | 3.956(0.016) | 0.901(0.001) | 0.002 | 0.049 | 0.132 | 0/000 |

Note: All values are based on 500 replications on each parameter combination setting.

# Chapter 5

# Modeling LRT statistics

## 5.1 Empirical Null Distribution

In the previous chapters, all simulations were conducted under the alternative hypothesis. In a common hypothesis testing scenario, the null distribution of the test statistic is known. Thus we can conduct the statistical inference with this null distribution. While at the time when the theoretical null is unknown, it is necessary to estimate the null distribution through simulation and yield an "empirical null". In this section, we discuss how to obtain the empirical null distribution of LRT for various parameter settings.

### 5.1.1 Data simulation

In our null hypothesis study, we simulate samples from the single bivariate normal distribution with 15 different parameter combination settings. For each setting, we run $N_s (N_s = 1000)$ simulations. The parameter values used to simulate the data for sample size n = 50, 100, 200, 500, 1000 are:

$$\mu_{01} = \mu_{02} = 0,\ \sigma_{01} = \sigma_{02} = 1, \text{ and}$$

$$\rho_0 = 0, 0.5, 0.9.$$

The procedure to simulate observation from a single bivariate normal distribution is

same as described in Section 4.1.1.

## 5.1.2 The Null distribution of LRT

The framework of the LRT calculation under the null hypothesis is given in Figure 5.1. $\vec{X}_i$ denote a simulated sample under the null distribution specified by the parameter combination and sample size. The simulation results for the LRT under the null hypothesis are shown in Tables 5.1, 5.2 and 5.3 respectively. As the first step of analysis, the means, standard deviations, and selected percentiles at each parameter combination setting for each sample size reveal information about the distribution of the LRT under the null hypothesis. After examination of the three tables, we believe that the mean of the LRT decreases as sample size increases for each correlation coefficient setting. Also the selected percentile of the LRT decreases as sample size increases. Next, a transformation of LRT is performed to study the value of LRT.

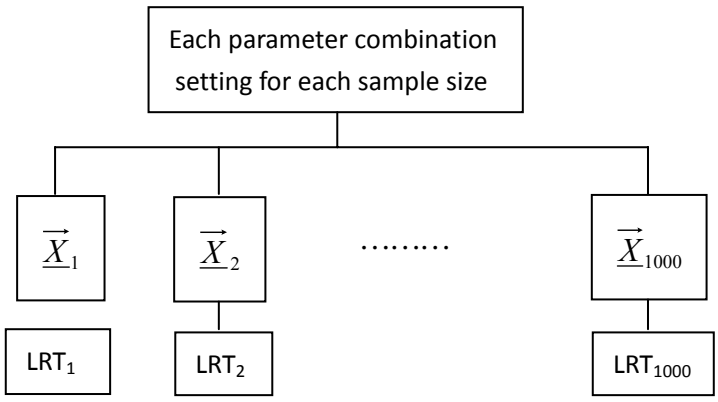



Figure 5.1 LRT calculation chart under the null hypothesis

Table 5.1 Simulation results of LRT under the null hypothesis when correlation coefficient $\rho_0 = 0.0$

| Sample size | Mean of LRT | SD of LRT | Percentiles of LRT | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 25 | 50 | 75 | 90 | 95 | 99 |
| n=50 | 5.7161 | 2.9001 | 3.5831 | 5.1463 | 7.3446 | 9.6514 | 11.1028 | 14.6090 |
| n=100 | 5.4001 | 2.9012 | 3.3158 | 4.8474 | 7.0471 | 9.2645 | 11.2678 | 14.5561 |
| n=200 | 5.2448 | 2.8506 | 3.1836 | 4.7676 | 6.6480 | 8.9409 | 10.4113 | 14.2156 |
| n=500 | 4.9943 | 2.8454 | 3.0306 | 4.4006 | 6.3299 | 8.6716 | 10.2290 | 14.3296 |
| n=1000 | 4.7176 | 2.7750 | 2.7649 | 4.1855 | 6.1597 | 8.4727 | 10.0951 | 12.5591 |

Note: All LRT statistic values are based on 1000 replicates

Table 5.2 Simulation results of LRT under the null hypothesis when correlation coefficient $\rho_0 = 0.5$

| Sample size | Mean of LRT | SD of LRT | Percentiles of LRT | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 25 | 50 | 75 | 90 | 95 | 99 |
| n=50 | 5.6242 | 2.6892 | 3.5796 | 5.1648 | 7.2231 | 9.3483 | 10.7222 | 13.4451 |
| n=100 | 5.2890 | 2.7348 | 3.2482 | 4.9006 | 6.7299 | 8.8973 | 10.6193 | 13.9617 |
| n=200 | 5.2759 | 2.7392 | 3.2343 | 4.7343 | 6.8760 | 9.1204 | 10.3667 | 13.3782 |
| n=500 | 5.0299 | 2.8359 | 3.0263 | 4.5058 | 6.3992 | 8.6458 | 10.4202 | 14.1423 |
| n=1000 | 4.8458 | 3.0307 | 2.6980 | 4.2627 | 6.2772 | 8.9385 | 10.3526 | 14.9373 |

Note: All LRT statistic values are based on 1000 replicates

Table 5.3 Simulation results of LRT under the null hypothesis when correlation coefficient $\rho_0 = 0.9$

| Sample size | Mean of LRT | SD of LRT | Percentiles of LRT | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 25 | 50 | 75 | 90 | 95 | 99 |
| n=50 | 5.7436 | 2.9061 | 3.6891 | 5.1811 | 7.1253 | 9.3564 | 11.5273 | 15.5504 |
| n=100 | 5.4305 | 2.7444 | 3.3588 | 4.8785 | 7.0793 | 8.9788 | 10.5205 | 13.3610 |
| n=200 | 5.2680 | 2.8273 | 3.2370 | 4.7191 | 6.6237 | 9.0477 | 10.7774 | 14.2209 |
| n=500 | 4.8295 | 2.6862 | 2.8004 | 4.3333 | 6.2346 | 8.4840 | 10.1762 | 13.1974 |
| n=1000 | 4.7523 | 2.8748 | 2.7787 | 4.1945 | 6.1061 | 8.7556 | 10.3406 | 13.7168 |

Note: All LRT statistic values are based on 1000 replicates

### 5.1.3 Modeling the LRT under the null hypothesis

A linear regression model is used to model the distribution of the LRT under the null hypothesis. Since the values of the LRT are apparently not normally distributed, directly using the LRT values as the dependent variable violate the normality assumption of the linear regression model. Our solution is to perform a cube root transformation on LRT due to the fact that the distribution of the cube root of a chi-square random variable is closer to a normal distribution. Then we retest the normality after performed cube root transformation on LRT for each parameter combination settings. Figure 5.2 shows the Q-Q plot of $\sqrt[3]{LRT}$ with sample size $n$ = 500. Most points fall on the straight line in the plot, which shows that the distribution of the cube root of LRT is closer to a normal distribution.

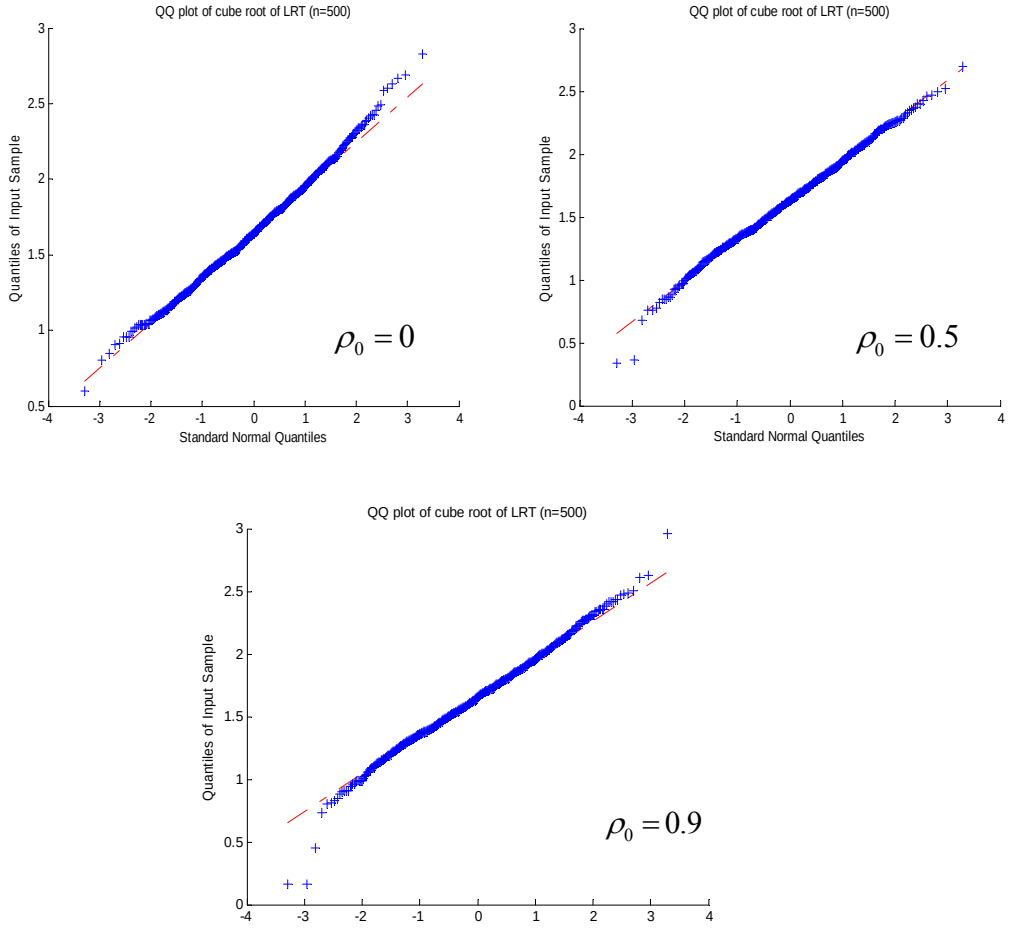



Figure 5.2 Normal Q-Q plot of $\sqrt[3]{LRT}$ at sample size n =500

Note: 1. All values of $\sqrt[3]{LRT}$ are based on 1000 replicates;

2. Parameter settings are $\mu_{01} = \mu_{02} = 0$, $\sigma_{01} = \sigma_{02} = 1$.

In our regression analysis, the independent variables are the sample size $n$ and the correlation coefficient $\rho_0$. The dependent variable is the mean of the value of $\sqrt[3]{LRT}$ for each sample size and correlation coefficient, which can be expressed as $E(\sqrt[3]{LRT}) = f(n, \rho_0) + error$. We measure each factor's contribution to the null distribution of the LRT statistic by performing a balanced two-way ANOVA for

comparing the means of the observed values. The sample size $n$ has five levels, and the correlation coefficient $\rho_0$ has three levels. There are 1000 replicates per contribution. Table 5.4 contains the ANOVA table for comparing the means of $\sqrt[3]{LRT}$. The sample size $n$ is significant, while neither the correlation coefficient $\rho_0$ nor the interaction term $\rho_0 \times n$ are significant with $p-values = 0.214, 0.743$ respectively. Therefore, the mean of $\sqrt[3]{LRT}$ appears to be only a function of sample size. In addition, to test the linearity assumption of regression analysis, we have plotted the dependent variable against independent variables and notice that there is not a straight line relationship between the value of $\sqrt[3]{LRT}$ and sample size *n*. Hence, we consider transforming sample size *n* to log *n*, which become our independent variables.

Table 5.4 ANOVA table for comparing the means of $\sqrt[3]{LRT}$

| Source | SS | df | MS | F | Prob>F |
|---|---|---|---|---|---|
| sample size n | 31.99 | 4 | 7.99748 | 77.01 | 0.000 |
| correlation coefficient $\rho_0$ | 0.32 | 2 | 0.16013 | 1.54 | 0.214 |
| $\rho_0 \times n$ | 0.53 | 8 | 0.06671 | 0.64 | 0.7426 |
| Error | 1556.26 | 14985 | 0.10385 | | |
| Total | 1589.11 | 14999 | | | |

Note: All values of $\sqrt[3]{LRT}$ are based on 1000 replicates.

Table 5.5 reports the linear regression results of the mean of $\sqrt[3]{LRT}$. The coefficient results for the independent variable give us the following regression model:

$$E(\sqrt[3]{LRT}) = 1.901(0.022) - 0.042(0.004)\ln n \qquad (5.1.1)$$

with R-square equal to 0.973, which indicates the model has good of fit. Table 5.6 reports summary statistics of observed $\sqrt[3]{LRT}$ and fitted $\sqrt[3]{LRT}$. The estimated standard deviation of $\sqrt[3]{LRT}$ is

$$\hat{\sigma} = \sqrt{\frac{\sum s_i^2}{5}} = 0.322$$

The estimated standard error is 0.01.

Table 5.5 Linear regression results for the mean of $\sqrt[3]{LRT}$

| Model | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|
| | B | Std. Error | Beta | | |
| (Constant) | 1.901 | .022 | | 85.377 | .000 |
| Log *n* | -.042 | .004 | -.987 | -10.450 | .002 |

Table 5.6 Summary statistics of $\sqrt[3]{LRT}$ and Type I error rate

| n | $\rho_0$ | observed mean $\sqrt[3]{LRT}$ | Fitted mean $\sqrt[3]{LRT}$ | observed sd $\sqrt[3]{LRT}$ | type I error rate | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Fitted null distribution (α = 0.05) | Bootstrap null distribution (α = 0.05) | Fitted null distribution (α = 0.01) | Bootstrap null distribution (α = 0.01) |
| 50 | 0.0 | 1.723 | 1.737 | 0.291 | 0.036 | 0.063 | 0.002 | 0.020 |
| | 0.5 | 1.739 | | 0.309 | 0.052 | 0.061 | 0.014 | 0.021 |
| | 0.9 | 1.743 | | 0.303 | 0.053 | 0.064 | 0.012 | 0.023 |
| 100 | 0.0 | 1.6967 | 1.707 | 0.295 | 0.029 | 0.060 | 0.008 | 0.017 |
| | 0.5 | 1.704 | | 0.313 | 0.049 | 0.057 | 0.012 | 0.017 |
| | 0.9 | 1.694 | | 0.298 | 0.033 | 0.059 | 0.008 | 0.021 |
| 200 | 0.0 | 1.689 | 1.678 | 0.295 | 0.053 | 0.053 | 0.007 | 0.016 |
| | 0.5 | 1.683 | | 0.305 | 0.045 | 0.054 | 0.013 | 0.013 |
| | 0.9 | 1.688 | | 0.302 | 0.046 | 0.057 | 0.010 | 0.020 |
| 500 | 0.0 | 1.642 | 1.640 | 0.312 | 0.046 | 0.053 | 0.006 | 0.014 |
| | 0.5 | 1.643 | | 0.331 | 0.057 | 0.057 | 0.015 | 0.016 |
| | 0.9 | 1.653 | | 0.331 | 0.053 | 0.054 | 0.009 | 0.017 |
| 1000 | 0.0 | 1.589 | 1.611 | 0.372 | 0.052 | 0.050 | 0.011 | 0.012 |
| | 0.5 | 1.594 | | 0.380 | 0.061 | 0.051 | 0.011 | 0.015 |
| | 0.9 | 1.618 | | 0.378 | 0.071 | 0.053 | 0.019 | 0.014 |

Note: Fitted results are based on 3000 values for each sample size

We use the following expression to calculate the *p*-value estimated by the linear regression model:

$$1-\hat{p}=P(Z\le\frac{\sqrt[3]{LRT}-\hat{\mu}(n)}{\hat{\sigma}}) \tag{5.1.2}$$

where $\hat{\mu}(n)$ are the fitted mean of $\sqrt[3]{LRT}$ with sample size $n=50$, 100, 200, 500 and 1000 and $\hat{\sigma}=0.322$. Based on 1000 observed $\sqrt[3]{LRT}$ values per parameter with each sample size, we obtain 1000 *p*-values from (5.1.2). The type I error rate on $\alpha$ level is

$$\text{Type I error}=\frac{\sum_{i=1}^{1000}I(\hat{p}_i\le\alpha)}{1000} \tag{5.1.3}$$

Also we calculate a critical value for each sample size with $\alpha$ significance level as

$$\text{Critical value}=z_\alpha\hat{\sigma}+\hat{\mu}(n) \tag{5.1.4}$$

We include the critical values of $\sqrt[3]{LRT}$ at each sample size in Table 5.7, which will be used to calculate the fitted power by modeling the LRT under the alternative hypothesis in the next section.

Table 5.7 Critical values of $\sqrt[3]{LRT}$ based on regression model

| $\alpha$ level | Sample size $n$ | | | | |
|---|---|---|---|---|---|
| | 50 | 100 | 200 | 500 | 1000 |
| 0.05 | 2.367 | 2.337 | 2.308 | 2.270 | 2.241 |
| 0.01 | 2.566 | 2.536 | 2.507 | 2.469 | 2.440 |

We also calculate the Type I error rates using the bootstrap method described in section 3.2 when we simulated data under the null hypothesis. Table 5.6 includes the Type I error rates of the fitted null distribution and bootstrap null distribution for each $\alpha$ level. As we expect, the Type I error rates are close to the nominal level 0.05 and 0.01 respectively. Figures 5.3 and 5.4 compare the Type I error rates from the regression model critical values with the Type I error rates obtained by bootstrap method. However, one special case is when $\rho_0 = 0$ at sample size 100. The fitted null distribution Type I error rate (0.029) is significantly smaller than the nominal level (0.05). This is because the regression model does not fit the values of $\sqrt[3]{LRT}$ very well at sample size $n = 100$. Therefore, for Type I error rates of fitted null distribution, whether the Type I error rates are close to desired value only depends on the reliability of the fit of the model. For Type I error rates of bootstrap null distribution, the results show that the proximity of the significance level to the nominal level depends not only on sample sizes but also on the correlation coefficient. The Type I error rates from the bootstrap method in large sample size is slightly closer to the nominal level than it is for a small sample size. Apparently, bootstrap methods are more reliable than our linear regression model in estimation of the Type I error. However, in general our fitted null distribution appears to work very well for samples of 200 or more.

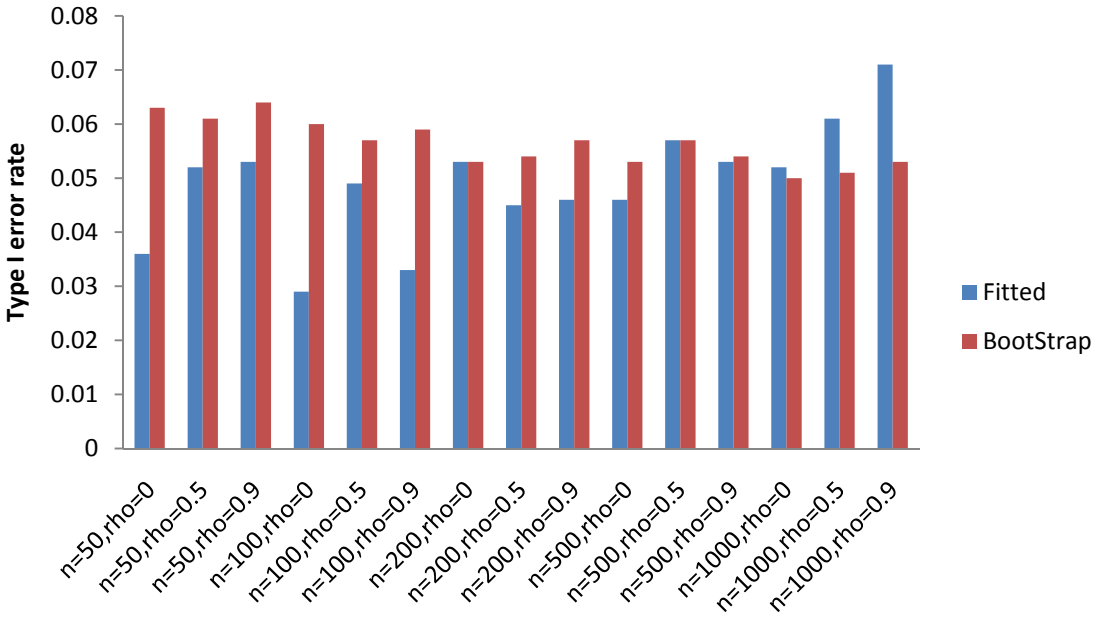



Figure 5.3 Comparison of Type I error rate using regression model and bootstrap method ($\alpha = 0.05$)

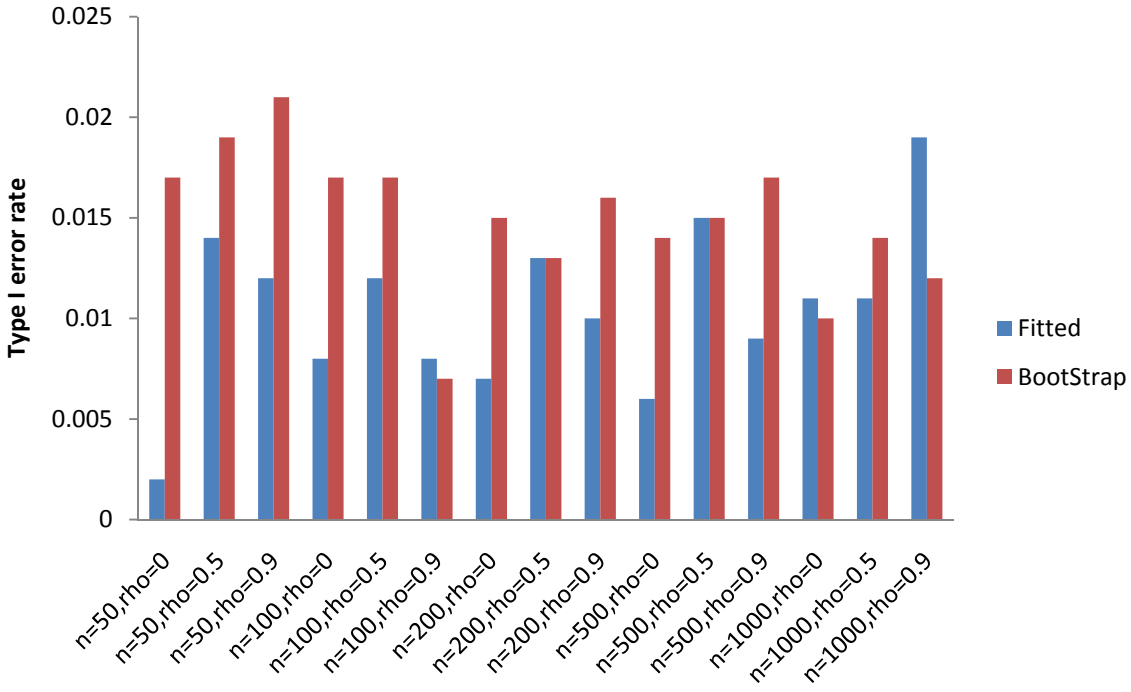



Figure 5.4 Comparison of Type I error rate using regression model and bootstrap method ($\alpha = 0.01$)

## 5.2 Distribution of LRT under the alternative hypothesis

We next consider the distribution of LRT under the alternative hypothesis. Table 5.8 presents the simulation results of this consideration.

Table 5.8: Simulation results of LRT under the alternative hypothesis when $\pi = 0.5$, $\overline{D} = 2 \sim 2.5$

| $\overline{D}$ | $\rho$ | $n$ | Mean LRT | SD LRT | Percentile of LRT | | |
|---|---|---|---|---|---|---|---|
| | | | | | 5 | 10 | 50 |
| 2 | 0 | 50 | 9.622 | 4.888 | 3.351 | 4.395 | 8.491 |
| | | 100 | 13.797 | 6.104 | 5.422 | 6.989 | 12.891 |
| | | 200 | 23.374 | 8.760 | 9.940 | 12.452 | 22.843 |
| | | 500 | 51.151 | 15.785 | 27.107 | 34.320 | 51.377 |
| | 0.5 | 50 | 6.951 | 3.371 | 2.696 | 3.216 | 6.447 |
| | | 100 | 8.838 | 4.276 | 3.172 | 4.028 | 8.032 |
| | | 200 | 11.976 | 5.943 | 4.453 | 5.550 | 10.866 |
| | | 500 | 22.443 | 9.803 | 5.599 | 10.328 | 22.284 |
| | 0.9 | 50 | 6.478 | 3.597 | 2.366 | 2.886 | 5.702 |
| | | 100 | 6.917 | 3.564 | 2.164 | 2.998 | 6.296 |
| | | 200 | 8.843 | 4.639 | 2.560 | 3.629 | 7.957 |
| | | 500 | 14.448 | 7.197 | 3.370 | 5.767 | 13.832 |
| 2.5 | 0 | 50 | 16.352 | 6.790 | 7.070 | 8.410 | 15.429 |
| | | 100 | 28.296 | 9.205 | 14.062 | 17.232 | 27.049 |
| | | 200 | 52.009 | 13.357 | 32.376 | 35.123 | 51.447 |
| | | 500 | 124.601 | 26.740 | 93.262 | 102.624 | 126.086 |
| | 0.5 | 50 | 10.889 | 4.953 | 4.123 | 5.088 | 10.059 |
| | | 100 | 17.287 | 7.206 | 7.033 | 8.559 | 16.641 |
| | | 200 | 30.976 | 9.842 | 16.891 | 19.395 | 29.731 |
| | | 500 | 68.005 | 20.450 | 39.661 | 47.890 | 69.557 |
| | 0.9 | 50 | 14.224 | 5.793 | 6.785 | 7.656 | 13.425 |
| | | 100 | 24.413 | 8.483 | 11.807 | 13.903 | 23.757 |
| | | 200 | 44.381 | 12.142 | 26.283 | 29.329 | 43.439 |
| | | 500 | 101.906 | 23.289 | 71.246 | 80.008 | 103.177 |

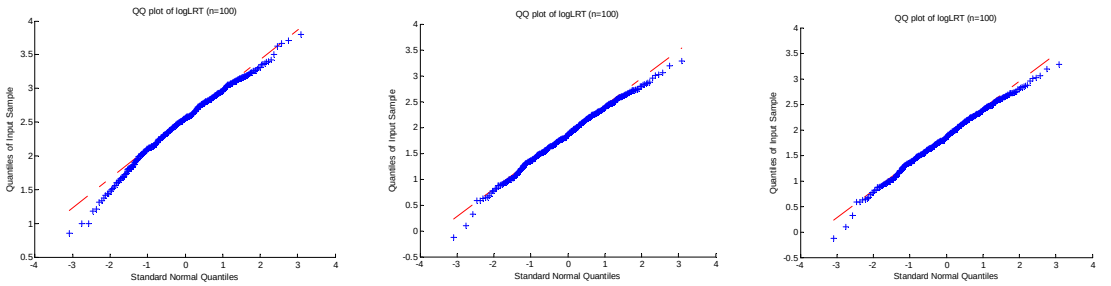
A similar analysis is performed with LRT under the null hypothesis. In the case of the study of the distribution of the LRT under the alternative, we have a multi-factor

design. The variables are sample size $n$, mixing proportion $\pi$, the average of the two mixture effect size $\overline{D}$ and within component correlation $\rho$. Since the linear model has the assumption of normality of errors, we check whether the value of LRT under the alternative hypothesis is normal distributed before we are doing analysis. Unfortunately, the values of LRT under the alternative hypothesis are in fact not approximately normally distributed. Hence, we look for an appropriate transformation of these values that have a distribution that is closer to the normal distribution using the Box-Cox transformation.

From the Box-Cox transformation, we found the transformation power $\lambda = 0.0488$. Since the value of $\lambda$ is close to 0, we can choose natural logarithm of LRT values under the alternative hypothesis as the transformation. Then we obtain transformed values for LRT statistics under the alternative hypothesis. Figure 5.5 shows a selected Q-Q plot of the log of the LRT. It illustrates that these transformed LRT values are approximately normal distributed. We then model the mean of the log of the LRT considering parameters and the sample sizes. The model also include all two factor interactions:

$$E(\log LRT) = f(n, \pi, \overline{D}, \rho, n*\pi, n*\overline{D}, n*\rho, \pi*\overline{D}, \pi*\rho, \overline{D}*\rho) + error$$

Figure 5.5 Selected Q-Q plot of $\log LRT$ with sample size $n$ = 100




$\pi = 0.5, \rho = 0, \overline{D} = 2$ $\pi = 0.7, \rho = 0.9, \overline{D} = 2$ $\pi = 0.9, \rho = 0.5, \overline{D} = 2.5$

Note: All values of $\log LRT$ are based on 1000 replicates.

Table 5.9: ANOVA table for the mean of $\log LRT$

| Source | DF | ANOVA SS | Mean Square | F value | Pr > F |
|---|---|---|---|---|---|
| $\pi$ | 2 | 1.015 | 0.507 | 60.37 | <.0001 |
| $\overline{D}$ | 3 | 113.050 | 37.683 | 4483.20 | <.0001 |
| $\rho$ | 2 | 7.191 | 3.595 | 427.74 | <.0001 |
| $n$ | 3 | 110.445 | 36.815 | 4379.91 | <.0001 |
| $\pi * \overline{D}$ | 6 | 0.160 | 0.027 | 3.18 | 0.0069 |
| $\pi * \rho$ | 4 | 0.007 | 0.002 | 0.20 | 0.9369 |
| $\pi * n$ | 6 | 0.184 | 0.031 | 3.65 | 0.0026 |
| $\overline{D} * \rho$ | 6 | 8.334 | 1.389 | 165.26 | <.0001 |
| $\overline{D} * n$ | 9 | 5.103 | 0.567 | 67.46 | <.0001 |
| $\rho * n$ | 6 | 0.382 | 0.064 | 7.58 | <.0001 |

Table 5.9 reports the ANOVA table for the mean of $\log LRT$. It shows that each main effect was significant with $p - value < 0.0001$. For the interaction variables, only the interaction of the mixing proportion and within component

correlation is not significant with $p-value=0.9369$. As a result, we can include all main effects and interaction effects except $\pi*\rho$. We consider the transformation of sample size $\log n$ and average of effect sizes $\sqrt{\overline{D}}$ in our model. The regression results of the model are reported in Table 5.10. Our model for mean of $\log LRT$ can be written as below with R-square equal to 0.889:

$$E(\log LRT)=0.05-0.586\log n+0.248\pi+0.028\sqrt{\overline{D}}-2.096\rho+0.209\pi*\log n+0.816\sqrt{\overline{D}}*\log n$$
$$-0.121\rho*\log n-0.927\pi*\sqrt{\overline{D}}+1.454\sqrt{\overline{D}}*\rho$$

(5.2.1)

The same procedure is performed to standard deviation of logLRT. The final model for SD of $\log LRT$ is below with R-square equal to 0.526:

$$sd(\log LRT)=-1.409+0.515\log n+3.352\pi-0.22\sqrt{\overline{D}}+0.24\rho-0.675\pi*\log n-0.039\sqrt{\overline{D}}*\log n$$
$$-0.094\pi*\sqrt{\overline{D}}-0.121\sqrt{\overline{D}}*\rho$$

(5.2.2)

Table 5.10: Regression results for mean of $\log LRT$ under alternative hypothesis

| Model | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|
| | B | Std. Error | Beta | | |
| (Constant) | .050 | 2.426 | | .021 | .983 |
| $\log n$ | -.586 | .395 | -.457 | -1.483 | .141 |
| $\pi$ | .248 | 2.190 | .037 | .113 | .910 |
| $\sqrt{\overline{D}}$ | .028 | 1.376 | .004 | .020 | .984 |
| $\rho$ | -2.096 | .971 | -.706 | -2.158 | .033 |
| $\log n * \pi$ | .209 | .226 | .195 | .926 | .356 |
| $\log n * \sqrt{\overline{D}}$ | .816 | .217 | 1.233 | 3.767 | .000 |
| $\log n * \rho$ | -.121 | .100 | -.211 | -1.204 | .231 |
| $\pi * \sqrt{\overline{D}}$ | -.927 | 1.131 | -.251 | -.820 | .414 |
| $\sqrt{\overline{D}} * \rho$ | 1.454 | .502 | .819 | 2.898 | .004 |

Since the mean and standard deviation of log(LRT) are known, we can use the above regression model to calculate the power of fitted model. From Table 5.6, we obtain the critical values of $\log LRT$ for each sample size and significance level. The equation for computing fitted power is:

$$\begin{aligned} Power &= P(\log LRT \geq \log LRT_{cv}) \\ &= P\left(\frac{\log LRT - \hat{\mu}(n,\pi,\overline{D},\rho)}{\hat{\sigma}(n,\pi,\overline{D},\rho)} \geq \frac{\log LRT_{cv} - \hat{\mu}(n,\pi,\overline{D},\rho)}{\hat{\sigma}(n,\pi,\overline{D},\rho)}\right) \\ &= 1 - \phi\left(\frac{\log LRT_{cv} - \hat{\mu}(n,\pi,\overline{D},\rho)}{\hat{\sigma}(n,\pi,\overline{D},\rho)}\right) \end{aligned} \tag{5.2.3}$$

Table 5.11 Summary statistics of $\log LRT$ and fitted power with sample size n=200

| $\pi$ | $\overline{D}$ | $\rho$ | Observed mean $\log LRT$ | Observed SD $\log LRT$ | Fitted mean $\log LRT$ | Fitted SD $\log LRT$ | Fitted power ($\alpha = 0.05$) | Fitted power ($\alpha = 0.01$) | difference between fitted and observe power $\Delta_{(0.05)}$ | $\Delta_{(0.01)}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.5 | 2 | 0 | 3.074 | 0.413 | 3.161 | 0.572 | 0.873 | 0.760 | -0.073 | -0.094 |
| | | 0.5 | 2.363 | 0.501 | 2.839 | 0.572 | 0.518 | 0.357 | 0.186 | 0.203 |
| | | 0.9 | 2.031 | 0.580 | 2.582 | 0.572 | 0.351 | 0.180 | 0.243 | 0.226 |
| | 2.5 | 0 | 3.917 | 0.268 | 3.779 | 0.483 | 0.996 | 0.983 | 0.002 | -0.009 |
| | | 0.5 | 3.379 | 0.346 | 3.579 | 0.483 | 0.987 | 0.956 | 0.009 | 0.009 |
| | | 0.9 | 3.753 | 0.291 | 3.419 | 0.483 | 0.970 | 0.915 | -0.037 | -0.104 |
| | 3 | 0 | 4.577 | 0.175 | 4.337 | 0.402 | 1.000 | 1.000 | 0.005 | 0.005 |
| | | 0.5 | 4.260 | 0.272 | 4.247 | 0.402 | 1.000 | 1.000 | 0.010 | 0.010 |
| | | 0.9 | 5.038 | 0.122 | 4.175 | 0.402 | 1.000 | 1.000 | 0.005 | 0.005 |
| | 3.5 | 0 | 4.855 | 0.141 | 4.851 | 0.328 | 1.000 | 1.000 | 0.000 | 0.000 |
| | | 0.5 | 4.374 | 0.217 | 4.862 | 0.328 | 1.000 | 1.000 | 0.005 | 0.005 |
| | | 0.9 | 4.421 | 0.182 | 4.870 | 0.328 | 1.000 | 1.000 | 0.007 | 0.007 |
| 0.7 | 2 | 0 | 3.172 | 0.404 | 3.114 | 0.496 | 0.888 | 0.764 | -0.059 | -0.118 |
| | | 0.5 | 2.435 | 0.589 | 2.792 | 0.496 | 0.716 | 0.528 | 0.108 | 0.158 |
| | | 0.9 | 2.163 | 0.507 | 2.535 | 0.496 | 0.421 | 0.227 | 0.156 | 0.132 |
| | 2.5 | 0 | 4.004 | 0.244 | 3.732 | 0.407 | 0.999 | 0.992 | -0.001 | -0.005 |
| | | 0.5 | 3.472 | 0.324 | 3.532 | 0.407 | 0.994 | 0.972 | -0.006 | -0.019 |
| | | 0.9 | 3.811 | 0.280 | 3.372 | 0.407 | 0.983 | 0.935 | -0.024 | -0.081 |
| | 3 | 0 | 4.612 | 0.166 | 4.290 | 0.326 | 1.000 | 1.000 | 0.000 | 0.000 |
| | | 0.5 | 4.316 | 0.209 | 4.200 | 0.326 | 1.000 | 1.000 | 0.000 | 0.000 |
| | | 0.9 | 5.059 | 0.113 | 4.128 | 0.326 | 1.000 | 1.000 | 0.000 | 0.000 |
| | 3.5 | 0 | 4.866 | 0.147 | 4.804 | 0.252 | 1.000 | 1.000 | 0.000 | 0.000 |
| | | 0.5 | 4.411 | 0.190 | 4.815 | 0.252 | 1.000 | 1.000 | 0.000 | 0.000 |
| | | 0.9 | 4.448 | 0.185 | 4.823 | 0.252 | 1.000 | 1.000 | 0.000 | 0.000 |
| 0.9 | 2 | 0 | 3.024 | 0.439 | 3.067 | 0.420 | 0.908 | 0.769 | 0.039 | 0.014 |
| | | 0.5 | 2.375 | 0.576 | 2.745 | 0.420 | 0.613 | 0.389 | 0.170 | 0.220 |
| | | 0.9 | 2.039 | 0.609 | 2.488 | 0.420 | 0.480 | 0.261 | 0.179 | 0.129 |
| | 2.5 | 0 | 3.805 | 0.317 | 3.685 | 0.331 | 1.000 | 0.997 | 0.000 | -0.001 |
| | | 0.5 | 3.286 | 0.385 | 3.485 | 0.331 | 0.998 | 0.986 | 0.011 | 0.040 |
| | | 0.9 | 3.635 | 0.366 | 3.325 | 0.331 | 0.993 | 0.957 | -0.007 | -0.037 |
| | 3 | 0 | 4.420 | 0.226 | 4.243 | 0.250 | 1.000 | 1.000 | 0.000 | 0.000 |
| | | 0.5 | 4.122 | 0.269 | 4.153 | 0.250 | 1.000 | 1.000 | 0.000 | 0.000 |
| | | 0.9 | 4.861 | 0.170 | 4.081 | 0.250 | 1.000 | 1.000 | 0.000 | 0.000 |
| | 3.5 | 0 | 4.701 | 0.191 | 4.757 | 0.175 | 1.000 | 1.000 | 0.000 | 0.000 |
| | | 0.5 | 4.213 | 0.250 | 4.768 | 0.175 | 1.000 | 1.000 | 0.000 | 0.000 |
| | | 0.9 | 4.275 | 0.244 | 4.776 | 0.175 | 1.000 | 1.000 | 0.000 | 0.000 |

The results of fitted power are also reported in Table 5.11. We compare fitted power and the observed power of the bootstrap test. Figures 5.6 and 5.7 show comparison of the power results of fitted and observed for each parameter combination setting and each sample size with significance level of 0.05 and 0.01 respectively. It appears most points fall close to the x=y line, which indicates that the

fitted power is close to the bootstrap power. Most fitted and observed powers estimates are roughly equal as the power increases. However, it does not tell any quantitative information on how well fitted power is. In order to investigate it, we compute the values of the difference between the powers for each method. Figure 5.8 illustrates these values of the difference for 36 parameter combination settings and each sample size with significance level of 0.05. It shows the value of the difference decreases as sample size increases from the red trend line, which indicates the fitted power value gets closer to observed power value when sample size increases. When the sample size $n \leq 100$, most fitted powers are under estimated based on the corresponding observed power. However, the largest difference of 0.32 corresponds to the case where the mixing proportion equals 0.7 and sample size equals 50 should be looked into. Additionally, the power difference corresponds to the case where the mixing proportion equals 0.5 even sample size equals 500 is consistently large. The reason is that the mean and SD of $\log LRT$ from regression model do not fit very well in these parameter settings.

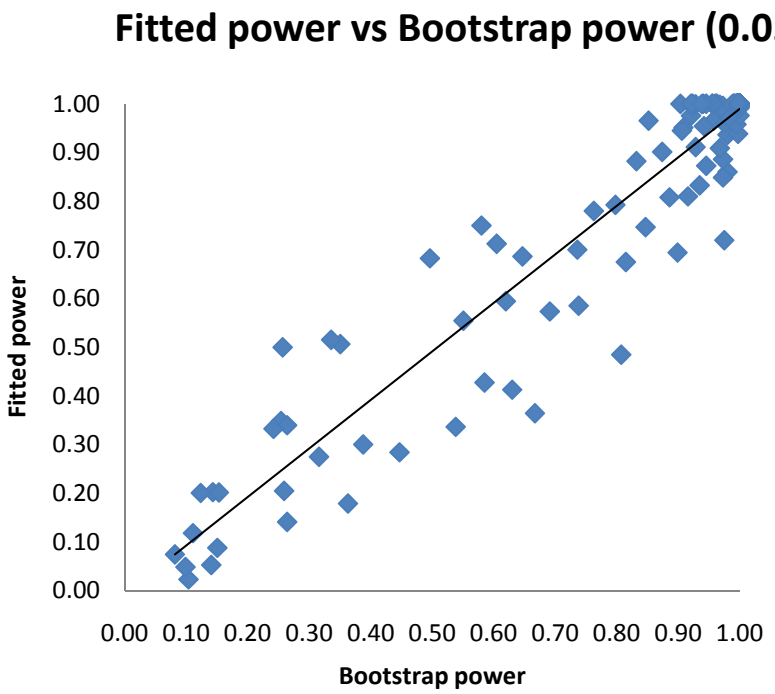



Figure 5.6: Comparison of fitted and observed power at significance level $\alpha = 0.05$

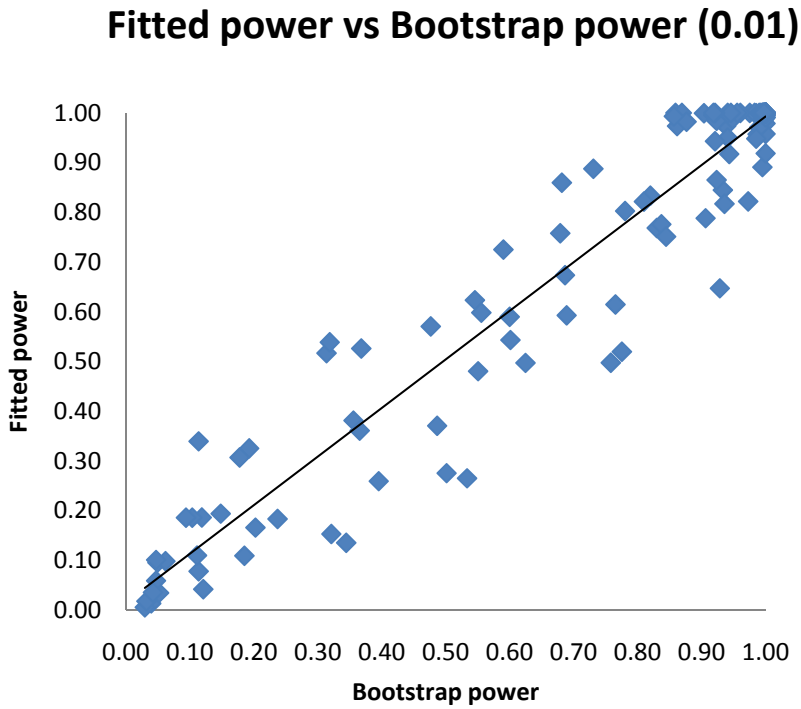



Figure 5.7: Comparison of fitted and observed power at significance level $\alpha = 0.01$

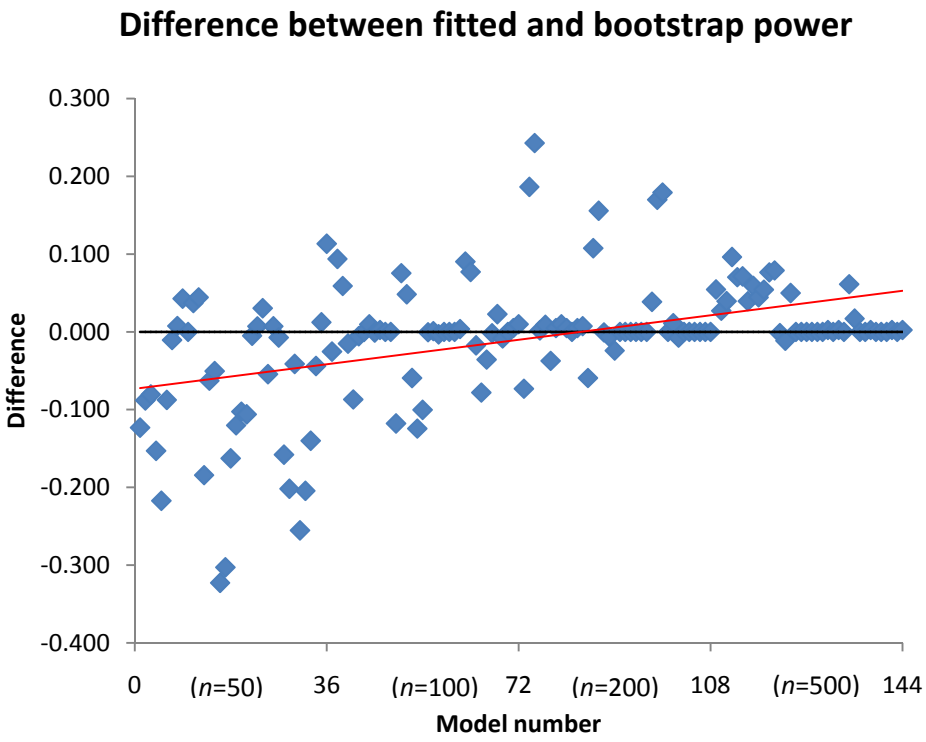



Figure 5.8: Power difference between fitted and observed at $\alpha = 0.05$

Note: 1. Model number = number of parameter combination settings (36) * number of sample size (4) (total 144) (Appendix E).

2. Fitted power = power calculation under regression model.

# Chapter 6

# Discussion and Conclusion

The research in this dissertation include the power analysis of the parametric bootstrap test to detect a two component bivariate normal mixture distribution and modeling the distribution of the LRT statistic under null and alternative hypothesis using simulated data sets of various parameter combination settings. The analysis of the LRT statistic is based on MLEs of unknown parameters obtained using EM algorithm. In the process, software programs have been developed for a simulation study of the power of the bootstrap statistic based on the LRT in this dissertation.

The alternative hypothesis is a two component bivariate normal mixture distribution. As expect, the power of the parametric bootstrap test was sensitive to sample size, average mixture effect size, and slightly sensitive to mixing proportion. We also note that power depends on the within component correlation. The power to detect a bivariate normal mixture increases as sample size increases and the average mixture effect sizes increases. As the mixing proportion approaches 0.5, the power slightly increases in some cases. Power increases as the within component correlation gets closer to 0.0 and is significantly lower when the within component correlation. We think the reason for this is that if the mixture accounts for all of the correlation. The power is high with zero within component correlation. Conversely when the within component correlation is high then the mixture accounts for a small proportion

of the correlation between the two variables. To obtain a reasonable power ( power $\geq 80\%$), one needs a sample size of 200 or more, average of the two mixture effect size, $\overline{D}$, of 2.5 or more and $0.1 \leq \pi \leq 0.9$ to detect a bivariate mixture. When the average of the two mixture effect size $\overline{D}$ is 3.0 or greater, the power is essentially 1.0 regardless of sample size and the remaining parameter values.

The null hypothesis is a single bivariate normal distribution. The mean of LRT under the null hypothesis is only sensitive to the sample size. The distribution of the test appears to be insensitive to the within component correlation under the null hypothesis. We use the mean of the cube root transformation of LRT as the dependent variable to fit a simple linear regression model on $\log n$. We obtain the following expression for mean of $\sqrt[3]{LRT}$ : $E(\sqrt[3]{LRT}) = 1.901(0.022) - 0.042(0.004)\ln n$. Thus, the pdf of $\sqrt[3]{LRT}$ is approximated by a normal distribution with mean of $\sqrt[3]{LRT}$ and variance $0.322^2$. We use this property of LRT statistic under the null hypothesis to calculate the estimated p-value and observed that this approach resulted in the satisfactory estimate of the type I error rate.

For the LRT statistics under the alternative hypothesis, the mean and the standard deviation of LRT is sensitive to all parameter settings. We use log transformation of LRT as dependent variable to fit regression model which included the significant two factor interaction effects. Then we have our final model for the mean and SD of $\log LRT$. Based on the critical values of LRT under the null distribution, we compute our fitted power. Most fitted and observed powers estimates are roughly equal as the power increases. However, it is not accurate enough that one would use it in practice

( there are several situations where the error of the fitted power is greater than 0.05). The equations for obtaining fitted power work better as sample size increased, i.e. sample size n=500 and mixing proportion is greater than 0.5.

From our experience with examples in which there is evidence of a bivariate mixture resulting from a major gene affecting the phenotype, the mixing proportion estimates probably between 0.8 and 1.0 (or equivalently, 0.0 and 0.2), and the within component correlation is unequal to 0. For this reason, we propose to consider these ranges of parameter values.

The parametric bootstrap method has been used in this dissertation. We expect that this would be the method of choice for evaluating the likelihood ratio test as applied to the null and alternative considered here. It would be interesting to compare the power of the method used to other methods. Alternative methods to consider might be the nonparametric bootstrap or a permutation test.

In this dissertation, one would expect that the assumption of equal mixing proportion for both traits implies that the mixture results from two correlated traits determined by a common factor that has not been measured in the individual subjects. Thus this would be an appropriate alternative, for example for two quantitative traits determined by a common gene and also for the joint distribution of a quantitative trait in a pairs of identical twins. This common factor could however be some unrecorded common environmental factor. It is possible and often the case that different factors determine two correlated traits. In this case the marginal distribution of both traits would be a mixture and the mixing proportions would not be equal. In fact one of the

traits might have a mixture distribution, and the other have a trait with a single component normal distribution. It would be interesting to consider all of these situations in the context of nested hypotheses. In the case of testing for a common factor for both traits vs. different factors with (unequal marginal mixing proportions) for each trait for each trait, one could consider the following two hypotheses.

$$H_0 : g(\underline{X};\underline{\theta_0}) = \pi_0 N_1(\underline{X};\begin{pmatrix}\mu_{01}\\ \mu_{02}\end{pmatrix},\begin{pmatrix}\sigma_{01}^2 & \rho_0\sigma_{01}\sigma_{02}\\ \rho_0\sigma_{01}\sigma_{02} & \sigma_{02}^2\end{pmatrix}) + (1-\pi_0)N_2(\underline{X};\begin{pmatrix}\mu_{01}+D_{01}\sigma_{01}\\ \mu_{02}+D_{02}\sigma_{02}\end{pmatrix},\begin{pmatrix}\sigma_{01}^2 & \rho_0\sigma_{01}\sigma_{02}\\ \rho_0\sigma_{01}\sigma_{02} & \sigma_{02}^2\end{pmatrix})$$

versus the alternative hypothesis

$$H_1 : g(\underline{X};\underline{\theta_0}) = \pi_{11} N_1(\underline{X};\begin{pmatrix}\mu_{11}\\ \mu_{12}\end{pmatrix},\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}) + (\pi_1-\pi_{11})N_2(\underline{X};\begin{pmatrix}\mu_{11}\\ \mu_{12}+D_2\sigma_2\end{pmatrix},\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix})$$
$$+(\pi_2-\pi_{11})N_3(\underline{X};\begin{pmatrix}\mu_{11}+D_1\sigma_1\\ \mu_{12}\end{pmatrix},\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}) + (1-\pi_1-\pi_2+\pi_{11})N_4(\underline{X};\begin{pmatrix}\mu_{11}+D_1\sigma_1\\ \mu_{12}+D_2\sigma_2\end{pmatrix},\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix})$$

## Bibliography

[1] Aitkin, M., Anderson, D., and Hinde, J. (1981). Statistical modeling of data on teaching styles (with discussion). Journal of Royal Statistical Society: Series A, 144, 419-461.

[2] Aitkin, M. and Rubin, D. B. (1985). Estimation and hypothesis testing in finite mixture models. Journal of Royal Statistical Society: Series B, 47, 67-75.

[3] Aitkin, M. and Wilson, G. T. (1980). Mixture models, outliers and the EM algorithm. Technometrics, 22, 325-331.

[4] Arcidiacono, P. and Jones, B. J. (2003). Finite mixture distributions, sequential likelihood and the EM algorithm. Econometrica, 71, 933-946.

[5] Basford, K.E. and McLachlan G.J. (1985) Likelihood estimation with normal mixture models. Journal of Applied Statistical Science, 34, 282-289.

[6] Berdai, A. and Garel, B. (1996). Detecting a univariate normal mixture with two components. Statistics and Decisions, 16, 35-51.

[7] Chen, J. and Tan, X. (2009). Inference for multivariate normal mixtures. Journal of Multivariate Analysis, 100, 1367-1383.

[8] Chuang, R. J. and Mendell, N. R. (1997). The approximate null distribution of the likelihood ratio test for a mixture of two bivariate normal distributions with equal covariance. Communication in Statistics B, 26, 631-648.

[9] Chatterjee, N. and Shih, J. (2001). A bivariate cure-mixture approach for modeling familial association in diseases. Biometrics, 57, 779-786.

[10] Cohen, A. C. (1967). Estimation in mixtures of two normal distributions. Technometrics, 9, 15-28.

[11] Day, N. E. (1969). Estimating the components of a mixture of normal distributions. Biometrika, 56, 463–474.

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

[13] Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. Journal of the Royal Statistical Society: Series B , 56,

363–375.

[14] Efron, B. (1979a). Bootstrap methods: another look at the jackknife. Annals of Statistics, 7, 1-26.

[15] Efron, B. (1979b). Computer and the theory of statistics: thinking the unthinkable. SIAM Review.

[16] Efron, B. (1982). The Jackknife, the bootstrap and other resampling plans. Philadelphia: SIAM.

[17] Engelman, L. and Hartigan J. A. (1969). *Percentage Points* of a *Test for Clusters.* Journal of the American Statistical Association, 64, 1647-8.

[18] Everitt, B. S. and Hand, D. J. (1981). Finite mixture distribution. Chapman and Hall, London.

[19] Everitt, B. S. (1981). A Monte Carlo investigation of the likelihood ratio for the number of components in a mixture of normal distributions. Multivariate Behavioral Research, 16, 171-180.

[20] Feng, Z. D. and McCulloch, C. E. (1994). On the likelihood ratio test statistic for the number of components in a normal mixture with unequal variance. Biometrics, 50, 1158-1162.

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

[22] Finch, S. J., Mendell, N. R., and Thode, H. C. Jr. (1989). Probabilistic measures of adequacy of a numerical search for a global maximum. Journal of the American Statistical Association, 84, 1020-1023.

[23] Fowlkes E. B. (1979). *Some methods for studying the mixture of two normal* (lognormal) *distributions*. Journal of the American Statistical Association, 74, 561-575.

[24] Fraley, C. and Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. Technical Report No. 329. Seattle: Department of Statistics, University of Washington.

[25] Hartigan, J. A. (1977). Distribution problems in clustering. In Classification and Clustering. New York: academic Press.

[26] Hartigan, J. A. (1985a). Statistical theory in clustering. Journal of Classification

2, 63-76.

[27] Hartigan, J. A. (1985b). A failure of likelihood asymptotics for normal mixtures. In Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, 2, 807-810.

[28] He, Y., Pan, W. and Lin, J. (2006). Cluster analysis using multivariate normal mixture models to detect differential gene expression with microarray data. Computational Statistics & Data Analysis, 51, 641-658.

[29] Kang, H., Qin, Z. S., Niu, T. and Liu, J. S. (2004). Incorporating genotyping uncertainty in haplotype inference for single-nucletide polymorphisms. American Journal of Human Genetics, 74, 495-510.

[30] Karlis, D. and Xekalaki, E. (2003). Choosing initial values for the EM algorithm for finite mixtures. Computational Statistics & Data Analysis, 41, 577-590.

[31] Kriessler, J. R. and Beers, T. C. (1997). Substructure in galaxy clusters: a two-dimensional approach. The Astronomical Journal, 113, 80-100.

[32] Lindsay B. G. and Basak P. (1993). Multivariate normal mixtures: A fast consistent method of moments. Journal of the American Statistical Association, 88, 468-476.

[33] Lo Y., Mendell N.R. and Rubin D.B. (2001). Testing the number of components in a normal mixture. Biometrika, 88, 767-778.

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

[35] McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number components in a normal mixture. Journal of Applied Statistical Science, 36, 318-324.

[36] McLachlan, G. J. and Basford, K. E. (1988). Mixture models. Inference and Applications to Clustering. Dekker, New York.

[37] McLachlan, G. J., Peel, D., Basford, K. E. and Adams, P. (1999). Fitting of mixtures of normal and t-components. Journal of Statistical Software, 4, No.2.

[38] McLachlan, G. J. and Peel, D. (2000). Finite mixture models. Wiley, New York.

[39] McLachlan, G. and Krishman, T. (1997). The EM algorithm and extensions. Wiley series in probability and statistics. John Wiley & Sons.

[40] Mendell, N. R., Finch, S. J. and Thode, H. C. (1993). Where is the likelihood ratio test powerful for detecting two component normal mixtures? Biometrics, 49, 907-915.

[41] Mendell, N.R., Thode, H.C. and Finch, S. J. (1991). The likelihood ratio test for the two-component normal mixture problem: Power and sample size analysis. Biometrics, 47, 1143-1148.

[42] Pearson, K. (1894). Contributions to the mathematical theory of evolution. Philosophical Transactions of the Royal Society of London A, 185, 71-110.

[43] Quandt, R. E. and Ramsey J. B. (1978). Estimating Mixtures of normal distributions and Switching regressions. Journal of the American statistical association, 73, 730-738.

[44] Quinn, B. G., McLachlan, G. J. and Hjort, N. L. (1987). A note on the Aitkin-Rubin approach to hypothesis testing in mixture models. Journal of the Royal Statistical Society B, 49, 311-314.

[45] Render, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. SIAM Review 26, 195-236.

[46] Schlattmann, P. (2005). On bootstrapping the number of components in finite mixtures of poisson distribution. Statistics and Computing, 15, 179-188.

[47] Schork, N., Allison, D. and Thiel, B. (1996). Mixture distributions in human genetics research. Statistical Methods in Medical Research, 5, 155-178.

[48] Self, S. G. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association, 82, 605-610.

[49] Thode H. C., Finch S. J. and Mendell N. R. (1988). Simulated percentage points for the null distribution of the likelihood ratio test for a mixture of two normals. Biometrics, 44, 1195-1201.

[50] Titterington, D. M. (1984). Recursive parameter estimation using incomplete data. Journal of the Royal Statistical Society B, 46, 257-267.

[51] Titterington, D., Smith, A. and Makov, U.(1985). Statistical analysis of finite mixture Distributions. New York: Wiley

[52] Weldon, W. F. R. (1893). On certain correlated variations in carcinus maenas. Proceedings of the Royal Society of London, 54, 318-329.

[53] Wienke, A., Lichtenstein, P. and Yashin, A. I. (2003). A bivariate frailty model with a cure fraction for modeling familial correlations in disease. Biometrics, 59, 1178-1183.

[54] Wolfe, J. H. (1967). NORMIX: computational methods for estimating the parameters of multivariate normal mixtures of distributions. Research Memorandum SRM 68-2. San diego: U. S. Naval Personnel Research Activity.

[55] Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. Multivariate Behavioral Research, 5, 329-350.

[56] Wolfe, J. H. (1971). A Monte Carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distributions. Technical Bulletin, STB 72-2. San Diego: U.S. Naval Personnel and Training Research Laboratory.

# Appendices

## Appendix A. Derivation of the correlation coefficient for the bivariate normal mixture

Let $\underline{X}$ ~ Bivariate normal mixture with joint pdf

$$g(\underline{X};\underline{\theta_1}) = \pi N_1(\underline{X};\begin{pmatrix}\mu_{11}\\ \mu_{12}\end{pmatrix},\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}) + (1-\pi)N_2(\underline{X};\begin{pmatrix}\mu_{11}+D_1\sigma_1\\ \mu_{12}+D_2\sigma_2\end{pmatrix},\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix})$$

The means and variances are given by

$$\mu_{X_1} = \mu_{11} + (1-\pi)D_1\sigma_1$$

$$\mu_{X_2} = \mu_{12} + (1-\pi)D_2\sigma_2$$

$$\sigma_{X_1}^2 = \sigma_1^2[1 + D_1^2\pi(1-\pi)]$$

$$\sigma_{X_2}^2 = \sigma_2^2[1 + D_2^2\pi(1-\pi)]$$

Next we calculate the expected value of the product to be

$$\begin{aligned}
&E(X_1X_2) = \pi E(X_1X_2 \mid GP1) + (1-\pi)E(X_1X_2 \mid GP2)\\
&= \pi[\mathrm{cov}(X_1X_2 \mid GP1) + E(X_1 \mid GP1)E(X_2 \mid GP1)] + (1-\pi)[\mathrm{cov}(X_1X_2 \mid GP2) + E(X_1 \mid GP2)E(X_2 \mid GP2)]\\
&= \pi(\rho\sigma_1\sigma_2 + \mu_{11}\mu_{12}) + (1-\pi)[\rho\sigma_1\sigma_2 + (\mu_{11}+D_1\sigma_1)(\mu_{12}+D_2\sigma_2)]\\
&= (\rho\sigma_1\sigma_2 + \mu_{11}\mu_{12}) + (1-\pi)(\mu_{11}D_2\sigma_2 + \mu_{12}D_1\sigma_1 + D_1D_2\sigma_1\sigma_2)
\end{aligned}$$

Thus

$$\begin{aligned}
&\mathrm{cov}(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2)\\
&= \rho\sigma_1\sigma_2 + \mu_{11}\mu_{12} + (1-\pi)(\mu_{11}D_2\sigma_2 + \mu_{12}D_1\sigma_1 + D_1D_2\sigma_1\sigma_2) - [\mu_{11} + (1-\pi)D_1\sigma_1][\mu_{12} + (1-\pi)D_2\sigma_2]\\
&= \rho\sigma_1\sigma_2 + \pi(1-\pi)D_1D_2\sigma_1\sigma_2
\end{aligned}$$

Finally, we can express the correlation coefficient for the bivariate normal mixture

$$\begin{aligned}
&\rho_{X_1X_2} = cor(X_1, X_2) = \mathrm{cov}(X_1, X_2)/\sigma_{X_1}\sigma_{X_2}\\
&= \frac{\rho\sigma_1\sigma_2 + \pi(1-\pi)D_1D_2\sigma_1\sigma_2}{\sigma_1\sigma_2\sqrt{[1 + D_1^2\pi(1-\pi)][1 + D_2^2\pi(1-\pi)]}}\\
&= \frac{\rho + \pi(1-\pi)D_1D_2}{\sqrt{[1 + D_1^2\pi(1-\pi)][1 + D_2^2\pi(1-\pi)]}}
\end{aligned}$$

## Appendix B. The derivatives equations of maximum likelihood estimation for the bivariate normal distribution

The maximum likelihood function is

$$L(\hat{\underline{\theta}_0};\underline{X}) = \prod_{i=1}^{n} g(\underline{x_i};\hat{\underline{\theta}_0}) = \prod_{i=1}^{n} N(\underline{X};\underline{\mu_0},\Sigma_0)$$

Then the log of the maximum likelihood function is

$$\log L(\hat{\underline{\theta}_0};\underline{X}) = \log \prod_{i=1}^{n} N(\underline{X};\underline{\mu_0},\Sigma_0)$$

$$= \log \prod_{i=1}^{n} \frac{1}{2\pi |\Sigma|^{1/2}} \exp[-\frac{1}{2}(\underline{x_i}-\underline{\mu_0})^T \Sigma(\underline{x_i}-\underline{\mu_0})]$$

$$= -n\log(2\pi |\Sigma|^{1/2}) - \frac{1}{2}\sum_{i=1}^{n}[(\underline{x_i}-\underline{\mu_0})^T \Sigma(\underline{x_i}-\underline{\mu_0})]$$

Differentiating the above equation with respect to $\underline{\mu_0}$ we get

$$\frac{\partial \log L(\hat{\underline{\theta}_0};\underline{X})}{\partial \underline{\mu_0}} = \sum_{i=1}^{n}\{-\frac{1}{2}\frac{\partial}{\partial \underline{\mu_0}}[(\underline{x_i}-\underline{\mu_0})^T \Sigma(\underline{x_i}-\underline{\mu_0})]\}$$

$$= \sum_{i=1}^{n}(\underline{x_i}-\underline{\mu_0})^T \Sigma^{-1}$$

We set the equation equal to 0, then the $\hat{\underline{\mu_0}}$ is

$$\hat{\underline{\mu_0}} = \frac{\sum_{i=1}^{n}\underline{x_i}}{n}$$

Similarly, differentiating the log of the maximum likelihood function with respect to $\Sigma^{-1}$ we get

$$\frac{\partial \log L(\hat{\underline{\theta}_0};\underline{X})}{\partial \Sigma^{-1}} = \frac{n}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{n}(\underline{x_i}-\underline{\mu_0})(\underline{x_i}-\underline{\mu_0})^T$$

We set the equation equal to 0, then the $\hat{\Sigma}$ is

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\underline{x_i}-\underline{\bar{x}})(\underline{x_i}-\underline{\bar{x}})^T$$

## Appendix C. The expansion of updated values of unknown parameters in EM algorithm

$$\underline{\mu_1}^{(k+1)} = \frac{\sum_{i=1}^{n} \tau_1(\underline{x_i}; \underline{\theta}^{(k)}) \underline{x_i}}{\sum_{i=1}^{n} \tau_1(\underline{x_i}; \underline{\theta}^{(k)})} = \frac{\tau_1(i=1)\underline{x_1} + \tau_1(i=2)\underline{x_2} + ... + \tau_1(i=n)\underline{x_n}}{\tau_1(i=1) + \tau_1(i=2) + ... + \tau_1(i=n)}$$

$$\mu_{11}^{(k+1)} = \frac{\tau_1(i=1)x_{11} + \tau_1(i=2)x_{21} + ... + \tau_1(i=n)x_{n1}}{\tau_1(i=1) + \tau_1(i=2) + ... + \tau_1(i=n)}$$

$$\mu_{12}^{(k+1)} = \frac{\tau_1(i=1)x_{12} + \tau_1(i=2)x_{22} + ... + \tau_1(i=n)x_{n2}}{\tau_1(i=1) + \tau_1(i=2) + ... + \tau_1(i=n)}$$

Similarly,

$$\underline{\mu_2}^{(k+1)} = \frac{\sum_{i=1}^{n} \tau_2(\underline{x_i}; \underline{\theta}^{(k)}) \underline{x_i}}{\sum_{i=1}^{n} \tau_2(\underline{x_i}; \underline{\theta}^{(k)})}$$

$$\mu_{21}^{(k+1)} = \frac{\tau_2(i=1)x_{11} + \tau_2(i=2)x_{21} + ... + \tau_2(i=n)x_{n1}}{\tau_2(i=1) + \tau_2(i=2) + ... + \tau_2(i=n)}$$

$$\mu_{22}^{(k+1)} = \frac{\tau_2(i=1)x_{12} + \tau_2(i=2)x_{22} + ... + \tau_2(i=n)x_{n2}}{\tau_2(i=1) + \tau_2(i=2) + ... + \tau_2(i=n)}$$

The updated covariance matrix is

$$\sigma_1^{2^{(k+1)}} = \sum_{i=1}^{n} \tau_1(i=i)(x_{i1} - \mu_{11}^{(k)})^2 + \sum_{i=1}^{n} \tau_2(i=i)(x_{i1} - \mu_{21}^{(k)})^2$$

$$\sigma_2^{2^{(k+1)}} = \sum_{i=1}^{n} \tau_1(i=i)(x_{i1} - \mu_{12}^{(k)})^2 + \sum_{i=1}^{n} \tau_2(i=i)(x_{i2} - \mu_{22}^{(k)})^2$$

$$\rho\sigma_1^{(k+1)}\sigma_2^{(k+1)} = \sum_{i=1}^{n} \tau_1(i=i)(x_{i1} - \mu_{11}^{(k)})(x_{i2} - \mu_{12}^{(k)}) + \sum_{i=1}^{n} \tau_2(i=i)(x_{i1} - \mu_{21}^{(k)})(x_{i2} - \mu_{22}^{(k)})$$

The updated mixing proportion is

$$\pi^{(k+1)} = \frac{\tau_1(i=1) + \tau_1(i=2) + ... + \tau_1(i=n)}{n}$$

## Appendix D. The empirical power results of the bootstrap test

Table D. 1 The empirical power of the bootstrap test for sample size $n = 50$

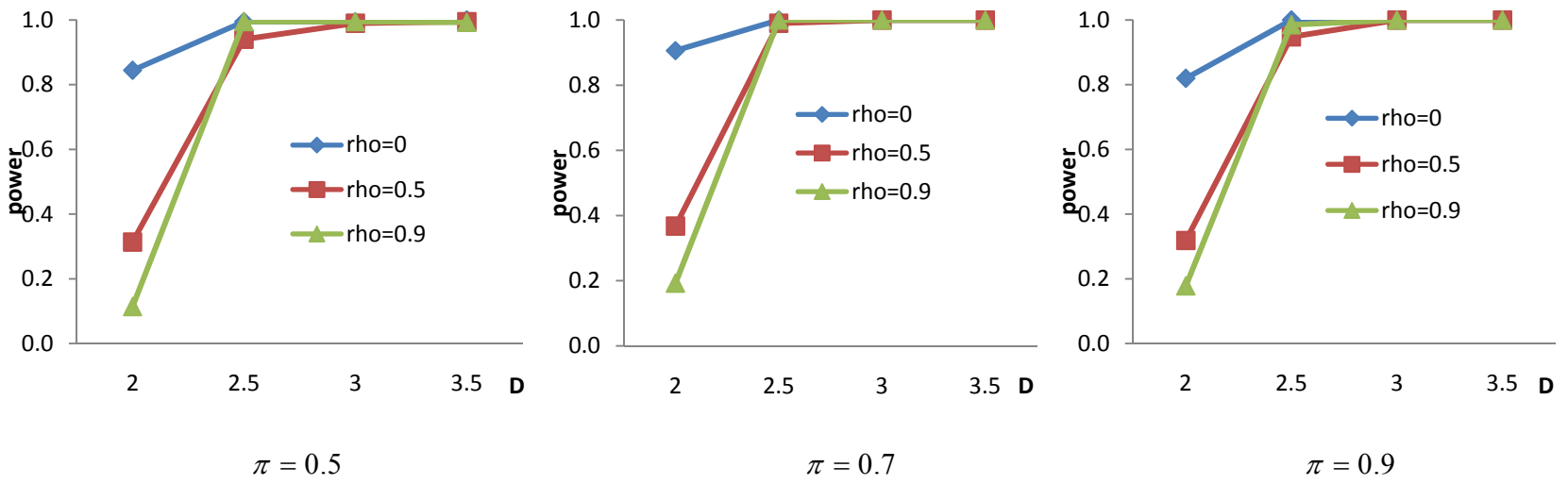
| Mixing proportion | Mean of effect size | within component correlation | Power($\alpha = 0.05$) | Power($\alpha = 0.01$) |
|---|---|---|---|---|
| 0.5 | 2 | 0 | 0.264 | 0.121 |
| | | 0.5 | 0.141 | 0.040 |
| | | 0.9 | 0.104 | 0.030 |
| | 2.5 | 0 | 0.738 | 0.501 |
| | | 0.5 | 0.630 | 0.202 |
| | | 0.9 | 0.388 | 0.385 |
| | 3 | 0 | 0.990 | 0.933 |
| | | 0.5 | 0.958 | 0.780 |
| | | 0.9 | 0.909 | 0.857 |
| | 3.5 | 0 | 1.000 | 0.985 |
| | | 0.5 | 0.963 | 0.859 |
| | | 0.9 | 0.956 | 0.869 |
| 0.7 | 2 | 0 | 0.363 | 0.114 |
| | | 0.5 | 0.151 | 0.052 |
| | | 0.9 | 0.099 | 0.032 |
| | 2.5 | 0 | 0.807 | 0.533 |
| | | 0.5 | 0.667 | 0.237 |
| | | 0.9 | 0.447 | 0.344 |
| | 3 | 0 | 0.980 | 0.928 |
| | | 0.5 | 0.935 | 0.765 |
| | | 0.9 | 0.916 | 0.810 |
| | 3.5 | 0 | 1.000 | 0.988 |
| | | 0.5 | 0.990 | 0.862 |
| | | 0.9 | 0.968 | 0.877 |
| 0.9 | 2 | 0 | 0.259 | 0.111 |
| | | 0.5 | 0.111 | 0.047 |
| | | 0.9 | 0.081 | 0.042 |
| | 2.5 | 0 | 0.585 | 0.395 |
| | | 0.5 | 0.538 | 0.148 |
| | | 0.9 | 0.316 | 0.275 |
| | 3 | 0 | 0.975 | 0.775 |
| | | 0.5 | 0.899 | 0.625 |
| | | 0.9 | 0.815 | 0.751 |
| | 3.5 | 0 | 0.980 | 0.936 |
| | | 0.5 | 0.942 | 0.681 |
| | | 0.9 | 0.852 | 0.731 |

Table D.2 The empirical power of the bootstrap test for sample size $n = 100$

| Mixing proportion | Mean of effect size | within component correlation | Power( $\alpha = 0.05$ ) | Power( $\alpha = 0.01$ ) |
|---|---|---|---|---|
| 0.5 | 2 | 0 | 0.620 | 0.356 |
| | | 0.5 | 0.254 | 0.119 |
| | | 0.9 | 0.143 | 0.049 |
| | 2.5 | 0 | 0.975 | 0.826 |
| | | 0.5 | 0.973 | 0.590 |
| | | 0.9 | 0.798 | 0.489 |
| | 3 | 0 | 1.000 | 1.000 |
| | | 0.5 | 0.990 | 0.990 |
| | | 0.9 | 1.000 | 0.995 |
| | 3.5 | 0 | 0.998 | 0.998 |
| | | 0.5 | 1.000 | 1.000 |
| | | 0.9 | 1.000 | 1.000 |
| 0.7 | 2 | 0 | 0.691 | 0.486 |
| | | 0.5 | 0.264 | 0.104 |
| | | 0.9 | 0.153 | 0.062 |
| | 2.5 | 0 | 0.998 | 0.973 |
| | | 0.5 | 0.973 | 0.686 |
| | | 0.9 | 0.847 | 0.801 |
| | 3 | 0 | 1.000 | 1.000 |
| | | 0.5 | 0.998 | 0.998 |
| | | 0.9 | 1.000 | 1.000 |
| | 3.5 | 0 | 1.000 | 1.000 |
| | | 0.5 | 1.000 | 1.000 |
| | | 0.9 | 1.000 | 1.000 |
| 0.9 | 2 | 0 | 0.551 | 0.365 |
| | | 0.5 | 0.242 | 0.094 |
| | | 0.9 | 0.123 | 0.047 |
| | 2.5 | 0 | 0.928 | 0.837 |
| | | 0.5 | 0.886 | 0.546 |
| | | 0.9 | 0.736 | 0.458 |
| | 3 | 0 | 1.000 | 0.998 |
| | | 0.5 | 0.973 | 0.938 |
| | | 0.9 | 1.000 | 1.000 |
| | 3.5 | 0 | 1.000 | 1.000 |
| | | 0.5 | 0.995 | 0.975 |
| | | 0.9 | 0.990 | 0.983 |

Table D.3 The empirical power of the bootstrap test for sample size $n = 500$

| Mixing proportion | Mean of effect size | within component correlation | Power( $\alpha = 0.05$ ) | Power( $\alpha = 0.01$ ) |
|---|---|---|---|---|
| 0.5 | 2 | 0 | 0.921 | 0.921 |
| | | 0.5 | 0.874 | 0.810 |
| | | 0.9 | 0.647 | 0.477 |
| | 2.5 | 0 | 0.904 | 0.904 |
| | | 0.5 | 0.928 | 0.919 |
| | | 0.9 | 0.923 | 0.923 |
| | 3 | 0 | 0.960 | 0.960 |
| | | 0.5 | 0.941 | 0.941 |
| | | 0.9 | 0.956 | 0.956 |
| | 3.5 | 0 | 0.946 | 0.946 |
| | | 0.5 | 0.923 | 0.921 |
| | | 0.9 | 0.921 | 0.919 |
| 0.7 | 2 | 0 | 1.000 | 1.000 |
| | | 0.5 | 0.983 | 0.943 |
| | | 0.9 | 0.832 | 0.679 |
| | 2.5 | 0 | 1.000 | 1.000 |
| | | 0.5 | 1.000 | 1.000 |
| | | 0.9 | 1.000 | 1.000 |
| | 3 | 0 | 1.000 | 1.000 |
| | | 0.5 | 1.000 | 1.000 |
| | | 0.9 | 1.000 | 1.000 |
| | 3.5 | 0 | 0.998 | 0.998 |
| | | 0.5 | 1.000 | 1.000 |
| | | 0.9 | 0.998 | 0.998 |
| 0.9 | 2 | 0 | 1.000 | 0.998 |
| | | 0.5 | 0.938 | 0.857 |
| | | 0.9 | 0.763 | 0.556 |
| | 2.5 | 0 | 1.000 | 1.000 |
| | | 0.5 | 1.000 | 1.000 |
| | | 0.9 | 0.998 | 0.998 |
| | 3 | 0 | 1.000 | 1.000 |
| | | 0.5 | 1.000 | 1.000 |
| | | 0.9 | 1.000 | 1.000 |
| | 3.5 | 0 | 0.998 | 0.998 |
| | | 0.5 | 1.000 | 1.000 |
| | | 0.9 | 0.998 | 0.998 |




Note: 1. The power results are based on 1000 bootstrap test and 1000 replicates.
2. Significance level $\alpha = 0.01$.
3. D=the average of the two mixture effect sizes, rho=within component correlation.

Figure D.1: Power results of the parametric bootstrap test for sample size $n = 200$

## Appendix E. Model Number Reference

Table E.1 The model numbers at sample size $n = 50$

| mixing proportion | mean of effect size | within component correlation | model number |
|---|---|---|---|
| 0.5 | 2 | 0 | 1 |
| | | 0.5 | 2 |
| | | 0.9 | 3 |
| | 2.5 | 0 | 4 |
| | | 0.5 | 5 |
| | | 0.9 | 6 |
| | 3 | 0 | 7 |
| | | 0.5 | 8 |
| | | 0.9 | 9 |
| | 3.5 | 0 | 10 |
| | | 0.5 | 11 |
| | | 0.9 | 12 |
| 0.7 | 2 | 0 | 13 |
| | | 0.5 | 14 |
| | | 0.9 | 15 |
| | 2.5 | 0 | 16 |
| | | 0.5 | 17 |
| | | 0.9 | 18 |
| | 3 | 0 | 19 |
| | | 0.5 | 20 |
| | | 0.9 | 21 |
| | 3.5 | 0 | 22 |
| | | 0.5 | 23 |
| | | 0.9 | 24 |
| 0.9 | 2 | 0 | 25 |
| | | 0.5 | 26 |
| | | 0.9 | 27 |
| | 2.5 | 0 | 28 |
| | | 0.5 | 29 |
| | | 0.9 | 30 |
| | 3 | 0 | 31 |
| | | 0.5 | 32 |
| | | 0.9 | 33 |
| | 3.5 | 0 | 34 |
| | | 0.5 | 35 |
| | | 0.9 | 36 |