Stony Brook University

The official electronic file of this thesis or dissertation is maintained by the University Libraries on behalf of The Graduate School at Stony Brook University.

Application of Asymptotic LRTS Results to Mixture Problems in Genetic Research

A Dissertation Presented

by

Wonkuk Kim

to

The Graduate School in Partial Fulfillment of the Requirements for the

Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Statistics

Stony Brook University

May 2007

Stony Brook University

The Graduate School

Wonkuk Kim

We, the dissertation committee for the above candidate for the Doctor of Philosophy
degree, hereby recommend acceptance of this dissertation.

Stephen J. Finch
Professor of Applied Mathematics and Statistics
Dissertation Advisor

Nancy R. Mendell
Professor of Applied Mathematics and Statistics
Chairperson of Defense

Wei Zhu
Associate Professor of Applied Mathematics and Statistics

Derek Gordon
Associate Professor of Genetics, Rutgers University
Outside Member

This dissertation is accepted by the Graduate School.

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Application of Asymptotic LRTS Results to Mixture Problems in Genetic Research

by

Wonkuk Kim

Doctor of Philosophy

in

Applied Mathematics and Statistics

Statistics

Stony Brook University

2007

This dissertation considers two mixture problems. The first has known mixing proportions, fixed but arbitrary number of components, and component density function from the one parameter exponential family. The asymptotic null distribution of the likelihood ratio test statistic (LRTS) of the null hypothesis of a single component distribution versus the alternative of two or more components in a mixture is proven to be $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$. The result is shown to hold when the component density function has a nuisance parameter satisfying a linearly independence condition on the partial derivatives. These results are applied to an F-2 breeding experiment in which the mixing proportions are known to be 1/4, 1/2, and 1/4. The scientific problem underlying the second problem is that there is a continuous variable used in assigning a subject in a genetic study to a genotype, which can be modeled by a mixture dis-

tribution with the same known number of components in controls and cases. The component parameters are assumed to be equal but unknown in the two groups. The null hypothesis is that the case mixing proportions are equal to the control mixing proportions. The LRTS for this hypothesis is presented, and the non-centrality parameter of the power function is derived. We compare the power of the LRTS to the power of the chi-square test of independence using genotype classifications from two classification rules, a half-way rule and a Bayesian rule. The LRTS is more powerful asymptotically than the test of independence using either classification rule, with increasing superiority as the frequency of the least common component becomes small.

Key Words: F-2 breeding experiment, genetic association study, case control study, statistical power, non-central parameter, classification rule.

Table of Contents

List of Figures

List of Tables

Acknowledgements

My long journey to a Ph.D. finally comes to the end. It is obviously not possible without the personal and academic support from numerous people around me. Thus my sincere gratitude must go to my family, all faculty members at Stony Brook University and Seoul National University, and all my friends for their love, support, and advice.

I would like to thank my advisor, Professor Stephen J. Finch for his dedication to my dissertation. During my work on the mixture models in genetic research, his unending drive and dedication never cease to amaze me.

It was fortunate for me to meet Professor Derek Gordon at Rutgers University. His insight and outlook on statistical genetics made it possible for me to expand my results. I am grateful to my Ph.D. committee members, Professor Nancy R. Mendell and Professor Wei Zhu for helping me complete this dissertation.

My graduate studies would not have been the same without the social and academic challenges and diversions provided by all my student-colleagues in the department of Applied Mathematics and Statistics. I am particularly thankful to Dr. Seungtai Yoon, Dr. Heejong Sung and Melissa Fazzari. We not only studied, relaxed, and shared the great moments in our lives well together.

Finally, I could not find any expression to my family and InYoung Baek for their endless support, encouragement and love.

Chapter 1

Introduction

Finite mixture models are useful to understand hidden heterogeneous structures in data. Mixture models have been widely applied in many fields including genetics, social sciences, biology, medicine, economics, engineering.

In 1893, Pearson [25] applied a mixture of two normal density functions to data of the ratio of forehead to body length of 1000 crabs sampled from the Bay of Naples. He used the method of moments to estimate the five parameters of the two component heteroscedastic normal mixture model. Rao [26] used the maximum likelihood to estimate the parameters in a mixture distribution by Fisher's scoring method. Tan and Chang [29] showed that maximum likelihood estimation (MLE) is superior to the method of moments for this problem. Dempster et al. [5] formalized an iterative algorithm, which is called the Expectation-Maximization (EM) algorithm. Wu [34] showed the convergence of the EM algorithm.

A number of papers have been written on the testing of the number of components of a mixture problem using the likelihood ratio test (LRT). However, the classical regularity conditions [16] for the asymptotic theory to the LRT do not hold in these mixture testing problems. Hartigan [11] and Liu et al. [19] showed that the asymptotic null distribution of the likelihood ratio test statistic (LRTS) is asymptoti-

cally unbounded in probability at a rate of $\log \log n$. Titterington [31] proved that the asymptotic null distribution of a single univariate density against a mixture of two known densities is a 50:50 mixture of a point mass at zero and a chi-square of degrees of freedom one. Ghosh and Sen [8] showed that the asymptotic LRTS is distributed as a functional form of a Gaussian process when the component parameter space is compact. Other approaches to obtain the percentiles of LRTS of mixtures are to use the bootstrap method by McLachlan [22] and the simulation results by Thode et al. [30].

In genetic research, one approach to detect a major gene is to use a mixture model with three component normal distributions [21]. For example, if the distribution of a phenotype is consistent with three normals with Hardy-Weinberg weights $p^2, 2p(1-p)$, and $(1-p)^2$, where the allele frequency p is estimated from the data, then we may have empirical support for a major locus model for the phenotype. Kruglyak and Lander [14] applied a mixture model to identity by descent (IBD) data for quantitative trait locus (QTL) analysis. They used the likelihood of the sib-pair difference. Later, Fulker and Cherny [7] suggested that a more powerful test can be obtained by considering the likelihood of a bivariate model.

Goffinet et al. [9] considered the LRTS of two component mixtures of normal distributions when the mixing proportions are known. Later, Loisel et al. [20] partially extended this work to three component mixtures of normal distributions when the mixing proportions are $p = (1/4, 1/2, 1/4)$. They applied these results to detect major genes of a QTL in an F2 breeding experiment.

This dissertation extends the asymptotic results of the LRTS of mixtures with known mixing proportions to the case in which the component is not a normal probability density. The rigorous mathematical proof is given in Chapter 3. The results show that the asymptotic null distribution of the LRTS does not depend on the

number of components of the alternative hypothesis. The degeneracy of the Fisher information matrix can cause the collapse of the dimension of the hypothesis space.

Kang et al. [12] suggested using a mixture of the bivariate t-distributions for clustering the single nucleotide polymorphism (SNP) data that are continuously measured by the fluorescence intensity signal. They showed that the use of the maximum likelihood estimation of mixtures performs better than the use of the k-means clustering. They conducted a test of independence followed by the MLE of the parameters of the mixtures.

We recommend the use of the LRTS testing the equality of mixing proportions in two groups under the assumption that the number of components and their parameters are equal but unknown. The major theorem of Chapter 4 provides the asymptotic power of the LRTS of mixtures in terms of the non-centrality parameter (NCP) of a chi-square distribution. The theoretical comparison of the asymptotic power function of the test of independence to the LRTS is given.

In summary, Chapter 2 presents the current asymptotic results on the distribution of the LRTS and other background material. In Chapter 3, our main results in the asymptotic null distribution of a mixture model with known mixing proportions is proven. In Chapter 4, my other main result about a mixture model in a genetic case/control study is proven. We also compare the power of the mixture LRTS to the chi-square test of independence.

Chapter 2

Quadratic Expansions and Asymptotic Calculations in Statistics

2.1 Likelihood Expansion and Likelihood Ratio Test

Suppose that a sample $X_1, \cdots, X_n$ is drawn from $f(\cdot|\theta_0)$. We wish to test the null hypothesis $H_0 : \theta \in \Theta_0$ versus the alternative hypothesis $H_1 : \theta \in \Theta_1$. Let $\Theta = \Theta_0 \cup \Theta_1$. An extension of the Neyman-Pearson test is based on the log-likelihood ratio [32]

$$\Lambda_n^* = \log \frac{\sup_{\Theta_1} \prod_{i=1}^n f(X_i|\theta)}{\sup_{\Theta_0} \prod_{i=1}^n f(X_i|\theta)}. \tag{2.1}$$

If the hypotheses are simple, then the test statistic (2.1) is equivalent to the Neyman-Pearson most powerful test.

Definition 2.1 (LRTS). *The log-likelihood ratio statistic (LRTS) is defined by*

$$LRTS = 2\Lambda_n = 2 \max(\Lambda_n^*, 0) = 2 \log \frac{\sup_{\Theta} \prod_{i=1}^n f(X_i|\theta)}{\sup_{\Theta_0} \prod_{i=1}^n f(X_i|\theta)}. \tag{2.2}$$

The following notation applies throughout this dissertation: $l(\theta|X) = \log f(X|\theta)$ and $L(\theta|X) = \sum_{k=1}^n l(\theta|X_k) = \sum_{k=1}^n \log f(X_k|\theta)$. We sometimes use the shortened notation $l = l(\theta|X)$, $l_0 = l(\theta_0|X)$, $L = L(\theta|X)$, and $L_0 = L(\theta_0|X)$. Every parameter

or random variable may be a vector depending on the problem of interest. Before we discuss the asymptotic results of the LRTS, we illustrate the exact null distribution of the LRTS applied to a random sample from a normal distribution.

Example 2.1 (Exact Case)

Suppose that $X_1, \cdots, X_n$ is drawn from $N(\theta_0, \Sigma)$ with a known non-singular covariance matrix Σ, where $\theta_0 \in \mathbb{R}^d$. We can diagonalize the covariance matrix $\Sigma = P^T \text{diag}(\sigma_1^2, \cdots, \sigma_d^2) P = P^T \Sigma_d P = (\Sigma_d^{1/2} P)^T \Sigma_d^{1/2} P$ with an orthogonal matrix P. Therefore, $\text{cov}(PX, PX) = E_{\theta_0} PX(PX)^T = P(E_{\theta_0} XX^T)P^T = \Sigma_d$. We define $W = \Sigma_d^{-1/2} P(X - \theta_0) \sim N(0, I_{d \times d})$, where $I_{d \times d}$ is the d dimensional identity matrix, and $\phi = \Sigma_d^{-1/2} P(\theta - \theta_0)$. The log-likelihood can be expanded as

$$
\begin{aligned}
2(L - L_0) &= -\sum_{i=1}^{n}(X_i - \theta)^T \Sigma^{-1}(X_i - \theta) + \sum_{i=1}^{n}(X_i - \theta_0)^T \Sigma^{-1}(X_i - \theta_0) \\
&= -n(\theta - \theta_0)^T \Sigma^{-1}(\theta - \theta_0) + \sum (X_i - \theta_0)^T \Sigma^{-1}(\theta - \theta_0) \\
&\quad + \sum (\theta - \theta_0)^T \Sigma^{-1}(X_i - \theta_0) \\
&= -\left(\sqrt{n}(\theta - \theta_0) - \frac{1}{\sqrt{n}}\sum(X_i - \theta_0)\right)^T \Sigma^{-1}\left(\sqrt{n}(\theta - \theta_0) - \frac{1}{\sqrt{n}}\sum(X_i - \theta_0)\right) \\
&\quad + \left(\frac{1}{\sqrt{n}}\sum(X_i - \theta_0)\right)^T \Sigma^{-1}\left(\frac{1}{\sqrt{n}}\sum(X_i - \theta_0)\right) \\
&= -\left(\sqrt{n}\phi - \frac{1}{\sqrt{n}}\sum W_i\right)^T\left(\sqrt{n}\phi - \frac{1}{\sqrt{n}}\sum W_i\right) + \left(\frac{1}{\sqrt{n}}\sum W_i\right)^T\left(\frac{1}{\sqrt{n}}\sum W_i\right) \\
&= -\left(\sqrt{n}\phi - Z\right)^T\left(\sqrt{n}\phi - Z\right) + Z^T Z.
\end{aligned}
$$

The supremum of $2(L(\theta|X) - L_0)$ is equal to $Z^T Z$ when $\phi = Z/\sqrt{n}$. Therefore, the LRTS of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ is a χ_d^2 when $\sqrt{n}(\theta - \theta_0) = \frac{1}{\sqrt{n}}\sum(X_i - \theta_0) \sim N(0, \Sigma)$.

When H_0 is a composite hypothesis such as $\Theta_0 = \{(\theta_1, \cdots, \theta_{d_0}, \theta_{0,d_0+1}, \cdots, \theta_{0,d})\}$, there exists an orthogonal matrix Q such that $\psi = Q\phi$ and $\Psi_0 = Q\Sigma_d^{-1/2} P(\Theta_0 - \theta_0) =$

$\{(\psi_1, \cdots, \psi_{d_0}, 0, \cdots, 0)\}$. Therefore,

$$2(L_{\Theta_0} - L_0) = - \left(\sqrt{n}(\psi_1, \cdots, \psi_{d_0}) - (Z_1, \cdots, Z_{d_0}) \right) \begin{pmatrix} \sqrt{n}\psi_1 - Z_1 \\ \vdots \\ \sqrt{n}\psi_{d_0} - Z_{d_0} \end{pmatrix}$$

$$+ (Z_1, \cdots, Z_{d_0})(Z_1, \cdots, Z_{d_0})^T \text{ and}$$

$$2(L_\Theta - L_0) = - \left(\sqrt{n}\psi - Z \right)^T \left(\sqrt{n}\psi - Z \right) + Z^T Z.$$

The LRTS is equal to $Z_{d_0+1}^2 + \cdots + Z_d^2 = \chi_{d-d_0}^2$. $\qquad \square$

As seen in the example, the distribution of LRTS is obtained by completing the square. To consider the cases in which a sample is drawn from a non-normal distribution, we use definitions of O_P and o_P [18, 32]. The notation $X_n = O_P(1)$ means that a sequence of random variables $\{X_n\}$ is bounded in probability and $X_n = o_P(1)$ means that $\{X_n\}$ converges to zero in probability.

Definition 2.2 (Bounded in probability $O_P(1)$). *A sequence of random variables $\{X_n\}$ is said to be bounded in probability (uniformly tight or $O_P(1)$) if for all $\epsilon > 0$, there exists $M = M(\delta) > 0$ such that $\sup_n P(|X_n| > M) < \delta$. Note that if for all $\delta > 0$, there exist $M = M(\delta) > 0$ and $N = (\delta, M) \in \mathbb{N}$ such that $P(|X_n| > M) < \delta$ for $n \geq N$, then $X_n = O_P(1)$.*

Definition 2.3 (Convergence to zero in probability $o_P(1)$). *A sequence of random variables $\{X_n\}$ converges to zero in probability, denoted by $X_n \xrightarrow{P} 0$, if for all $\epsilon > 0$ and $\delta > 0$, there exists $N = N(\delta, \epsilon) \in \mathbb{N}$ such that $P(|X_n| > \epsilon) < \delta$ for $n \geq N$.*

Using the definitions of boundedness in probability and convergence in probability, the notations O_P and o_P are defined as:

Definition 2.4 (O_P). *Define $X_n = O_P(R_n)$ if $X_n = Y_n R_n$ and $Y_n = O_P(1)$.*

Definition 2.5 (o_P). *Define $X_n = o_P(R_n)$ if $X_n = Y_n R_n$ and $Y_n \xrightarrow{P} 0$.*

Given a family of random sequences, uniform O_P and o_P are defined as follows.

Definition 2.6 (Uniform O_P and o_P). *A family of random sequences $\{X_n(\theta)|\Theta\}$ is said to be uniformly $O_P(1)$ if for all $\delta > 0$, there exist $M = M(\delta) > 0$ and $N = N(\delta, M) \in \mathbb{N}$ such that $P(\sup_{\theta \in \Theta} |X_n(\theta)| > M) < \delta$ for $n \geq N$. The sequence $\{X_n(\theta)\}$ is said to be uniformly $o_P(1)$ if for all $\epsilon > 0$ and $\delta > 0$, there exists $N = N(\delta, \epsilon) \in \mathbb{N}$ such that $P(\sup_{\theta \in \Theta} |X_n(\theta)| > \epsilon) < \delta$ for $n \geq N$.*

Using notation of the gradient $\nabla l = (\partial l / \partial \theta_1, \cdots, \partial l / \partial \theta_d)^T$ and the Hessian matrix $(\nabla^2 l)_{ij} = (\partial^2 l / \partial \theta_i \partial \theta_j)$, the log-likelihood function may be expanded as

$$
\begin{aligned}
L(\theta|X) =& L(\theta_0|X) + \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \left(\nabla l(\theta_0|X_k) \right)^T \sqrt{n}(\theta - \theta_0) \\
&+ \frac{1}{2} \sqrt{n}(\theta - \theta_0)^T \frac{1}{n} \sum_{k=1}^{n} \nabla^2 l(\theta_0|X_k) \sqrt{n}(\theta - \theta_0) \\
&+ \frac{1}{3!} \sum_{i_1,i_2,i_3} \sum_{k=1}^{n} \frac{\partial^3 L(\theta^*|X_k)}{\partial \theta_{i_1} \partial \theta_{i_2} \partial \theta_{i_3}} (\theta_{i_1} - \theta_{i_1 0})(\theta_{i_2} - \theta_{i_2 0})(\theta_{i_3} - \theta_{i_3 0}).
\end{aligned}
\tag{2.3}
$$

Definition 2.7. *The Fisher information $I(\theta)$ is defined as the symmetric quadratic form $I(\theta_0) = E_\theta(\nabla l \nabla l^T)$.*

By the central limit theorem, $\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \nabla l(\theta_0|X_k)$ converges in distribution to a mean zero normal random vector W. The covariance of the mean zero normal random vector W is equal to the Fisher information matrix $I(\theta_0)$. Using the weak law of large numbers, $-\frac{1}{n} \sum_{k=1}^{n} \nabla^2 l(\theta_0|X_k)$ converges in probability to the Fisher information matrix $I(\theta_0)$ at the true parameter value θ_0. By assuming the uniform boundedness of the third order derivatives (i.e., the third of term in equation (2.3) is

$o_P(1))$, the log-likelihood expansion (2.3) can be written as

$$
\begin{aligned}
L(\theta|X) =& L(\theta_0|X) + W^T\sqrt{n}(\theta - \theta_0) - \frac{1}{2}\sqrt{n}(\theta - \theta_0)^T I_{\theta_0}\sqrt{n}(\theta - \theta_0) + o_P(1) \\
=& L(\theta_0|X) - \frac{1}{2}(\sqrt{n}(\theta - \theta_0) - I_{\theta_0}^{-1}W)^T I_{\theta_0}(\sqrt{n}(\theta - \theta_0) - I_{\theta_0}^{-1}W) \qquad (2.4) \\
& + \frac{1}{2}W^T I_{\theta_0}^{-1}W + o_P(1).
\end{aligned}
$$

where $W \sim N(0, I(\theta_0))$. Therefore, the asymptotic null distribution of the LRTS can be obtained in the form of χ^2 distributions as in the example.

2.2 Consistency of the maximum likelihood estimators

As seen in previous sections, it is assumed that the third order terms are uniformly $o_P(1)$ in the parameter space. However, it may not hold or may not be easy to show in the whole parameter space. A common argument is to use the consistency of the maximum likelihood estimators to obtain the uniform $o_P(1)$ quadratic expansion of the log-likelihood function in a small neighborhood of the true parameter value. Because the consistency provides that the maximum likelihood estimators fall into the neighborhood with probability one as the sample size tends to infinity, one only needs to show the uniform quadratic expansion of the log-likelihood function in the neighborhood. Another usual technique is to obtain the expansion of the log-likelihood function in a $(1/\sqrt{n})$-neighborhood of the true parameter value and extend the expansion to the neighborhood by showing that $L - L_0 > 0$ implies the parameters fall into the $(1/\sqrt{n})$-neighborhood with probability one as the sample size tends to infinity.

Definition 2.8 (Consistency). *If the estimator $\hat{\theta}$ converges to the true parameter θ in probability, then the estimator is said to be consistent. That is, $\lim_{n\to\infty}\Pr(|\hat{\theta}_n - \theta| <$*

$\epsilon) = 1$ *for all* $\epsilon > 0$. *It is called strongly consistent, if the sequence of the estimator converges almost surely to the true value. If* $\lim_{n \to \infty} \Pr(\sqrt{n}|\hat{\theta}_n - \theta| < \epsilon) = 1$ *for all* $\epsilon > 0$, *the the sequence of the estimator* $\hat{\theta}_n$ *is called* $\sqrt{n}$*-consistent.*

Let $K = E_0 \log \frac{f}{f_0}$ be the negative Kullback-Leibler information [15]. Concavity of the logarithmic function gives that K is uniquely maximized when $f = f_0$. Identifiability of the parameter space means that $f = f_0$ implies $\theta = \theta_0$. Hence, K has a unique maximum when $\theta = \theta_0$. With $\sup_{\Theta} |K_n(\theta) - K(\theta)| \xrightarrow{P} 0$, where $K_n = \frac{1}{n} \sum_{i=1}^{n} \frac{f(X|\theta)}{f(X|\theta_0)}$, it is not hard to see that any sequence $\{\theta_n\}$ such that $K_n(\theta_n) \geq K_n(\theta_0) - o_P(1)$ converges to θ_0 in probability:

$$K(\theta_0) - K(\theta_n) \leq K_n(\theta_n) - K(\theta_n) + o_P(1)$$

$$\leq \sup_{\Theta} |K_n(\theta) - K(\theta)| + o_P(1) \xrightarrow{P} 0.$$

The consistency of the maximum likelihood estimators is proved by Wald [33] in 1949. Redner [27] extended the proof when the null space does not have identifiability. He established the consistency of the maximum likelihood estimators in a quotient topology with a quotient metric.

In 1954, Chernoff [3] proved the $\sqrt{n}$-consistency of the maximum likelihood estimators. For instance,

$$L - L_0 = \frac{1}{\sqrt{n}} \sum \frac{f'}{f}(\sqrt{n}\theta) - \frac{1}{2}\left[\frac{1}{n}\sum(\frac{f'}{f})^2 + \frac{1}{n}\sum R(\theta^*)_n \epsilon\right](\sqrt{n}\theta)^2$$

where $E_o \sup_{\Theta} |R_n| = M < \infty$ and $|\theta| < \epsilon$. With the consistency of the maximum likelihood estimators and positive definiteness of the information, one can choose $\epsilon > 0$ such that $E_0(\frac{f'}{f})^2 + M\epsilon > 0$. Then, $L - L_0 > 0$ implies that $\sqrt{n}\theta < C$ for some $C > 0$. Liu and Shao [18] also used the same argument to obtain the $\sqrt{n}$-consistency

9

with Hellinger distance of mixture problems.

2.3 Boundary Parameter Problem

Self and Liang [28] characterized the LRTS using a projection of a normal random variable onto a tangent cone of the parameter space. By letting $Z = I^{-1}W$ and absorbing $\sqrt{n}$ in θ in the equation (2.4), the log-likelihood function is in the form of

$$2(L - L_0) = -(\theta - Z)^T I(\theta - Z) + Z^T I Z + o_P(1), \tag{2.5}$$

where $Z \sim N(0, I^{-1})$. When the true value $(0, \cdots, 0)$ of the parameter is on the boundary of $[0, \infty) \times \mathbb{R}^{d-1}$, the maximum is not realized if $Z_1 < 0$. The maximum like-lihood estimate in the $o_P(1)$ sense can be obtained by projecting Z onto $[0, \infty) \times \mathbb{R}^{d-1}$ with the inner product $\langle a, b \rangle = a^T I b$. Let (I^{ij}) be the inverse matrix of I. Then, the first column of the inverse matrix is orthogonal to all vectors on the boundary of the parameter space since $(0, a_2, \cdots, a_d)I \begin{pmatrix} I^{11} \\ I^{21} \\ \vdots \\ I^{d1} \end{pmatrix} = 0$ for all a_j. Therefore $(Z_1, Z_2 - \theta_2, \cdots, Z_d - \theta_d)$ is parallel to $(I^{11}, I^{21}, \cdots, I^{d1}) = (Z_1, Z_2 I^{21}/I^{11}, \cdots, Z_d I^{d1}/I^{11})I^{11}/Z_1$ and $\theta_j = Z_j - (I^{j1}/I^{11})Z_1$ for $j = 2, \cdots, d$. The asymptotic maximum likelihood estimate (MLE) may be written as

$$\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_d \end{pmatrix} \mathbf{1}(Z_1 \geq 0) + \begin{pmatrix} 0 \\ Z_2 - (I^{21}/I^{11})Z_1 \\ \vdots \\ Z_d - (I^{d1}/I^{11})Z_1 \end{pmatrix} \mathbf{1}(Z_1 < 0). \tag{2.6}$$

After the MLE in (2.6) is substituted into the log-likelihood expansion (2.5), then

$$LRTS = \sup_{\theta} 2(L - L_0) = Z^T I Z - \frac{Z_1^2}{I^{11}} \mathbf{1}(Z_1 < 0) + o_P(1).$$

Using the spectral decomposition of Z with an orthogonal matrix P such that

$$PZ = (Z_1, Z_2^*, \cdots, Z_d^*)^T \sim N(0, \text{diag}(1/I^{11}, \text{var} Z_2^*, \cdots, \text{var} Z_d^*)),$$

the asymptotic distribution of $\sup_{\theta} 2(L - L_0)$ reduces to $\frac{1}{2}\chi_{d-1}^2 + \frac{1}{2}\chi_d^2$. Additional examples may be found in Self and Liang [28].

2.4   Asymptotic Power of the LRTS

In the asymptotic theory of the likelihood ratio test, the asymptotic power of the LRTS along a contiguous family of hypotheses $H_n : \theta = \theta_0 + c/\sqrt{n}$ can be expressed in terms of the vector c and the information matrix I_0 at θ_0. Let $W(X|\theta) = \frac{1}{\sqrt{n}} \sum \frac{\nabla f}{f}(X|\theta)$ and $\theta_n = \theta_0 + c/\sqrt{n}$. Under the assumption that H_n is true, $W(X|\theta_0) = W(X|\theta_n) + \frac{1}{n} \sum \frac{\nabla^2 f}{f}(X|\theta_n) \cdot c + o_P(1)$. From the expansion (2.4), the LRTS is still asymptotically equivalent to $W^T I_0^{-1} W$. Therefore, the LRTS asymptotically follows a non-central chi-square distribution with the non-centrality parameter $c^T I_0 c$ if the null hypothesis H_0 is simple. When the null hypothesis is composite, we may decompose $c = c^{\parallel} + c^{\perp}$ with respect to an inner product $\langle a, b \rangle = a I_0 b$ such that $\langle c^{\parallel}, c^{\perp} \rangle = 0$ and $\theta_0 + \frac{c^{\parallel}}{\sqrt{n}} \in \Theta_0$. Therefore, the non-centrality parameter can be obtained from calculating $(c^{\perp})^T I_0 c^{\perp}$.

2.5 Likelihood Ratio Test of Mixtures

Consider a mixture of two components:

$$h(x|\theta_1, \theta_2, p) = pf(x|\theta_1) + (1-p)f(x|\theta_2). \qquad (2.7)$$

It is natural to try to use the LRTS to detect a mixture: $H_0 : h(x|\theta_1, \theta_2, p) = f(x|\theta)$ against $H_1 : h(x|\theta_1, \theta_2, p) = pf(x|\theta_1) + (1-p)f(x|\theta_2)$ where $\theta_1 \neq \theta_2$ and $p \neq 0$ or 1. The first obstacle to apply the classical asymptotic theory is identifiability, i.e., in the null space $\Theta_0 = \{(\theta_1, \theta_2, p) : p = 0 \text{ or } p = 1 \text{ or } \theta_1 = \theta_2\}$, $h(x|\theta_1, \theta_2, p) = h(x|\theta_1^*, \theta_2^*, p^*)$ does not implies $(\theta_1, \theta_2, p) = (\theta_1^*, \theta_2^*, p^*)$. It makes one hard to find a uniform expansion of the log-likelihood function. Another problem comes from the singularity of the information matrix. For convenience, we use the following model instead of (2.7):

$$h(x|\theta_1, \theta_2, p) = pf(x|\theta_1 + q\theta_2) + qf(x|\theta_1 - p\theta_2)$$

such that $H_0 : p = 0$ or $p = 1$ or $\theta_2 = 0$. By restricting $p \in [0, 1/2]$, $H_0 : p\theta_2 = 0$. By direct differentiations, we have

$$\frac{\partial \log h}{\partial \theta_1}\bigg|_{p\theta_2=0} = \frac{f'(x|\theta_1)}{f(x|\theta_1)}$$

$$\frac{\partial \log h}{\partial \theta_2}\bigg|_{p\theta_2=0} = 0$$

$$\frac{\partial \log h}{\partial p}\bigg|_{p\theta_2=0} = \frac{f(x|\theta_1 + \theta_2) - f(x|\theta_1) - \theta_2 f'(x|\theta_1)}{f(x|\theta_1)}.$$

The information matrix on H_0 is singular with one or two dimensional degeneracies depending on $p = 0$ or $\theta_2 = 0$. Other problem for a mixture is the convergence of LRTS. It was proved that LRTS of mixture does not converge when the parameter space is unbounded. It diverges at a rate of $\log \log n$ for normal mixtures [19] and

Gamma mixtures [17] when one component is completely known. They showed that the convergent part of LRTS converges to an extreme value distribution. However, as they mentioned, the rate of convergence is too slow to apply to practical problems. Therefore, the parameter space is set to be compact to avoid divergent LRTS. Dacunha-Castelle and Gassiat [4] introduced a locally conic parametrization to characterize LRTS of mixture problems. They used $\mathcal{L}^2$ norm in the space of the mixture probability density functions and expressed LRTS in terms of the supremum of some Gaussian processes. Liu and Shao [18] introduced a generalized differentiability in quadratic mean and used a Hellinger distance. They also expressed LRTS of mixtures in terms the Gaussian processes.

Definition 2.9 (Gaussian Process). *A stochastic process $\{W_t : t \in A\}$ with an index set A is said to be a Gaussian process if any finite collection $(W_{t_1}, \cdots, W_{t_k})$ follows a multi-variate normal distribution. Note that a Gaussian process is completely characterized by its mean and covariance kernel.*

Th Gaussian process expression for the LRTS is

$$LRTS = \sup_{\theta \in \Theta} \{\max(G_\theta, 0)\}^2,$$

where θ is a parameter in the component probability density and G_θ is a Gaussian process with mean zero with covariance $r(\theta_1, \theta_2) = cov(G_{\theta_1}, G_{\theta_2})$ that must be computed from a given model. Unfortunately, this does not have a simple series expression to perform a test. In order to find the percentiles of this asymptotic null distribution of LRTS, one needs to simulate a Gaussian process by discretizing Θ, generating a large number of independent normal distributions and transforming them to realize the specified covariance kernel by the Choleski decomposition.

Bootstrapping was suggested by McLachlan [22]. The MLE under the null hy-

pothesis is obtained. Bootstrap samples are selected from the fitted null distribu-
tions. The LRTS is calculated for each bootstrap sample to generate the empirical
percentiles which are subsequently used for testing mixtures.

Chapter 3

Mixtures with known proportions

3.1 Notations, Assumptions and Definitions

In this chapter, we assume that the probability density function is in the form of

$$h(x|\theta_1, \cdots, \theta_d) = \sum_{i=1}^{d} p_i f(x|\theta_i), \tag{3.1}$$

where all $0 < p_i < 1$ are known and $\{f(x|\theta)\}$ is a family of one parameter probability densities with $\theta \in \Theta \subset \mathbb{R}^d$. The information matrix when $\theta_i = \theta_0$ for all $i = 1, \cdots, d$ can be easily computed as

$$I_{\theta_0} = E_0(\frac{f'}{f})^2 \begin{pmatrix} p_1^2 & \cdots & p_1 p_d \\ \vdots & & \vdots \\ p_d p_1 & \cdots & p_d^2 \end{pmatrix}$$

Since $(p_1, \cdots, p_d)^T (p_1, \cdots, p_d)$ is a $d \times d$ matrix with the rank one, the Fisher information has $(d-1)$ dimensional degeneracy. The assumptions used in this chapter are:

(A1) Θ is compact in $\mathbb{R}^d$.

(A2) h is identifiable when $\theta_i = \theta_0$ for all $i = 1, \cdots, d$, i.e.,

$h(\cdot|\theta_1, \cdots, \theta_d) = f(\cdot|\theta_0) \Leftrightarrow$ all $\theta_i = \theta_0$

(A3) $0 < E_{\theta_0}(\frac{f''}{f})^2 = M_1 < \infty$.

(A4) The log-likelihood function $\log h$ is continuously fifth differentiable with respect to θ, $\log h(\cdot|\theta) \in C^5(\Theta)$, and $E_{\theta_0} \sup_{\theta \in \Theta} |(\log h)^{(k)}(X|\theta)| < \infty$ for $k = 0, \cdots, 5$. Let $\max_{1 \le j,k,l,m,n \le d} E_{\theta_0} \sup_{\theta \in \Theta} |\frac{\partial^5 \log h(X|\theta)}{\partial \theta_j \partial \theta_k \partial \theta_l \partial \theta_m \partial \theta_n}| = M_2$.

(A5) $(\theta_0, \cdots, \theta_0)$ is an interior point in Θ.

3.2 Asymptotic LRTS for Mixture with Known Proportions

Let $X = (X_1, \cdots, X_n)$ be n independent random variables with probability density function $\{h(x|\theta)\}$. Assume that the true value $\theta_0 = (\theta_{01}, \cdots, \theta_{0d})$ of parameter is an interior point in a compact space $\Theta \subset \mathbb{R}^d$.

Lemma 3.1. *Assume* **(A1)**, **(A2)**, **(A4)**, *and* **(A5)**. *Suppose that $h_i = \partial h/\partial \theta_i = 0$ for all $i = 1, \cdots, d$ under the null hypothesis $H_0 : \theta = \theta_0$ (that is equivalent to zero information matrix $I(\theta_0) = 0$) and the Hessian matrix $h_{ij} = h_{11} \lambda_i \delta_{ij}$ with constants $\lambda_i > 0$. h_{11} is none-zero with positive probability at θ_0. Then the likelihood expansion can be written as:*

$$\max\left(0, L(\theta|X) - L(\theta_0|X)\right) = \max\left(0, \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{\frac{h_{11}}{h}(X|\theta_0)}{\sqrt{E_0(\frac{h_{11}}{h})^2}} \sqrt{n} D(\theta) - \frac{1}{2} n D^2(\theta)\right)$$

$$+ o_P(1), \tag{3.2}$$

provided $D(\theta) = \sqrt{E_0(\frac{h_{11}}{h})^2} \frac{1}{2} \sum_{j=1}^{d} \lambda_j (\theta_j - \theta_{0j})^2 < \epsilon$ for some small $\epsilon > 0$. Moreover, the likelihood ratio test statistic asymptotically follows $\{\max(0, Z)\}^2$ where Z is a one dimensional standard normal random variable.

Proof. We expand the log-likelihood function up to the fifth order to obtain an approximation of the log-likelihood in the sense of $o_P(1)$. For simplicity, all derivative functions are assumed evaluated at θ_0 unless otherwise stated. Without loss of generality, we assume $\theta_0 = 0 \in \mathbb{R}^d$.

$$
\begin{aligned}
L =& L_0 + \sum_{i=1}^{d}\sum_{r=1}^{n}\frac{\partial L}{\partial \theta_i}\theta_i + \frac{1}{2!}\sum_{i,j}\sum_{r=1}^{n}\frac{\partial^2 L}{\partial\theta_i\partial\theta_j}\theta_i\theta_j + \frac{1}{3!}\sum_{i,j,k}\sum_{r=1}^{n}\frac{\partial^3 L}{\partial\theta_i\partial\theta_j\partial\theta_k}\theta_i\theta_j\theta_k \\
&+ \frac{1}{4!}\sum_{i,j,k,l}\sum_{r=1}^{n}\frac{\partial^4 L}{\partial\theta_i\partial\theta_j\partial\theta_k\partial\theta_l}\theta_i\theta_j\theta_k\theta_l + \frac{1}{5!}\sum_{i,j,k,l,m}\sum_{r=1}^{n}\frac{\partial^5 L(\tilde\theta|X_r)}{\partial\theta_i\partial\theta_j\partial\theta_k\partial\theta_l\partial\theta_m}\theta_i\theta_j\theta_k\theta_l\theta_m \\
=& L_0 + \frac{1}{2!}\sum_{i}\sum_{r=1}^{n}\frac{h_{ii}}{h}\theta_i^2 + \frac{1}{3!}\sum_{i,j,k}\sum_{r=1}^{n}\frac{h_{ijk}}{h}\theta_i\theta_j\theta_k + \frac{1}{4!}\sum_{ijkl}\sum_{r=1}^{n}\frac{h_{ijkl}}{h}\theta_i\theta_j\theta_k\theta_l \\
&- \frac{1}{4!}\sum_{i}\sum_{r=1}^{n}3\Big(\frac{h_{ii}}{h}\Big)^2\theta_i^4 - \frac{6}{4!}\sum_{i<j}\sum_{r=1}^{n}\frac{h_{ii}}{h}\frac{h_{jj}}{h}\theta_i^2\theta_j^2 \qquad\qquad (3.3) \\
&+ \frac{1}{5!}\sum_{i,j,k,l,m}\sum_{r=1}^{n}\frac{\partial^5 L(\tilde\theta|X_r)}{\partial\theta_i\partial\theta_j\partial\theta_k\partial\theta_l\partial\theta_m}\theta_i\theta_j\theta_k\theta_l\theta_m \\
=& L_0 + \frac{1}{\sqrt{n}}\sum_{r=1}^{n}\frac{h_{11}}{h}\Big(\sqrt{n}\frac{1}{2}\sum_i\lambda_i\theta_i^2\Big) - \frac{1}{2n}\sum_{r=1}^{n}\Big(\frac{h_{11}}{h}\Big)^2\Big(\sqrt{n}\frac{1}{2}\sum_i\lambda_i\theta_i^2\Big)^2 \\
&+ \frac{1}{3!}\sum_{i,j,k}\sum_{r=1}^{n}\frac{h_{ijk}}{h}\theta_i\theta_j\theta_k + \frac{1}{4!}\sum_{ijkl}\sum_{r=1}^{n}\frac{h_{ijkl}}{h}\theta_i\theta_j\theta_k\theta_l \qquad\qquad (3.4) \\
&+ \frac{1}{5!}\sum_{i,j,k,l,m}\sum_{r=1}^{n}\frac{\partial^5 L(\tilde\theta|X_r)}{\partial\theta_i\partial\theta_j\partial\theta_k\partial\theta_l\partial\theta_m}\theta_i\theta_j\theta_k\theta_l\theta_m.
\end{aligned}
$$

In a neighborhood of θ_0 satisfying $D^*(\theta) = \sum_i \lambda_i\theta_i^2 \leq c/\sqrt{n}$ for any $c > 0$, the second line in the equation (3.4) is $o_P(1)$ since $\frac{h_{ijk}}{h}$ is a mean zero random variable and $|\theta_i\theta_j\theta_k| = O_P(1/n^{\frac{3}{4}})$. The third line is $o_P(1)$ since $\max E_0 \sup_{\theta\in\Theta}\big|\frac{\partial^5 L(\theta|X_i)}{\partial\theta_j\theta_k\theta_l\theta_m\theta_n}\big| = M_2 < \infty$. It remains to prove the expansion holds uniformly when $D^*(\theta) < \epsilon$ for some $\epsilon > 0$. For simplicity, we can assume that all $\lambda_i = 1$ because they are positive and finite. $g(X)$ denotes $h_{11}(X|\theta_0)/h(X|\theta_0)$. By the assumption, $0 < M_1 = E_0 g^2 < \infty$. First we choose a small $\epsilon > 0$ such that $(1/4)M_1 - (d^5/5!)M_2\sqrt{\epsilon} > 0$, where $\epsilon = \epsilon(M_1, M_2, d)$

depends only on the model. $L - L_0$ is bounded above in $D^*(\theta) < \epsilon$ by following inequality:

$$L - L_0 \leq \left(\frac{1}{\sqrt{n}} \left| \sum \frac{h_{11}}{h} \right| + \frac{d^3}{3!} \max_{i,j,k} \left| \sum \frac{h_{ijk}}{h} \right| \sqrt{\epsilon} \right) \sqrt{n} D^*(\theta)$$
$$- \frac{n}{2} D^*(\theta) \left(\frac{1}{n} \sum \left(\frac{h_{11}}{h} \right)^2 - \frac{2d^4}{4!} \max_{i,j,k,l} \frac{1}{n} \left| \sum \frac{h_{ijkl}}{h} \right| \right.$$
$$\left. - \frac{2d^5}{5!} \max_{i,j,k,l,m} \frac{1}{n} \sum \sup_{\Theta} \left| \frac{\partial^5 L}{\partial \theta_i \cdots \partial \theta_m} \right| \sqrt{\epsilon} \right)$$

For a given $\delta > 0$, we can choose $N_0 = N_0(M_1, \delta, d)$, $N_1 = N_1(M_1, \delta, d)$, $N_2 = N_2(M_1, \delta)$, $N_3 = N_3(M_2, \delta, d) \in \mathbb{N}$ such that

$$P_0 \left(\frac{1}{\sqrt{n}} \left| \sum \frac{h_{11}}{h} \right| + \frac{d^3}{3!} \max_{i,j,k} \left| \sum \frac{h_{ijk}}{h} \right| \sqrt{\epsilon} > M_1 \right) < \frac{\delta}{4} \text{ for } n \geq N_0$$

$$P_0 \left(\max_{1 \leq i,j,k,l \leq d} \left| \frac{1}{n} \sum_{r=1}^{n} \frac{h_{ijkl}}{h} (X_r | \theta_0) \right| > \frac{4! M_1}{8 d^4} \right) < \frac{\delta}{4} \text{ for } n \geq N_1.$$

$$P_0 \left(\left| \frac{1}{n} \sum_{r=1}^{n} g^2(X_r) - M_1 \right| > \frac{M_1}{2} \right) < \frac{\delta}{4} \text{ for } n \geq N_2,$$

$$P_0 \left(\max_{1 \leq i,j,k,l,m \leq d} \frac{1}{n} \sum_{r=1}^{n} \sup_{\theta \in \Theta} \left| \frac{\partial^5 L(\theta | X_r)}{\partial \theta_i \cdots \partial \theta_m} \right| > \frac{M_2}{2} \right) < \frac{\delta}{4} \text{ for } n \geq N_3.$$

Define $B_n(\delta | M_1, M_2, d) = \{(X_1, \cdots, X_n) : \frac{1}{\sqrt{n}} \left| \sum \frac{h_{11}}{h} \right| + \frac{d^3}{3!} \max_{i,j,k} \left| \sum \frac{h_{ijk}}{h} \right| \sqrt{\epsilon} < M_1,$ $\max_{1 \leq i,j,k,l \leq d} \left| \frac{1}{n} \sum_{r=1}^{n} \frac{h_{ijkl}}{h} (X_r | \theta_0) \right| < \frac{4! M_1}{8 d^4}, \left| \frac{1}{n} \sum_{r=1}^{n} g^2(X_r) - M_1 \right| < \frac{M_1}{2}, \max_{1 \leq i,j,k,l,m \leq d}$ $\frac{1}{n} \sum_{r=1}^{n} \sup_{\theta \in \Theta} \left| \frac{\partial^5 L(\theta | X_r)}{\partial \theta_i \cdots \partial \theta_m} \right| < \frac{M_2}{2}\}$, then $P_0(B_n) \geq 1 - \delta$ if $n \geq N = \max(N_0, N_1, N_2, N_3)$. For $X = (X_1, \cdots, X_n) \in B_n$, $L - L_0 > 0$ implies $D^*(\theta) < c/\sqrt{n}$ for some $c > 0$. This completes the proof. $\qquad \square$

If the signs of λ_i's are not the same, then the Lemma does not hold. In that case, $D(\theta)$ does not measure the distance from the true probability density function in the space of probability density functions.

Now we consider a hypothesis testing problem of a mixture family:

$$h(x|\theta_1, \cdots, \theta_d) = \sum_{i=1}^{d} p_i f(x|\theta_i) \tag{3.5}$$

$$H_0 : \theta \in \Theta_0 \text{ versus } H_1 : \theta \in \Theta - \Theta_0 \tag{3.6}$$

with known $0 < p_i < 1$ for $i = 1, \cdots, d$, and $\Theta_0 = \text{span}(e) \subset \Theta$, where $e = (1, \cdots, 1) \in \mathbb{R}^d$.

Theorem 3.1. *Assume that* **(A1)** *through* **(A5)** *hold and* $f' \neq f''$ *with a positive probability at* θ_0*. The LRTS to test (3.6) of the model (3.5) follows asymptotically* $\{\max(0, Z)\}^2$ *under* H_0*, where* Z *is a one dimensional standard normal random variable.*

Proof. It is enough to show that there exists a reparametrization satisfying conditions in Lemma 3.1 for $(d - 1)$ parameters across the one dimensional null space. Define a d-dimensional vector $p = (p_1, \cdots, p_d)$. Since we assumed that all mixing proportions are known and positive, it is possible to define an inner product $\langle a, b \rangle_p = a^T \text{diag}(p^T) b$ in $\mathbb{R}^d$. One can choose an orthogonal frame $\{w_1 = e, w_2, \cdots, w_d\}$ including $e = (1, \cdots, 1)$ with respect to the inner product just defined. Now introduce a new parameters ϕ defining $\theta_i - \theta_{0i} = \sum_{k=1}^{d} w_{ki} \phi_k$, where w_{ki} is the i-th element of w_k and θ_{0i} is the i-th element of the true value of the parameter θ. We also use notations $h_{i_1, \cdots, i_r}(x|\theta(\phi)) = \partial^r h(x|\theta(\phi))/(\partial \phi_{i_1} \cdots \partial \phi_{i_r})$. It is easy to see that $h_i(x|\theta_0) = \langle w_1, w_i \rangle_p f'(x|\theta_0)$ and $h_{ij}(x|\theta_0) = \langle w_i, w_j \rangle_p f''(x|\theta_0)$. From how we selected $\{w_i\}$, all $h_i(x|\theta_0)$ for $i \geq 2$ and non-diagonal terms $h_{ij}(x|\theta_0)$ for $i \neq j$ vanish. Moreover, every second order diagonal term has strictly positive coefficient $\langle w_i, w_i \rangle_p$ of the function $f''(x|\theta_0)$. Therefore, with $\lambda_i = \langle w_i, w_i \rangle_p / \langle w_1, w_1 \rangle_p$, Lemma 3.1 can be

applied to the log-likelihood function:

$$L(\theta(\phi)|X) = L(\theta_0|X) + \frac{1}{\sqrt{n}}\Big(\sum_{k=1}^{n}\frac{f'}{f}(X_k|\theta_0), \frac{1}{2}\sum_{k=1}^{n}\frac{f''}{f}(X_k|\theta_0)\Big)\sqrt{n}\big(\phi_1, D^*(\phi)\big)^T$$

$$- \frac{1}{2}\sqrt{n}\big(\phi_1, D^*(\phi)\big)E_0\begin{pmatrix}\left(\frac{f'}{f}\right)^2 & \left(\frac{f'}{f}\frac{f''}{2f}\right) \\ \left(\frac{f'}{f}\frac{f''}{2f}\right) & \left(\frac{f''}{2f}\right)^2\end{pmatrix}\sqrt{n}\big(\phi_1, D^*(\phi)\big)^T + o_P(1)$$

Because the null space $\Theta_0 = \mathrm{span}(e) = \{D^*(\phi) = 0\}$, this completes the proof. $\qquad\square$

Corollary 3.1. *Assume that* **(A1)** *through* **(A5)** *hold. Suppose that* $\{f(x|\theta)\}$ *is a one parameter exponential family, then the asymptotic distribution of LRTS is* $\{\max(0, Z)\}^2$ *where* Z *is a one dimensional standard normal random variable.*

Proof. Use a canonical form $f(x|\theta) = A(\theta)B(x)\exp(\theta\tau(x))$. It suffices to show that f' is not proportional to f''. Suppose not. Direct differentiation gives $A(\theta_0)\tau^2(x) + (2A'(\theta_0) - CA(\theta_0))\tau(x) + A''(\theta_0) - CA'(\theta_0) = 0$ for some constant C. Therefore, $\tau(x)$ must be a constant. $\qquad\square$

3.3   Presence of a Nuisance Parameter

When a mixture model includes one or more nuisance parameter, the asymptotic result depends on the linear dependency of the derivative of the log-likelihood function with respect to the nuisance parameter and the first or the second derivatives with respect to other parameters. Consider a mixture with known proportions when the probability density function includes a nuisance parameter η:

$$h(x|\theta_1, \cdots, \theta_d, \eta) = \sum_{i=1}^{d} p_i f(x|\theta_i, \eta), \tag{3.7}$$

20

If f_θ, $f_{\theta\theta}$ and f_η are linearly independent, then with assumptions (A0) through (A6) the log-likelihood function can be written as:

$$L = L_0 + W^T \sqrt{n} \begin{pmatrix} \phi_1 \\ D^*(\phi) \\ \eta \end{pmatrix} - \frac{1}{2}\sqrt{n}(\phi_1, D^*(\phi), \eta) I_0 \sqrt{n} \begin{pmatrix} \phi_1 \\ D^*(\phi) \\ \eta \end{pmatrix} + o_P(1), \quad (3.8)$$

where W asymptotically follows a normal distribution with mean zero and variance I_0:

$$W = \frac{1}{\sqrt{n}}\left(\sum \frac{f_\theta}{f}, \sum \frac{f_{\theta\theta}}{f}, \sum \frac{f_\eta}{f}\right)^T$$

$$I_0 = E_0 \begin{pmatrix} (\frac{f_\theta}{f})^2 & \frac{f_\theta f_{\theta\theta}}{f\cdot 2f} & \frac{f_\theta f_\eta}{f\cdot f} \\ \frac{f_\theta f_{\theta\theta}}{f\cdot 2f} & (\frac{f_{\theta\theta}}{2f})^2 & \frac{f_{\theta\theta} f_\eta}{2f\cdot f} \\ \frac{f_\theta f_\eta}{f\cdot f} & \frac{f_{\theta\theta} f_\eta}{2f\cdot f} & (\frac{f_\eta}{f})^2 \end{pmatrix}$$

Therefore, the asymptotic null distribution of LRTS follows $\{\max(0, Z)\}^2$, where $Z \sim N(0,1)$.

Example 3.1 Mixture of Weibull Distributions

When $f(x|\theta, \eta) = \theta\eta x^{\eta-1}\exp(-\theta x^\eta)$, we have

$$\begin{aligned}
\frac{\partial f}{\partial \theta} &= \eta x^{\eta-1}(1 - x^\eta)\exp(-\theta x^\eta) \\
\frac{\partial^2 f}{\partial \theta^2} &= -\eta x^{2\eta-1}(1 - x^\eta)\exp(-\theta x^\eta) \\
\frac{\partial f}{\partial \eta} &= \theta x^{\eta-1}(1 - x\log x)\exp(-\theta x^\eta) \\
\frac{\partial^2 f}{\partial \eta^2} &= -\theta^2 x^{2\eta-1}(1 - x\log x)(\log x)\exp(-\theta x^\eta)
\end{aligned}$$

There are two possible models: (1) the shape parameter η is a nuisance parameter.

21

(2) the scale parameter θ is a nuisance parameter. In either case, the asymptotic quadratic expansion can be written in the form of (3.8). Since they are linearly independent at any $(\theta, \eta) \in (0, \infty) \times (0, \infty)$, the asymptotic null distribution of LRTS of homogeneity of Weibull components when the mixing proportions are known is $\{\max(0, Z)\}^2$ with $Z \sim N(0, 1)$. □

When the first derivative of the probability density function with respect to a nuisance parameter is linearly dependent on one of the other derivatives. For instance, in a mixture of normal distribution components with a nuisance parameter standard deviation σ, f_σ is linearly dependent to $f_{\theta\theta}$ in H_0. Then, (3.8) does not hold for normal mixtures with a nuisance parameter σ. Let $f(x|\theta, \eta) = \frac{\eta}{\sqrt{2\pi}} \exp(-\eta^2(x - \theta)^2/2)$ with $\eta = 1/\sigma$.

$$f_\theta = \frac{1}{\sqrt{2\pi}}(x - \theta)\eta^3 \exp(-\eta^2(x - \theta)^2/2)$$

$$f_{\theta\theta} = \frac{1}{\sqrt{2\pi}}(-\eta^3 + (x - \theta)^2\eta^5) \exp(-\eta^2(x - \theta)^2/2)$$

$$f_\eta = \frac{1}{\sqrt{2\pi}}(1 - \eta^2(x - \theta)^2) \exp(-\eta^2(x - \theta)^2/2)$$

$$f_{\eta\eta} = \frac{1}{\sqrt{2\pi}}(-2\eta(x - \theta)^2 - \eta(x - \theta)^2 + \eta^3$$

Since $f_{\theta\theta} = -\eta^3 f_\eta$, the expansion (3.8) does not contain all limit points of the score vectors. We provide the simulation results for this case in following section.

3.4 Simulation

We use a random starting point (RSP) strategy in our simulation studies [6]. Given ordered data $X_{(1)}, \cdots, X_{(n)}$ chosen from $N(0, 1)$ for the null distribution to test a normal mixture, select a random sample $(N_1, \cdots, N_d)$ from a Multinomial(p).

We start an iteration algorithm such as expectation-maximization (EM) method with the following initial values:

$$\theta_j^{(0)} = \frac{1}{N_j} \sum_{k=N_1+\cdots+N_{j-1}+1}^{N_1+\cdots+N_j} X_{(k)}$$

When the standard deviation σ is a nuisance parameter, we calculate a pooled variance of the grouped data:

$$s_j^2 = \frac{1}{N_j - 1} \sum_{k=N_1+\cdots+N_{j-1}+1}^{N_1+\cdots+N_j} (X_{(k)} - \theta_j^{(0)})^2$$

$$(\sigma^2)^{(0)} = \frac{(N_1 - 1)s_1^2 + \cdots + (N_d - 1)s_d^2}{N_1 + \cdots + N_d - d}$$

We select 25 samples from the Multinomial(p) and maximize the likelihood with each RSP with the tolerance 10^{-16} and the maximum iteration 1000. To obtain the maximum of the likelihood function, we choose the maximum of 25 maxima. We used the statistical software R for our simulation study.

Table 3.1: Summary of null distribution of the LRTS of the null hypothesis of single normal versus the alternative of three component normal mixture with known mixing proportions $(1/4, 1/2, 1/4)$

p	Sample Size	Mean	Variance	90%	95%	99%
χ_2^2		2	4	4.6052	5.9915	9.2103
$\frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_2^2$		1.5	3.25	3.8078	5.1384	8.2733
χ_1^2		1	2	2.7055	3.8415	6.6349
$(0.25, 0.5, 0.25)$	100	2.031464	5.145351	5.029431	6.592596	10.32350
$(0.25, 0.5, 0.25)$	1000	1.599577	3.805618	4.019602	5.457903	9.001206
$(0.25, 0.5, 0.25)$	5000	1.535972	3.401620	4.005553	5.423889	8.242983
$(0.25, 0.5, 0.25)$	10000	1.495016	3.149016	3.898502	5.139938	7.992714

Notes: Simulation results based on 5,000 replicates per line.

Table 3.1 contains the mean, variance, and selected percentiles from the simulated null distribution of the LRTS for sample size between 100 and 10,000 for the test

23

discussed in Loisel et al. [20]. We also report the values for selected distributions. The means, variances, and percentiles constantly decrease. The null distribution appears to converge, possibly to the $\frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_2^2$ as stated in Loisel et al. However, at $n = 10,000$, simulated values are less than values for $\frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_2^2$, raising the question that this distribution may not the limit.

3.5   Dimension Loss

When the information matrix is singular, the parameter space may be folded and the dimensionality of the asymptotic results may reduce. The following example shows that the dimension of the LRTS may decrease when the hypothesis space is a cusp.

Example 3.2

Suppose that the log-likelihood can be expanded uniformly in the parameter space Θ as

$$L = (Z_1, Z_2)\sqrt{n}(\theta_1, \theta_2)^T - \frac{1}{2}n(\theta_1, \theta_2)I(\theta_1, \theta_2)^T + o_P(1), \tag{3.9}$$

where I is a 2×2 identity matrix. Consider hypothesis testing in a parameter space $\Theta = \{(\theta_1, \theta_2) : \theta_1^2 \geq \theta_2 \text{ and } \theta_i \geq 0 \text{ for } i = 1, 2\}$

$$H_0 : \theta = (0, 0)$$

$$H_1 : \theta \neq (0, 0)$$

In $\mathbb{R}^2$, the maximum likelihood estimate of $\hat{\theta} = (Z_1/\sqrt{n}, Z_2/\sqrt{n}) + o_P(1)$. When

24

$\hat{\theta}_1^2 < \hat{\theta}_2$, the restricted MLE can be obtained by solving

$$y = x^2$$

$$(x - a, y - b) \cdot (1, 2x) = 0,$$

where $(a, b) = \hat{\theta}$.

$$x = \frac{(108a + \sqrt{108^2 a^2 + 4(6 - 12b)^3})^{1/3}}{6 \cdot 2^{1/3}}$$
$$- \frac{6 - 12b}{3 \cdot 2^{2/3}(108a + \sqrt{108^2 a^2 + 4(6 - 12b)^3})^{1/3}} 6 \cdot 2^{1/3}$$
$$y = x^2$$

Using the Taylor series expansion in a and b,

$$x = \frac{Z_1}{\sqrt{n}} + 2\frac{Z_1 Z_2}{n} + o(1/n)$$
$$y = \frac{Z_1^2}{n} + o(1/n).$$

Therefore, we have

$$L = \frac{1}{2}Z_1^2 + Z_2 \frac{Z_1^2}{\sqrt{n}} - \frac{1}{2}\frac{Z_1^4}{n} + o_P(1).$$

The asymptotic null distribution of LRTS is $\{\max(0, Z_1)\}^2$. $\qquad\square$

The following example shows a nested mixture of a three component mixture from one parameter family densities, which is discussed in the lemma.

Example 3.3

Consider a three component mixture with two parameters $h(x|\theta_1, \theta_2) = (1/3)f(x|\theta_1 + \theta_2) + (1/3)f(x| - \theta_1 + \theta_2) + (1/3)f(x| - 2\theta_2)$. We assume that the true parameter

values are $\theta_0 = (0, 0)$. The first partial derivatives $\partial h/\partial \theta_i$ vanishes at θ_0. The second partial derivatives can be combined as $(f''/f)[2/3\theta_1^2 + 2\theta_2^2]$. The third partial derivatives are $(f'''/f)[(2/3)\theta_1^2\theta_2 - 2\theta_2^3]$. Using a non-linear transformation $(\phi_1, \phi_2) = (2/3\theta_1^2 + 2\theta_2^2, (2/3)\theta_1^2\theta_2 - 2\theta_2^3)$, the log-likelihood in some small ϵ neighborhood of θ_0 may be expanded with a new information matrix defined with f''/f and f'''/f if they are linearly independent. However, the third order terms decrease faster than $1/\sqrt{n}$ and these terms asymptotically vanish. $\hfill \square$

3.6 Detection of a Major Gene

Loisel et al. [20] applied the asymptotic results of mixtures of known proportions to detect a major gene of quantitative trait loci (QTL). They used a mixture model of normal distributions with known proportion $p = (1/4, 1/2, 1/4)$.

$$h(x|\theta_1, \theta_2, \theta_3, \sigma) = \frac{1}{4}f(x|\theta_1, \sigma) + \frac{1}{2}f(x|\theta_2, \sigma) + \frac{1}{4}f(x|\theta_3, \sigma)$$

They applied normal mixtures to analyze the square root of the number of grains in an F2 population of 150 plants. By Theorems in Chapter 3, we may apply mixtures of other exponential families. For illustration, a mixture of a gamma, Gamma$(\theta, 1)$, distribution is applied to the data to test hypotheses

$$H_0 : \theta_1 = \theta_2 = \theta_3$$

$$H_2 : \theta_2 = \theta_3 \text{ Dominance}$$

$$H_3 : (\theta_1, \theta_2, \theta_3) \text{ Saturated model}$$

The $LRTS = 163.65 > 1.92$ for H_0 against H_3 and the $LRTS < 10^{-12} < 1.92$ for H_2 against H_3, which suggest the dominance model, i.e., two component mixture

of gamma distributions. The estimated parameters are $\hat{\theta}_1 = 3.29, \hat{\theta}_2 = 11.08$ and $\hat{\theta}_3 = 11.08$. The critical value 1.92 is the 95th percentile of the 50:50 mixture of a point mass at zero and a chi-square of degrees of freedom one.

Figure 3.1: Histogram of data and fitted mixture
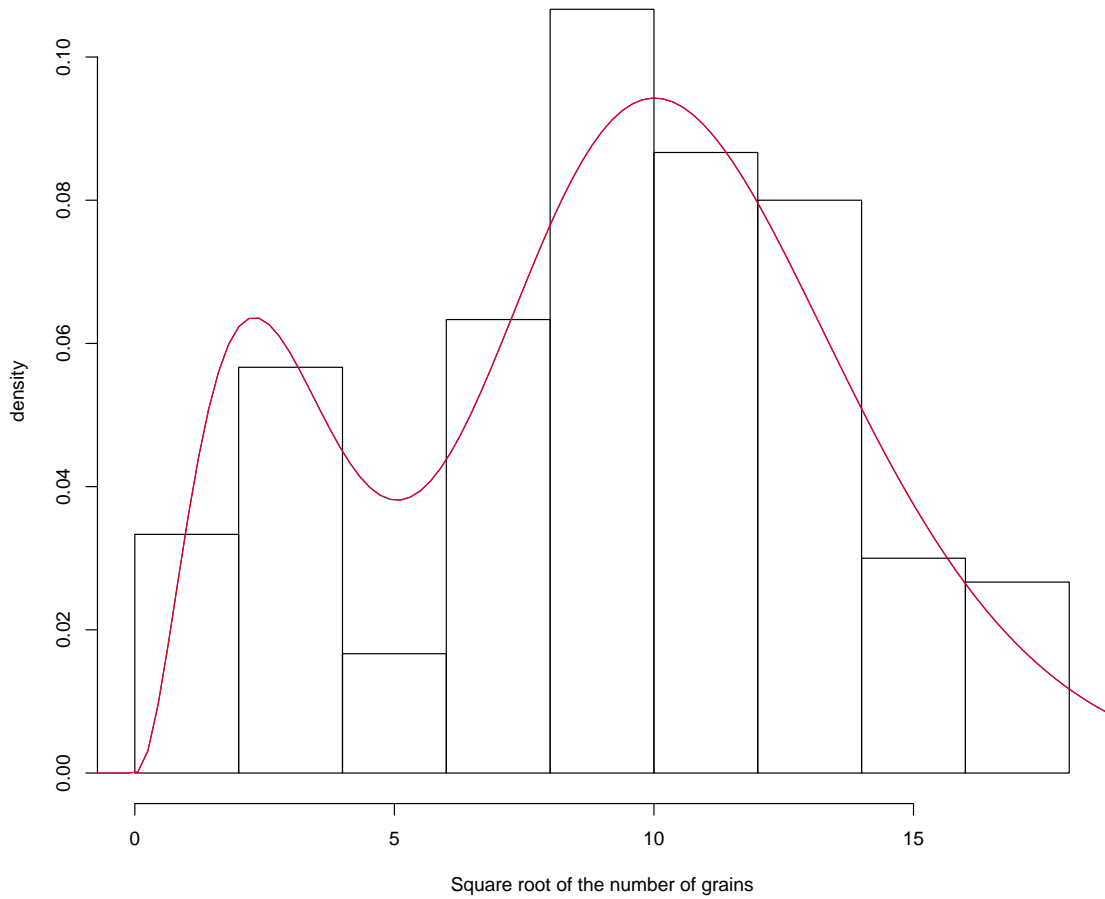
Source: Data are 150 values from Loisel et al. [20]

Chapter 4

A Mixture Model in Genetic Association Studies

In this chapter, we discuss the likelihood ratio test of a mixture model in a genetic association study with genotyping errors. Suppose that genotypes of a single nucleotide polymorphism (SNP) marker are determined by a continuous measurement such as the fluorescence intensity signal. When the signals are well separated among the different genotypes, a 2x3 test can be easily applied to test association between the marker and trait of interest. The distribution of intensities of a large number of samples is a mixture of a normal or other distributions. If the separation of the intensities are small, then it is not easy to call the genotypes of the sample. We propose the likelihood ratio test (LRT) without genotype calls when the separations are not large and compare the performance of LRT to 2×3 tests with two different types of genotype calls: (1) the half-way rule and (2) a Bayesian rule.

If the genotyping errors are non-differential across cases and controls, then the Type I error of the chi-square test is not affected, but the power of the test decreases. Two samples are randomly selected from two populations, cases and controls, independently. Although we assume that the dependent variable is univariate, this is not an important limitation. For example, a two dimensional intensity scatter plot can be projected using the angle from an axis or the ratio of intensities; so that a two

dimensional data problem is reduced to one dimension. All parameters θ_i and σ in a component probability density function are nuisance parameters and the mixing proportions are to be tested whether or not they are equal, that is, $H_0 : p_a = p_u$ against $H_1 : p_a \neq p_u$, where $p_b = (p_{b1}, \cdots, p_{bd})$, with $b = a$ or u. That is,

$$h_a(x|\theta, \sigma, p_a) = \sum_{i=1}^{d} p_{ai} f(x|\theta_i, \sigma)$$

$$h_u(x|\theta, \sigma, p_u) = \sum_{i=1}^{d} p_{ui} f(x|\theta_i, \sigma),$$

where $f(x|\theta_i, \sigma)$ is a probability density of a normal distribution with mean θ_i and standard deviation σ. If two dimensional data is analyzed without projection, then $\theta_i \in \mathbb{R}^2$ and $\sigma = \Sigma$ is a 2×2 covariance matrix. Since LRTS for the model satisfies classical regularity conditions, the asymptotic null distribution of LRTS follows a χ^2 with $(d-1)$ degrees of freedoms. The asymptotic power can be obtained by calculating the non-centrality parameter since the alternative distribution of the LRTS asymptotically follows a non-central chi-square distribution.

4.1   Calculation of the non-centrality parameter of LRTS

In this section, we derive the non-centrality parameter of the LRTS to test whether or not the mixing proportions of two populations are equal. We assume the number of components is fixed and known for both populations. Let $f(x|\theta_i, \eta)$ denote the probability density function of the ith component of a mixture distribution, where η is a vector of nuisance parameters. Suppose that the ith population

follows a mixture distribution with common components

$$h_i(x|p_i, \theta, \eta) = \sum_{j=1}^{d} p_{ij} f(x|\theta_j, \eta) \tag{4.1}$$

Theorem 4.1. *Assume two independent random samples are drawn from $h_i(x|p, \theta, \eta)$ for $i = 1, 2$. Let $p_i = (p_{i1}, \cdots, p_{1d})$ and $\theta = (\theta_1, \cdots, \theta_d)$. Let $H_0 : p_{1j} = p_{2j}$ for all $j = 1, \cdots, d$. Assume that the classical regularity conditions hold for the consistency of the maximum likelihood estimators and quadratic approximations of the likelihood function. Then, (1) the LRTS under H_0 follows a χ^2_{d-1} distribution and (2) the asymptotic power along a contiguous family of hypotheses $H_N : p_{ij} = p_{0j} + c_{ij}/\sqrt{N}$ can be calculated by a non-central chi-square distribution $\chi^2_{d-1,\lambda}$ with the non-centrality parameter:*

$$\lambda_{LRT} = N Q_1 Q_2 \left(p_{11} - p_{21}, \cdots, p_{1(d-1)} - p_{2(d-1)} \right) J_0 \begin{pmatrix} p_{11} - p_{21} \\ \cdots \\ p_{1(d-1)} - p_{2(d-1)}, \end{pmatrix} \tag{4.2}$$

where J_0 is a $(d-1) \times (d-1)$ symmetric matrix

$$J_0 = E_0 \begin{pmatrix} \left(\frac{f(x|\theta_1,\eta) - f(x|\theta_d,\eta)}{\sum p_{0i} f(x|\theta_i,\eta)} \right)^2 & \cdots & \frac{\prod_{k=1,d-1} \left(f(x|\theta_k,\eta) - f(x|\theta_d,\eta) \right)}{(\sum p_{0i} f(x|\theta_i,\eta))^2} \\ \vdots & \cdots & \vdots \\ \frac{\prod_{k=1,d-1} \left(f(x|\theta_k,\eta) - f(x|\theta_d,\eta) \right)}{(\sum p_{0i} f(x|\theta_i,\eta))^2} & \cdots & \left(\frac{f(x|\theta_{d-1},\eta) - f(x|\theta_d,\eta)}{\sum p_{0i} f(x|\theta_i,\eta)} \right)^2 \end{pmatrix}.$$

Proof. The first part of Theorem is a direct conclusion from the classical asymptotic result of the likelihood ratio test. To prove the second part of Theorem, let $p_0 = (p_{01}, \cdots, p_{0d})$ denote the true parameter under H_0 and $h_0 = h(x|p_0, \theta, \eta)$. Since $\sum p_{ij} = 1$, we use $p_{id} = 1 - \sum_{j=1}^{d-1} p_{ij}$. The information matrix $I_0 = I(p_0)$ can be

written as

$$I(p_0) = \begin{pmatrix} K(p_0) & A(p_0) \\ A^T(p_0) & B(p_0) \end{pmatrix} \quad \text{where } B(p_0) \text{ is symmetric.}$$

$$K(p_0) = \begin{pmatrix} J(p_0) & 0 \\ 0 & J(p_0) \end{pmatrix} \quad \text{and}$$

$$J_0 = J(p_0) = E_0 \begin{pmatrix} \left(\frac{f(x|\theta_1,\eta)-f(x|\theta_d,\eta)}{\sum p_{0i} f(x|\theta_i,\eta)}\right)^2 & \cdots & \frac{\prod_{k=1,d-1}\left(f(x|\theta_k,\eta)-f(x|\theta_d,\eta)\right)}{(\sum p_{0i} f(x|\theta_i,\eta))^2} \\ \vdots & \cdots & \vdots \\ \frac{\prod_{k=1,d-1}\left(f(x|\theta_k,\eta)-f(x|\theta_d,\eta)\right)}{(\sum p_{0i} f(x|\theta_i,\eta))^2} & \cdots & \left(\frac{f(x|\theta_{d-1},\eta)-f(x|\theta_d,\eta)}{\sum p_{0i} f(x|\theta_i,\eta)}\right)^2 \end{pmatrix}.$$

Let J_{ij} be the ij element of the matrix J_0. The total sample size $N = N_1 + N_2$ and the fraction of ith sample $Q_i = N_i/N$. Consider a contiguous family of hypotheses H_N : $p_{ij} = p_{0j} + \frac{c_{ij}}{\sqrt{N}}$. By decomposing a vector $v = (c_{11}\sqrt{Q_1}, \cdots, c_{1(d-1)}\sqrt{Q_1}, c_{21}\sqrt{Q_2}, \cdots, c_{2(d-1)}\sqrt{Q_2})^T$ into a sum of two orthogonal vectors with respect to the inner product generated by $K_0 = K(p_0)$,

$$\begin{pmatrix} c_{11}\sqrt{Q_1} \\ \vdots \\ c_{1(d-1)}\sqrt{Q_1} \\ c_{21}\sqrt{Q_2} \\ \vdots \\ c_{1(d-1)}\sqrt{Q_2} \end{pmatrix} = \begin{pmatrix} \frac{c_{11}Q_1+c_{21}Q_2}{Q_1+Q_2}\sqrt{Q_1} \\ \vdots \\ \frac{c_{1(d-1)}Q_1+c_{2(d-1)}Q_2}{Q_1+Q_2}\sqrt{Q_1} \\ \frac{c_{11}Q_1+c_{21}Q_2}{Q_1+Q_2}\sqrt{Q_2} \\ \vdots \\ \frac{c_{1(d-1)}Q_1+c_{2(d-1)}Q_2}{Q_1+Q_2}\sqrt{Q_2} \end{pmatrix} + \begin{pmatrix} \frac{(c_{11}-c_{21})\sqrt{Q_1 Q_2}}{Q_1+Q_2}\sqrt{Q_2} \\ \vdots \\ \frac{(c_{1(d-1)}-c_{2(d-1)})\sqrt{Q_1 Q_2}}{Q_1+Q_2}\sqrt{Q_2} \\ -\frac{(c_{11}-c_{21})\sqrt{Q_1 Q_2}}{Q_1+Q_2}\sqrt{Q_1} \\ \vdots \\ -\frac{(c_{1(d-1)}-c_{2(d-1)})\sqrt{Q_1 Q_2}}{Q_1+Q_2}\sqrt{Q_1} \end{pmatrix} \quad (4.3)$$

$$= v^{\parallel} + v^{\perp}.$$

Since $(v^\parallel)^T K_0 v^\perp = 0$ and $v = v^\parallel + v^\perp$, the non-centrality parameter can be obtained in a quadratic form $(v^\perp)^T K_0 v^\perp$.

$$\lambda_{LRT} = Q_1 Q_2 \sum_{i=1}^{d-1} \sum_{j=1}^{d-1} J_{ij}(c_{1i} - c_{2i})(c_{1j} - c_{2j})$$

Using the relationship $\sqrt{N}(p_{ij} - p_{0j}) = c_{ij}$,

$$\lambda_{LRT} = N Q_1 Q_2 \sum_{i=1}^{d-1} \sum_{j=1}^{d-1} J_{ij}(p_{1i} - p_{2i})(p_{1j} - p_{2j}) \tag{4.4}$$

This completes the proof. $\qquad\square$

Example 4.1

When $d = 3$ and $N_1 = N_2$, the non-centrality parameter of the power of the LRT may be written

$$\lambda_{LRT} = \frac{N}{4}\left(J_{11}(p_{11} - p_{21})^2 + 2J_{12}(p_{11} - p_{21})(p_{12} - p_{22}) + J_{22}(p_{12} - p_{22})^2\right). \tag{4.5}$$

$\square$

4.2 Classification Rules, Error Matrices, and Power for 2×3 Test

Assume that observations follow $h(x|p_i, \theta, \eta) = \sum_{j=1}^{d} p_i f(x|\theta_j, \eta)$ for $i = 1, 2$. We define two classification rules for clustering data: (1) the first classification method is to use a half-way rule proposed by Kang et al. [13]. Once a centroid $g(\theta_i, \eta)$ of each component is chosen, then determine the component of an observation X by a rule that X belongs to the component i if $D(X, g(\theta_i, \eta)) \leq \min_{1 \leq j \leq d}\{D(X, g(\theta_j, \eta))\}$, where $D(x, y)$ is the distance between x and y.

Example 4.2

Suppose that $d = 3$, f is a univariate normal probability density function, and $D(x, y) = |x - y|^k$ is the L^k distance in Euclidean space. Define the centroids $g(\theta_i, \eta) = \theta_i$, and two boundary points $\gamma_1^H = \frac{\theta_1 + \theta_2}{2}$, and $\gamma_2^H = \frac{\theta_2 + \theta_3}{2}$. The half-way rule can be written

$$X \leq \gamma_1^H \Leftrightarrow \text{Genotype } AA$$

$$\gamma_2^H < X \leq \gamma_2^H \Leftrightarrow \text{Genotype } AB$$

$$\gamma_2^H \leq X \Leftrightarrow \text{Genotype } BB.$$

(2) Another classification method is to use the Bayes rule [23]. An observation X is assigned to the ith component if and only if $p_{0i} f(X|\theta_i, \eta) \geq \max_{1 \leq j \leq d}(p_{0j} f(X|\theta_j, \eta))$, where $p_{0i} = Q_1 p_{1i} + Q_2 p_{2i}$ is the pooled mixing proportion to satisfy non-differential error mechanism.

Example 4.3

When $d = 3$ and f is a univariate normal probability density function, the boundaries of clusters can be obtained by straightforward calculations

$$\gamma_1^B = \min\left(\frac{\theta_2 + \theta_1}{2} - \frac{\sigma^2}{\theta_2 - \theta_1} \log \frac{p_{02}}{p_{01}}, \frac{\theta_3 + \theta_1}{2} - \frac{\sigma^2}{\theta_3 - \theta_1} \log \frac{p_{03}}{p_{01}}\right) \qquad (4.6)$$

$$\gamma_2^B = \max\left(\frac{\theta_3 + \theta_1}{2} + \frac{\sigma^2}{\theta_3 - \theta_1} \log \frac{p_{01}}{p_{03}}, \frac{\theta_3 + \theta_2}{2} + \frac{\sigma^2}{\theta_3 - \theta_2} \log \frac{p_{02}}{p_{03}}\right). \qquad (4.7)$$

□

We now formulate the error matrix $\mathcal{E} = (\epsilon_{ij} = \Pr(\text{ assigned to } i|\text{true component } j))$ in terms of the cumulative distribution function (CDF) F of a univariate random

Figure 4.1: Probability density function of three component normal mixture. Bayesian classification division points γ_i^B. Half-way rule division points γ_i^H.
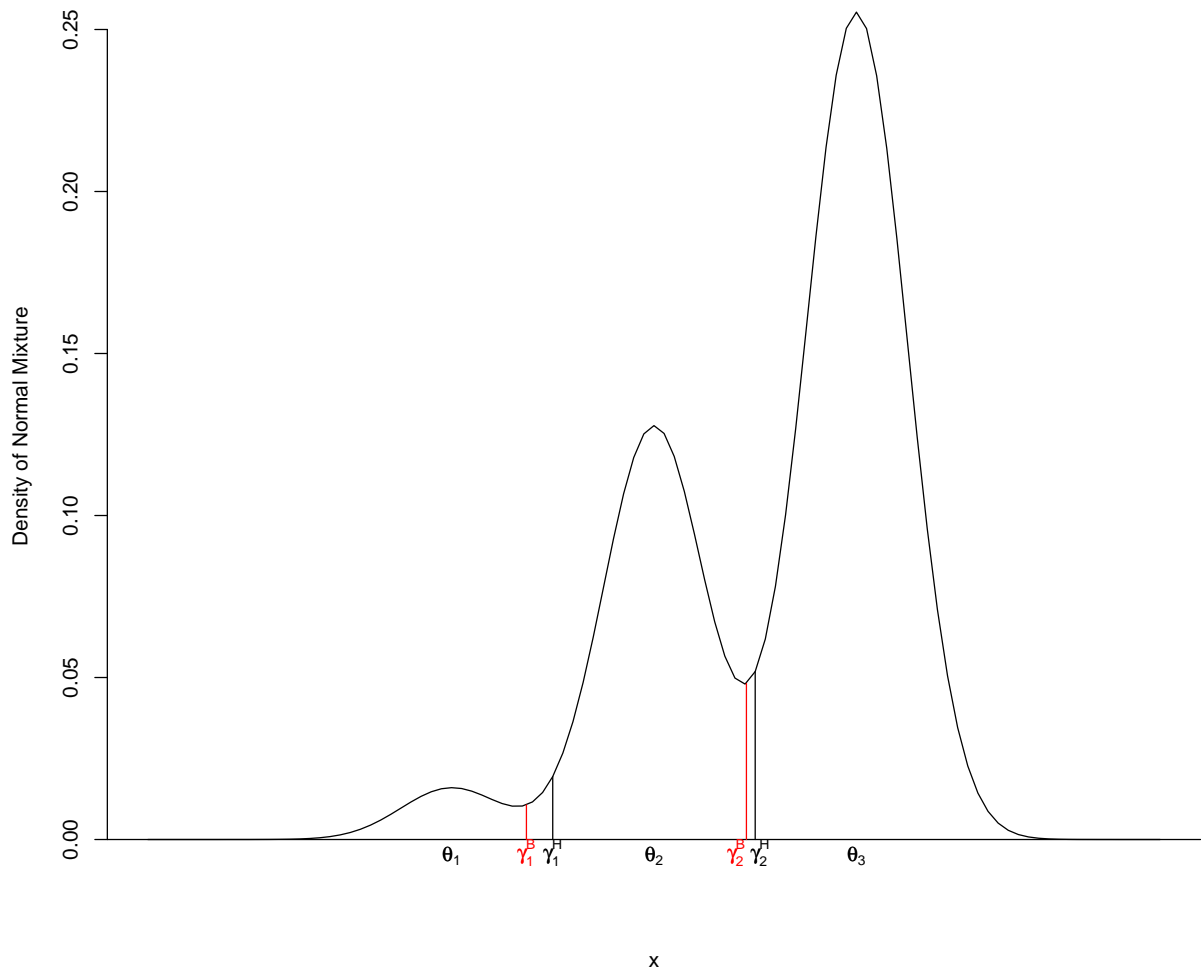
Notes: Means are $\theta_1 = -4$, $\theta_2 = 0$, and $\theta_3 = 4$ with common standard deviation $\sigma = 1$. Mixing proportions $(p_1, p_2, p_3) = (0.04, 0.32, 0.64)$.

variable X. Assuming $\theta_1 < \cdots < \theta_d$ and $-\infty = \gamma_0 < \gamma_1 < \cdots < \gamma_{d-1} < \gamma_d = \infty$,

$\epsilon_{ij} = F(\gamma_i|\theta_j, \eta) - F(\gamma_{i-1}|\theta_j, \eta)$ for $i = 1, \cdots, d$.

Example 4.4

Suppose that $f(x|\theta_i, \eta)$ is a normal density with mean θ_i and variance σ^2. In case $d = 3$, the error matrix has the form

$$\mathcal{E} = \begin{pmatrix} \Phi(\frac{\gamma_1-\theta_1}{\sigma}) & \Phi(\frac{\gamma_1-\theta_2}{\sigma}) & \Phi(\frac{\gamma_1-\theta_3}{\sigma}) \\ \Phi(\frac{\gamma_2-\theta_1}{\sigma}) - \Phi(\frac{\gamma_1-\theta_1}{\sigma}) & \Phi(\frac{\gamma_2-\theta_2}{\sigma}) - \Phi(\frac{\gamma_1-\theta_2}{\sigma}) & \Phi(\frac{\gamma_2-\theta_3}{\sigma}) - \Phi(\frac{\gamma_1-\theta_3}{\sigma}) \\ 1 - \Phi(\frac{\gamma_2-\theta_1}{\sigma}) & 1 - \Phi(\frac{\gamma_2-\theta_2}{\sigma}) & 1 - \Phi(\frac{\gamma_2-\theta_3}{\sigma}) \end{pmatrix},$$

where Φ is a CDF of a standard normal distribution. The misclassified proportions are computed from $p_i^{*Y} = \mathcal{E}^Y p_i$ for $Y = H$ or B, i.e.,

$$\begin{pmatrix} p_{i1}^* \\ p_{i2}^* \\ p_{i3}^* \end{pmatrix} = \begin{pmatrix} \Phi(\frac{\gamma_1-\theta_1}{\sigma}) & \Phi(\frac{\gamma_1-\theta_2}{\sigma}) & \Phi(\frac{\gamma_1-\theta_3}{\sigma}) \\ \Phi(\frac{\gamma_2-\theta_1}{\sigma}) - \Phi(\frac{\gamma_1-\theta_1}{\sigma}) & \Phi(\frac{\gamma_2-\theta_2}{\sigma}) - \Phi(\frac{\gamma_1-\theta_2}{\sigma}) & \Phi(\frac{\gamma_2-\theta_3}{\sigma}) - \Phi(\frac{\gamma_1-\theta_3}{\sigma}) \\ 1 - \Phi(\frac{\gamma_2-\theta_1}{\sigma}) & 1 - \Phi(\frac{\gamma_2-\theta_2}{\sigma}) & 1 - \Phi(\frac{\gamma_2-\theta_3}{\sigma}) \end{pmatrix} \begin{pmatrix} p_{i1} \\ p_{i2} \\ p_{i3} \end{pmatrix}.$$

$\square$

The non-centrality parameter λ_{CS}^Y [10, 24] with classification rule $Y = H$ or B can be written

$$\lambda_{CS}^Y = N Q_1 Q_2 \sum_{i=1}^{d} \frac{(p_{1i}^{*Y} - p_{2i}^{*Y})^2}{p_{0i}^{*Y}}. \tag{4.8}$$

4.3 Comparison of the LRT and the chi-square tests

We define the relative efficiency Eff^Y of the chi-square test to the LRT and the extra sample size N^Y that is needed so that the chi-square test has the same power as the LRT.

Definition 4.1. *The ratio of the non-centrality parameter of the chi-square test to the non-centrality parameter of the LRT is called the relative efficiency.*

$$\text{Eff}^Y = \frac{\lambda^Y_{CS}}{\lambda_{LRT}}.$$

By equating the non-centrality parameter of a chi-square test with sample size $N + N^*$ to the non-centrality parameter of the LRT with sample size N, we define the extra sample size N^*.

Definition 4.2. *The extra sample size N^{*Y} can be written as*

$$N^{*Y} = \left(\frac{1}{\text{Eff}^Y} - 1\right)N.$$

Example 4.5

Given mixing proportions $p_i \in \mathbb{R}^d$, $p_0 = Q_1 p_1 + Q_2 p_2$. The non-centrality parameters (4.8) and (4.4) can be used to derive the relative efficiency

$$
\begin{aligned}
\text{Eff}^Y &= \frac{\lambda^Y_{CS}}{\lambda_{LRT}} \\
&= \frac{\sum_{i=1}^d \frac{(p^{*Y}_{1i} - p^{*Y}_{2i})^2}{Q_1 p^{*Y}_{1i} + Q_2 p^{*Y}_{2i}}}{\sum_{i=1}^{d-1} \sum_{j=1}^{d-1} J_{ij}(p_{1i} - p_{2i})(p_{1j} - p_{2j})},
\end{aligned}
$$

where
$$J_{ij} = E_0\left[\frac{\big(f(x|\theta_i, \eta) - f(x|\theta_d, \eta)\big)\big(f(x|\theta_j, \eta) - f(x|\theta_d, \eta)\big)}{(\sum_{i=1}^d (Q_1 p_{1i} + Q_2 p_{2i})f(x|\theta_i, \eta))^2}\right].$$

$\square$

We illustrate tables and figures of the relative efficiencies and the extra sample sizes when the number of components of a univariate mixture distribution d is three and the components are normally distributed. In order to make parameter config-

urations simpler, we assume Hardy-Weinberg equilibrium (HWE) in both cases and controls. That is, $p_{11} = p_u^2$, $p_{12} = 2p_u(1 - p_u)$, $p_{13} = (1 - p_u)^2$ in control group and $p_{21} = p_a^2$, $p_{22} = 2p_a(1 - p_a)$, $p_{23} = (1 - p_a)^2$ in case group. We set $\theta_3 = -\theta_1$ and $\theta_2 = 0$ and call $S = (\theta_3/\sigma)$ the separation of the mixture distribution. The smaller the separation S is, the smaller the relative efficiency Eff^Y is. The relative efficiency becomes smaller as the minor allele frequency p_u decreases.

Table 4.1: Efficiency of 2×3 independence test relative to LRTS for equal proportions in cases and controls

| p_u	S	Eff^H	Eff^B	N^{*H}	N^{*B}
0.1	1	0.8357	0.6310	39.31	116.93
	2	0.8527	0.8434	34.54	37.13
	3	0.9046	0.9375	21.08	13.31
	4	0.9621	0.9793	7.88	4.21
	5	0.9891	0.9944	2.20	1.11
	6	0.9978	0.9989	0.45	0.21
0.2	1	0.8430	0.7384	37.26	70.83
	2	0.8808	0.8658	27.08	30.98
	3	0.9344	0.9466	14.03	11.26
	4	0.9753	0.9825	5.06	3.55
	5	0.9931	0.9953	1.40	0.93
	6	0.9985	0.9990	0.30	0.19
0.3	1	0.8443	0.7538	36.86	65.31
	2	0.8896	0.8747	24.79	28.64
	3	0.9438	0.9508	11.89	10.34
	4	0.9796	0.9839	4.16	3.25
	5	0.9943	0.9957	1.13	0.85
	6	0.9987	0.9990	0.24	0.18
0.4	1	0.8444	0.7505	36.85	66.45
	2	0.8923	0.8781	24.12	27.75
	3	0.9469	0.9525	11.19	9.96
	4	0.9811	0.9845	3.85	3.12
	5	0.9948	0.9959	1.04	0.82
	6	0.9988	0.9991	0.22	0.17

Notes: $\Delta p = p_a - p_u = 0.1$. N^* is calculated when $N_A = N_U = 200$.

Figure 4.2: Efficiency of 2×3 independence test with respect to LRTS for mixtures by separation of components. Two classification rules: Bayesian, Halfway.
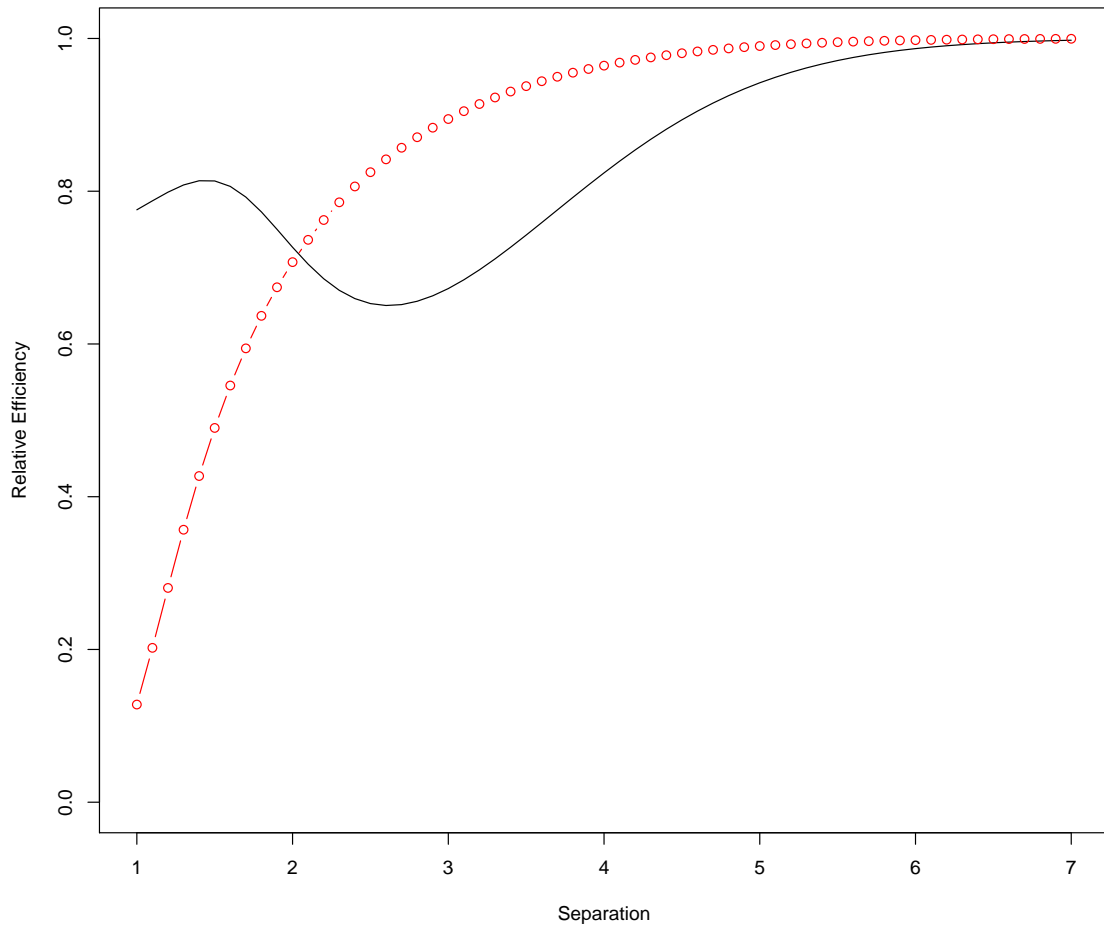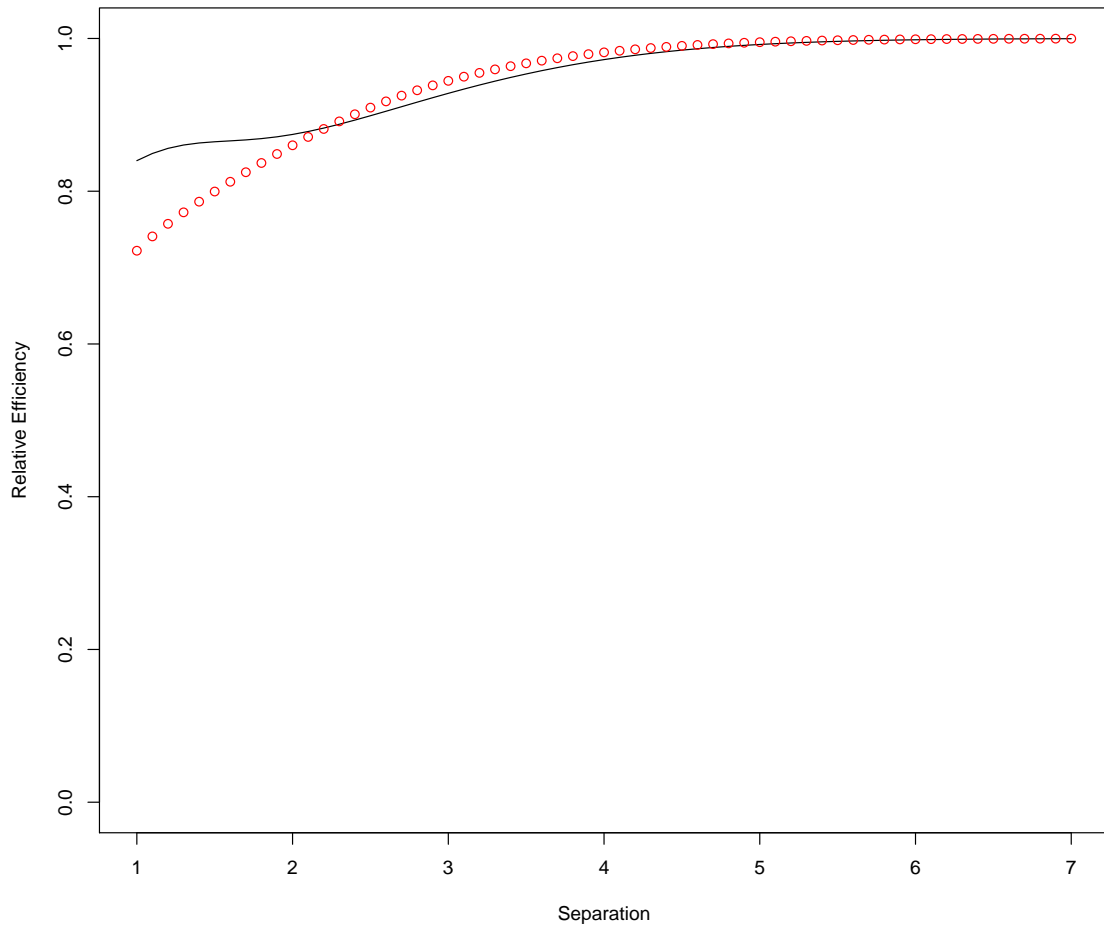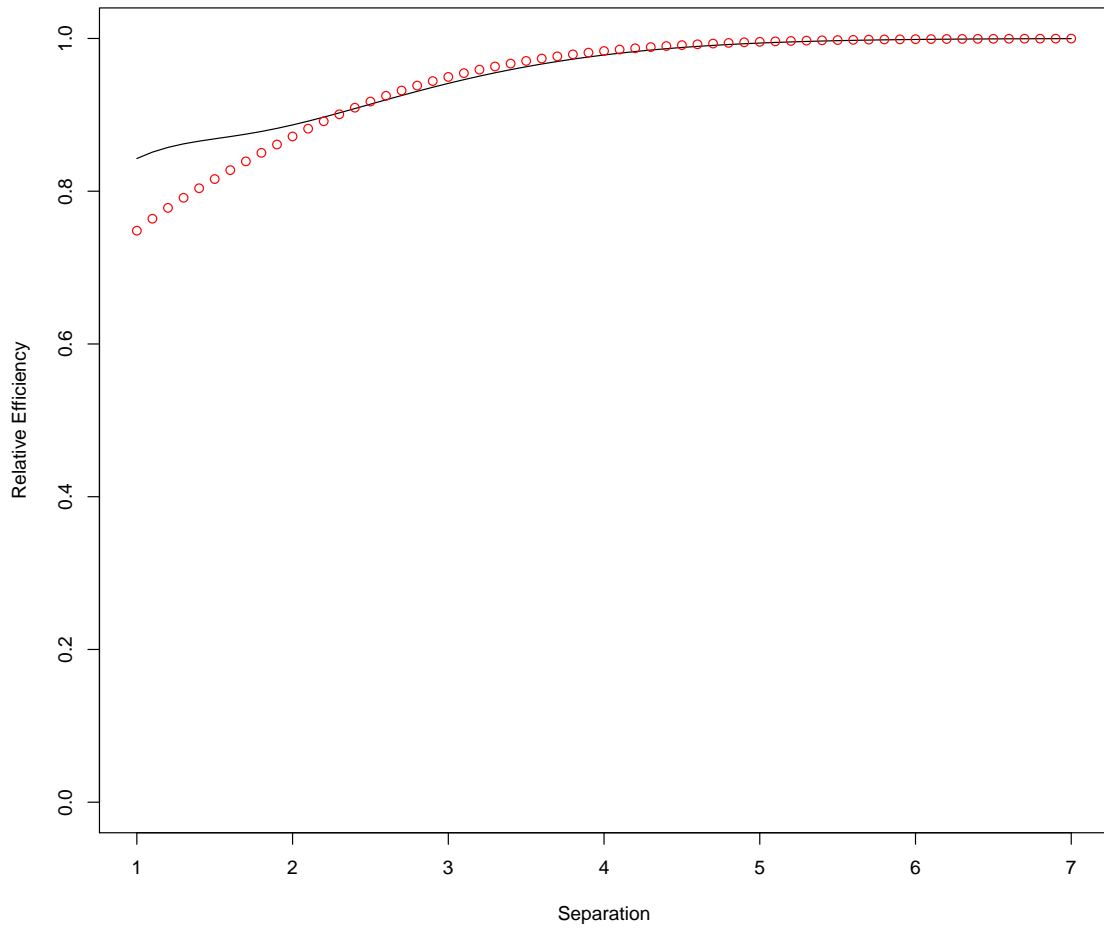

Notes: Minor control allele frequency 0.01 under HWE. Minor case allele frequency 0.06 under HWE. Dotted curve for Bayesian classification. Solid curve for half-way rule.

Figure 4.3: Efficiency of 2×3 independence test with respect to LRTS for mixtures by separation of components. Two classification rules: Bayesian, Halfway.
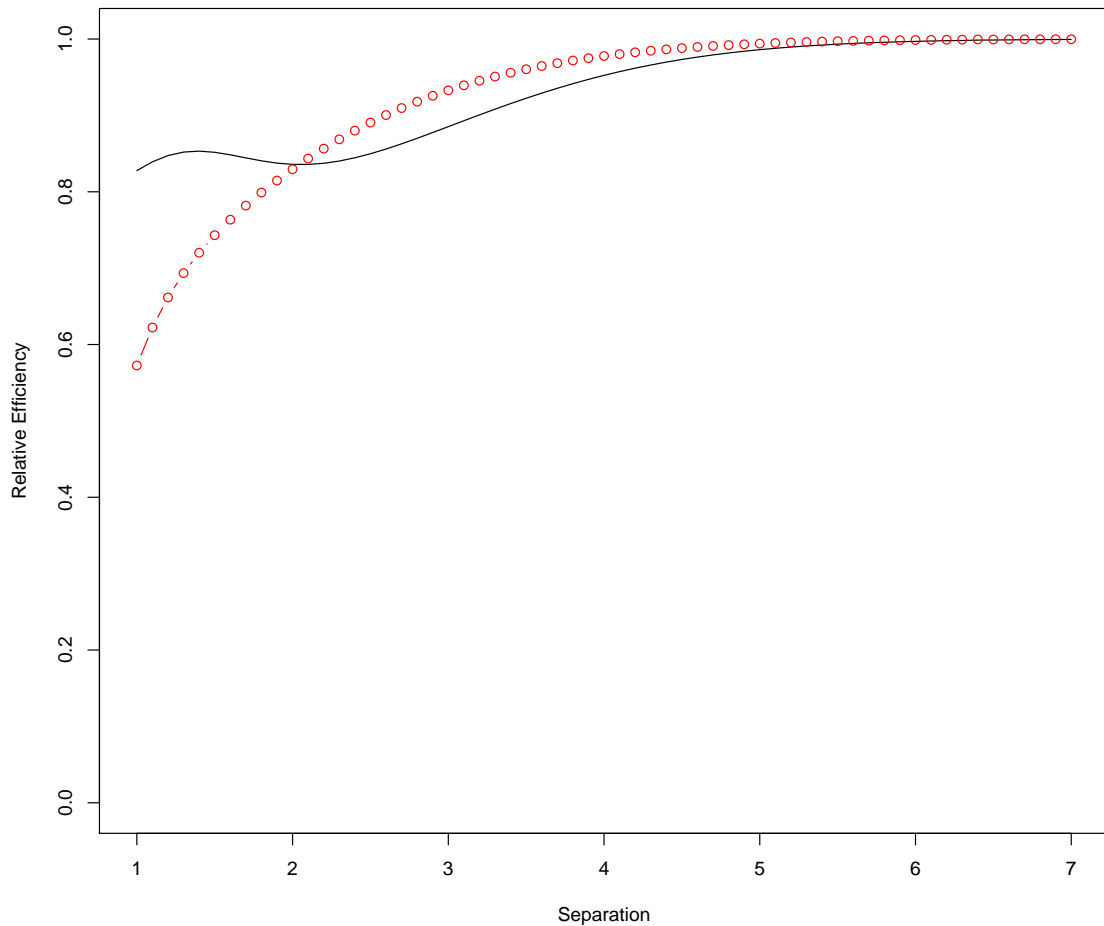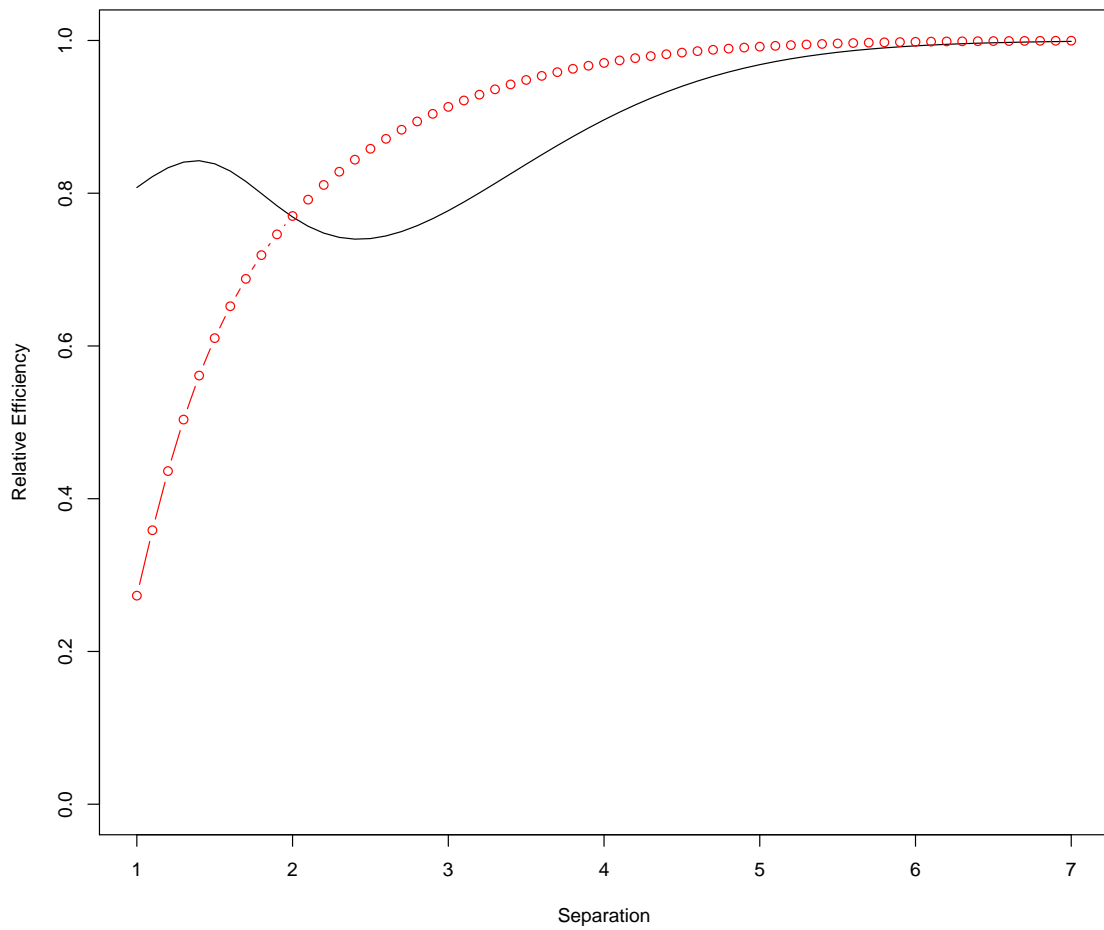

Notes: Minor control allele frequency 0.1 under HWE. Minor case allele frequency 0.15 under HWE. Dotted curve for Bayesian classification. Solid curve for half-way rule.

Figure 4.4: Efficiency of 2×3 independence test with respect to LRTS for mixtures by separation of components. Two classification rules: Bayesian, Halfway.
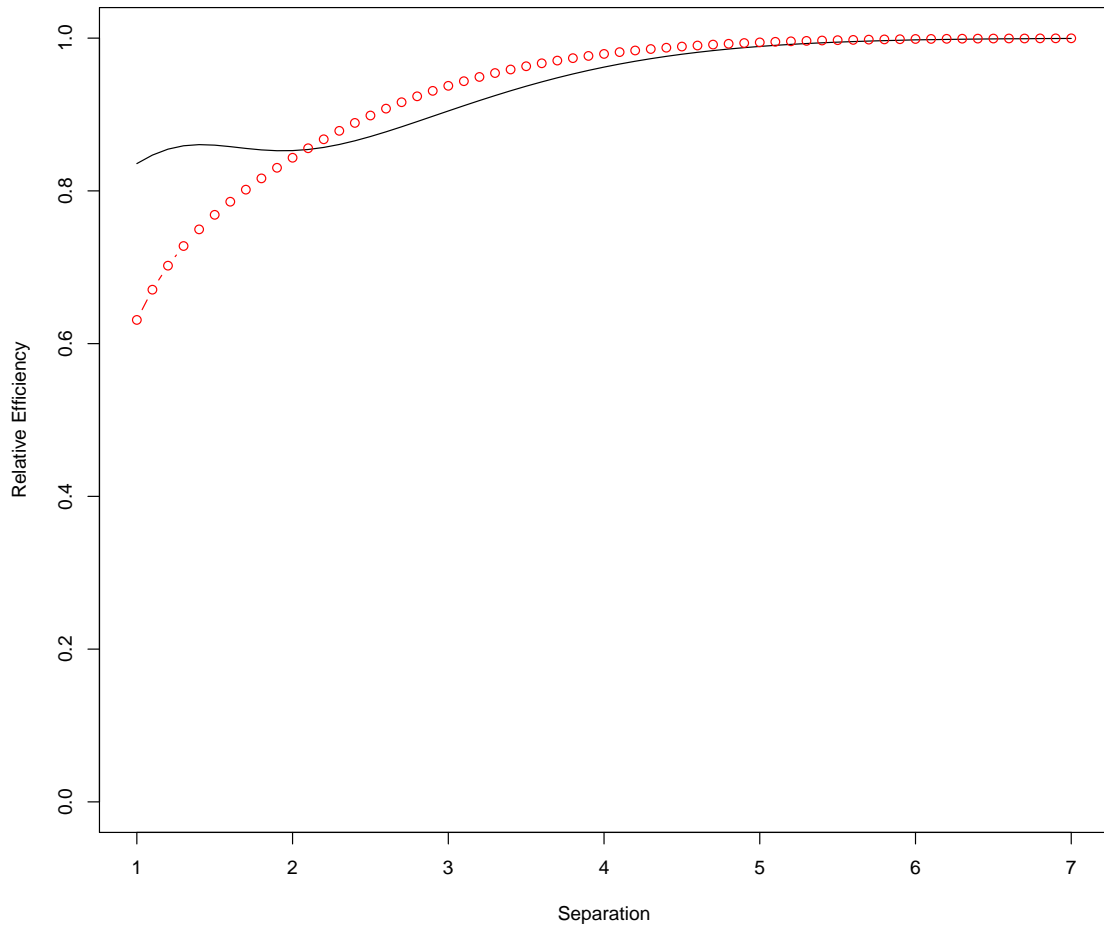

Notes: Minor control allele frequency 0.2 under HWE. Minor case allele frequency 0.25 under HWE. Dotted curve for Bayesian classification. Solid curve for half-way rule.

Figure 4.5: Efficiency of 2×3 independence test with respect to LRTS for mixtures by separation of components. Two classification rules: Bayesian, Halfway.

Notes: Minor control allele frequency 0.3 under HWE. Minor case allele frequency 0.35 under HWE. Dotted curve for Bayesian classification. Solid curve for half-way rule.

Figure 4.6: Efficiency of 2×3 independence test with respect to LRTS for mixtures by separation of components. Two classification rules: Bayesian, Halfway.
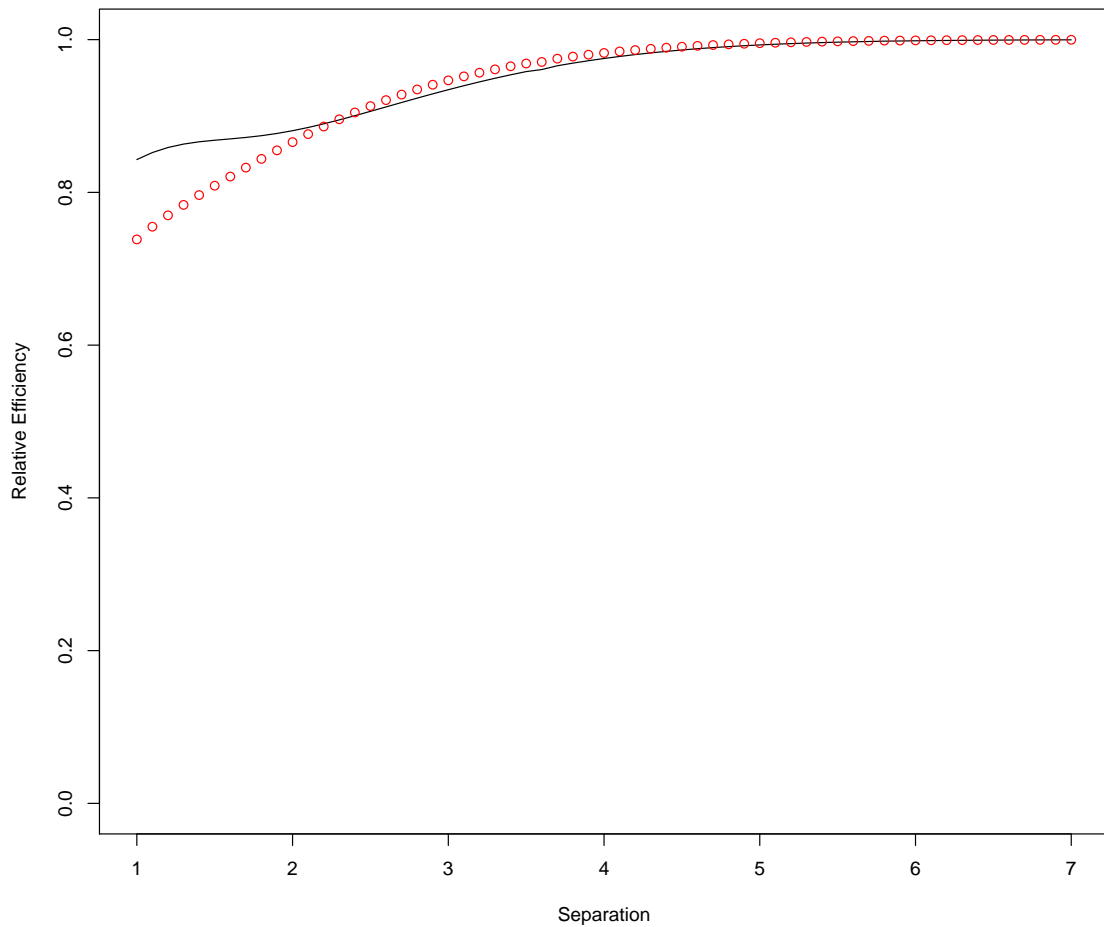

Notes: Minor control allele frequency 0.01 under HWE. Minor case allele frequency 0.11 under HWE. Dotted curve for Bayesian classification. Solid curve for half-way rule.

Figure 4.7: Efficiency of 2×3 independence test with respect to LRTS for mixtures by separation of components. Two classification rules: Bayesian, Halfway.
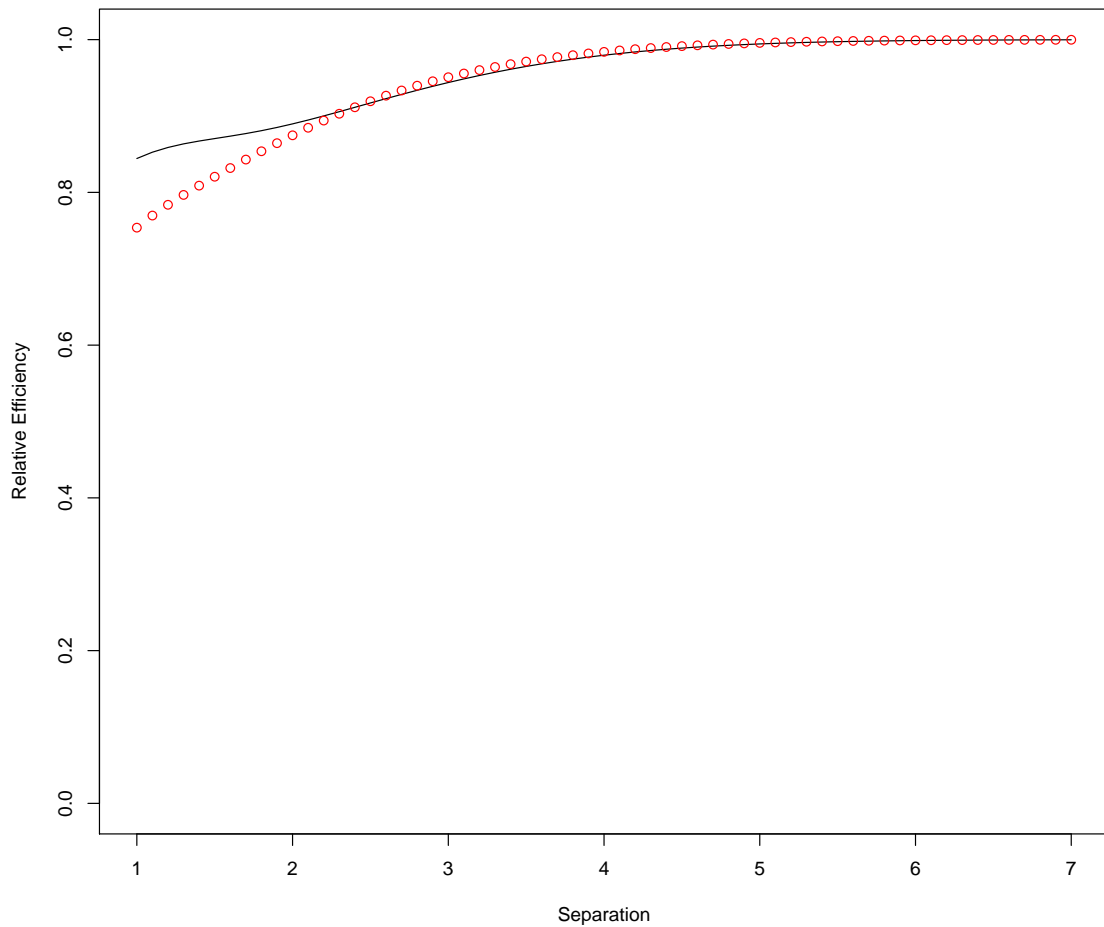

Notes: Minor control allele frequency 0.1 under HWE. Minor case allele frequency 0.2 under HWE. Dotted curve for Bayesian classification. Solid curve for half-way rule.

Figure 4.8: Efficiency of 2×3 independence test with respect to LRTS for mixtures by separation of components. Two classification rules: Bayesian, Halfway.
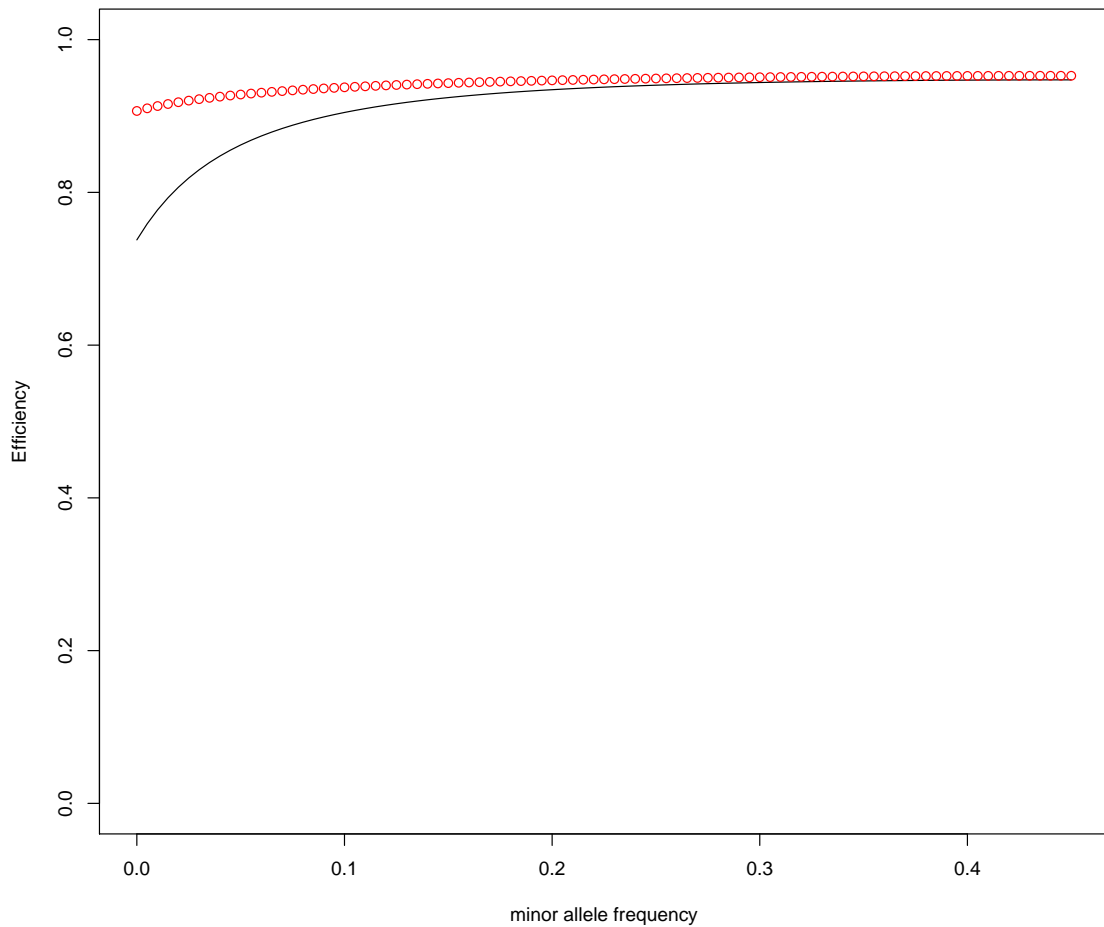

Notes: Minor control allele frequency 0.2 under HWE. Minor case allele frequency 0.3 under HWE. Dotted curve for Bayesian classification. Solid curve for half-way rule.

Figure 4.9: Efficiency of 2×3 independence test with respect to LRTS for mixtures by separation of components. Two classification rules: Bayesian, Halfway.
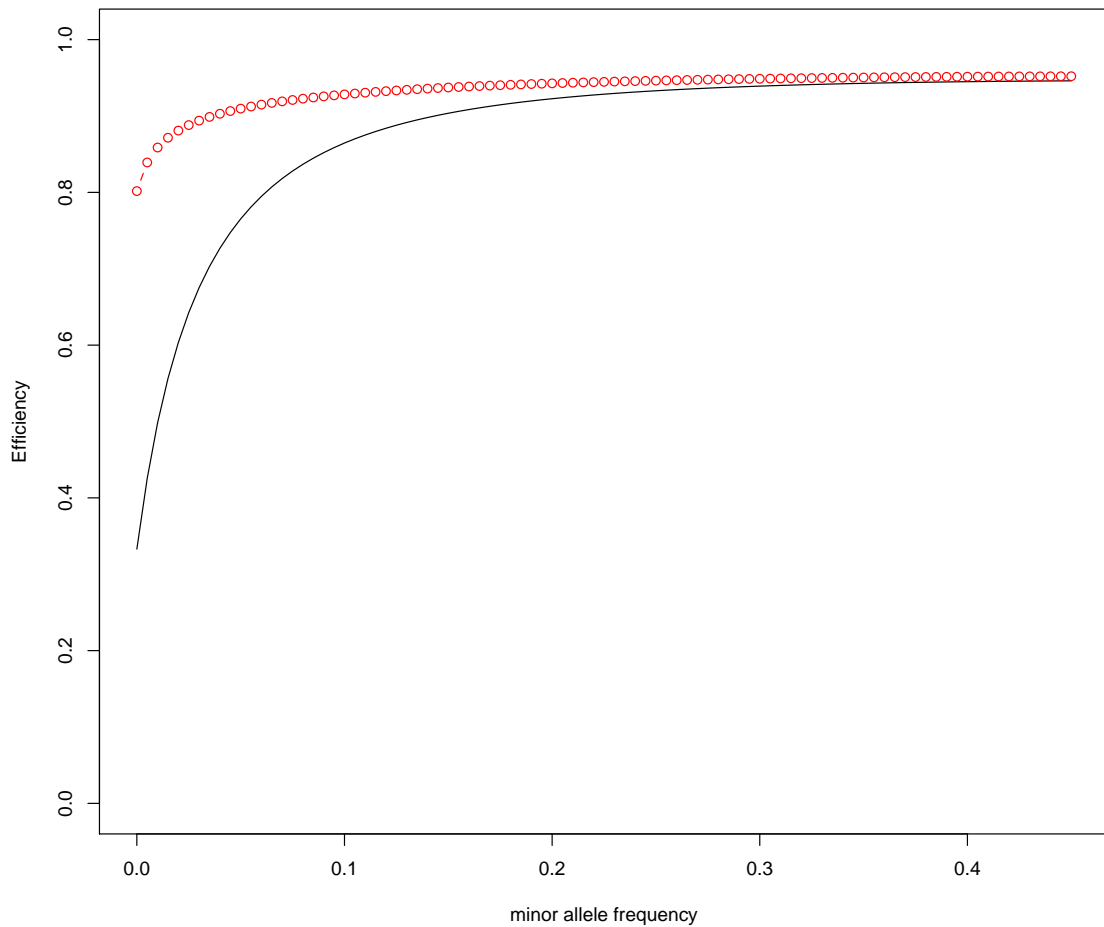

Notes: Minor control allele frequency 0.4 under HWE. Minor case allele frequency 0.5 under HWE. Dotted curve for Bayesian classification. Solid curve for half-way rule.

Figure 4.10: Efficiency of 2×3 independence test with respect to LRTS for mixtures by minor control allele frequency. Two classification rules: Bayesian, Halfway.

Notes: Separation of components $D = 3$. Difference in minor allele frequencies $\Delta p = 0.1$. Dotted curve for Bayesian classification. Solid curve for half-way rule.

Figure 4.11: Efficiency of 2×3 independence test with respect to LRTS for mixtures by minor control allele frequency. Two classification rules: Bayesian, Halfway.
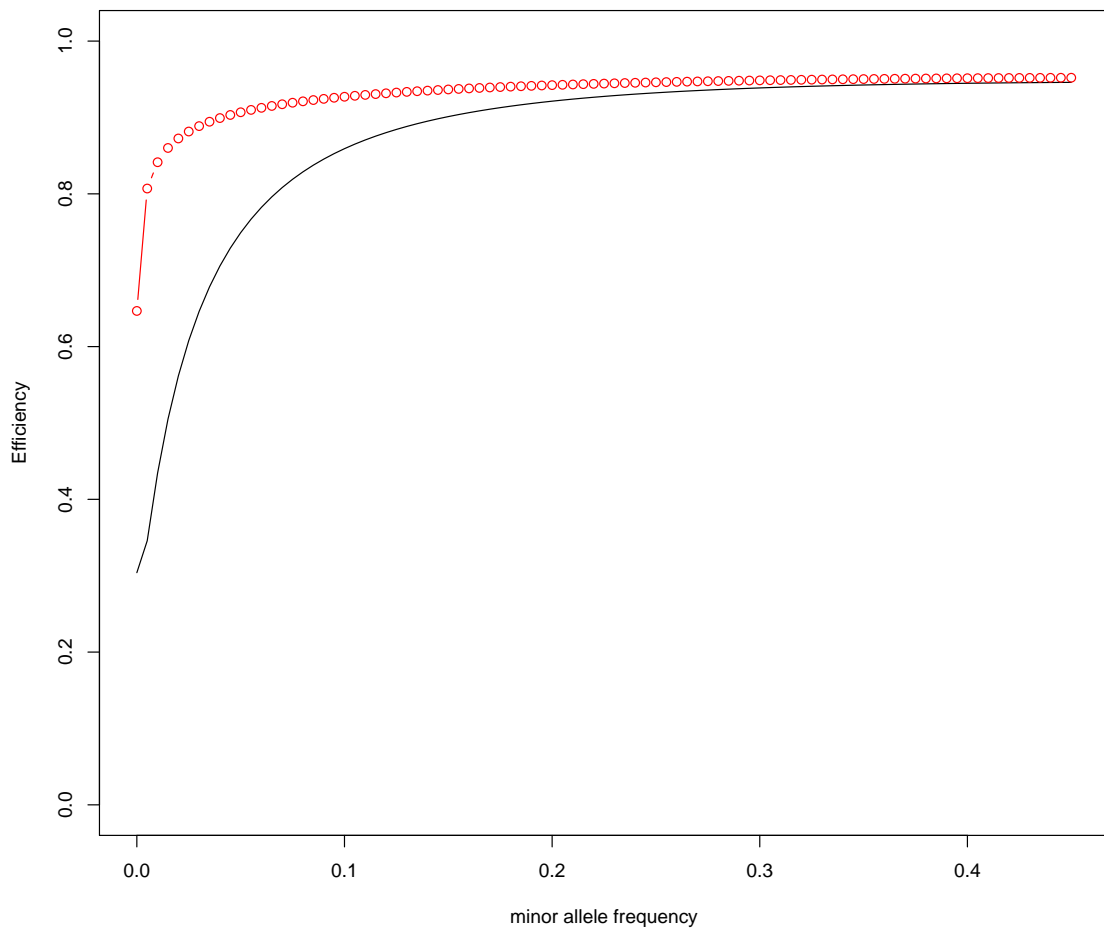

Notes: Separation of components $D = 3$. Difference in minor allele frequencies $\Delta p = 0.01$. Dotted curve for Bayesian classification. Solid curve for half-way rule.

Figure 4.12: Efficiency of 2×3 independence test with respect to LRTS for mixtures by minor control allele frequency. Two classification rules: Bayesian, Halfway.

Notes: Separation of components $D = 3$. Difference in minor allele frequencies $\Delta p = 0.001$. Dotted curve for Bayesian classification. Solid curve for half-way rule.

4.4 Limiting Behaviors of the Relative Efficiency

In this section, we calculate the limiting relative efficiencies for three component mixtures. Assume the HWE in both case and control groups, and let $\Delta = p_a - p_u$ be the difference of minor allele frequencies.

$$p_{a1} - p_{u1} = \Delta(2p_u + \Delta)$$

$$p_{a2} - p_{u2} = 2\Delta(1 - 2p_u - \Delta)$$

$$p_{a3} - p_{u3} = \Delta(-2 + 2p_u + \Delta).$$

Then the non-centrality parameter for the LRT can be written

$$\lambda_{LRT} = \frac{(\Delta)^2 N}{4}\left[J_{11}(2p_u + \Delta)^2 + 2J_{12}(2p_u + \Delta)(-2 + 2p_u + \Delta) + J_{22}(-2 + 2p_u + \Delta)^2\right].$$

The non-centrality parameter of a 2×3 chi-square test

$$\lambda_{CS} = \frac{(\Delta)^2 N}{2}\sum_{i=1}^{3}\frac{A_i}{B_i},$$

where $A_i = [\epsilon_{i1}(2p_u + \Delta) + 2\epsilon_{i2}(1 - 2p_u - 2\Delta) + \epsilon_{i3}(-2 + 2p_u + \Delta)]^2$ and $B_i = \epsilon_{i1}[(p_u+\Delta)^2+p_u^2]+2\epsilon_{i2}[(p_u+\Delta)(1-p_u-\Delta)+p_u(1-p_u)]+\epsilon_{i3}[(1-p_u-\Delta)^2+(1-p_u)^2].$

The limit of the relative efficiency as $\Delta \to 0$

$$\lim_{\Delta \to 0}\frac{\lambda_{CS}}{\lambda_{LRT}} = \frac{\sum_{i=1}^{3}\frac{[\epsilon_{i1}(2p_u)+2\epsilon_{i2}(1-2p_u)+\epsilon_{i3}(-2+2p_u)]^2}{\epsilon_{i1}p_u^2+2\epsilon_{i2}p_u(1-p_u)+\epsilon_{i3}(1-p_u)^2}}{\left[J_{11}(2p_u)^2 + 2J_{12}(2p_u)(-2 + 2p_u) + J_{22}(-2 + 2p_u)^2\right]}.$$

By taking $\lim_{p_u \to 0}\lim_{\Delta \to 0}$ or $\lim_{\Delta \to 0}\lim_{p_u \to 0}$ to the relative efficiency,

$$\lim_{\Delta \to 0}\lim_{p_u \to 0}\frac{\lambda_{CS}}{\lambda_{LRT}} = \lim_{p_u \to 0}\lim_{\Delta \to 0}\frac{\lambda_{CS}}{\lambda_{LRT}} = \frac{\sum_{i=1}^{3}\frac{(\epsilon_{i3}-\epsilon_{i2})^2}{\epsilon_{i3}}}{\int\frac{(f_3-f_2)^2}{f_3}}, \tag{4.9}$$

where f_i is the ith component probability density function.

Example 4.6

Suppose that $f_i = f(x|\mu_i, \sigma)$ is a normal density with mean μ_i and standard deviation σ. For simplicity, let $D = \mu_3 = -\mu_1$, $\mu_2 = 0$ and $\sigma = 1$. The limit of the relative efficiency (4.9) is

$$\frac{\sum_{i=1}^{3} \frac{(\epsilon_{i3}-\epsilon_{i2})^2}{\epsilon_{i3}}}{\int \frac{(f_3-f_2)^2}{f_3}} = \frac{1}{e^{D^2}-1} \sum_{i=1}^{3} \frac{(\epsilon_{i3}-\epsilon_{i2})^2}{\epsilon_{i3}}. \tag{4.10}$$

We have different limiting behaviors depending upon which classification rule is applied.

(1) The limit of the relative efficiency of the half-way rule:

The two boundary points $\gamma_1^H = -D/2$ and $\gamma_2^H = D/2$ do not depend on the minor allele frequencies or the difference in the minor allele frequencies in Example 4.2. The limit of the relative efficiency (4.10) is always positive because $(\epsilon_{i3} - \epsilon_{i2})^2 > 0$ and $\epsilon_{i3} > 0$.

(2) The limit of the relative efficiency of Bayes rule:

From the equation (4.7), γ_i^B tends to the negative infinity as p_u and Δ decrease to zero. Therefore, $\epsilon_{i3} - \epsilon_{i2} \to 0$ and the limit of the relative efficiency (4.10) approaches zero. $\square$

Chapter 5

Conclusion

The asymptotic null distribution of LRTS of the null hypothesis of single component versus the alternative of two or more components in a mixture with known mixing proportions is a 50:50 mixture of a point mass at zero and a chi-square of degree of freedom one. The asymptotic null distribution does not depend on the number of components in the alternative hypothesis. A mixture model with the non-normal component probability density functions with known mixing proportions has potential applications to association mapping in population genetics. For example, an exponential or Weibull distribution can be a possible model for a survival type quantitative trait locus (QTL) such as the flowering time [2]. This dissertation extended the asymptotic results of the LRTS of a mixture with known mixing proportions to the case in which the component probability density function is from an exponential family with one parameter. When the component probability density has more than one parameter, the asymptotic results are still valid if the partial derivatives are linearly independent. The next research is to find the asymptotic power of the LRTS of mixtures with known mixing proportions.

In the second part of the dissertation, we calculated the non-centrality parameter of the LRTS of the mixture models in case/control studies. The tables and figures

in Chapter 4 show that the LRTS without genotype calls has better power than the 2×3 independence test in which subjects were assigned to a genotype by a half-way rule or a Bayesian clustering rule. Specifically, the relative efficiency of the independence test to the LRT decreases as the minor allele frequency becomes small and the separation of the components of mixtures becomes small. The non-centrality parameter is derived for an arbitrary fixed number of components. We have not assumed that the component probability is normal. This result can be applied to a genetic association case/control study such as use of the copy number variants (CNV.) In these studies, the number of components may be greater than three. The technique of the non-centrality parameter calculations can be extended to the likelihood ratio test for a trend, and it can be compared to the non-centrality parameter of the Cochran-Armitage test [1].

Bibliography

[1] K. Ahn, C. Haynes, W. Kim, R. St Fleur, D. Gordon, and S. J. Finch. The effects of SNP genotyping errors on the power of the cochran-armitage linear trend test for case/control association studies. *Annals of Human Genetics*, 71:249–U4, Mar 2007.

[2] M. J. Aranzana, S. Kim, K. Y. Zhao, E. Bakker, M. Horton, K. Jakob, C. Lister, J. Molitor, C. Shindo, C. L. Tang, C. Toomajian, B. Traw, H. G. Zheng, J. Bergelson, C. Dean, P. Marjoram, and M. Nordborg. Genome-wide association mapping in arabidopsis identifies previously known flowering time and pathogen resistance genes. *Plos Genetics*, 1(5):531–539, Nov 2005.

[3] Herman Chernoff. On the distribution of the likelihood ratio. *Ann. Math. Statistics*, 25:573–578, 1954.

[4] D. Dacunha-Castelle and É. Gassiat. Testing in locally conic models, and application to mixture models. *ESAIM Probab. Statist.*, 1:285–317 (electronic), 1995/97.

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

[6] Stephen J. Finch, Nancy R. Mendell, and Jr. Thode, Henry C. Probabilistic measures of adequacy of a numerical search for a global maximum. *Journal of the American Statistical Association*, 84(408):1020–1023, 1989.

[7] D. W. Fulker and S. S. Cherny. An improved multipoint sib-pair analysis of quantitative traits. *Behavior Genetics*, 26(5):527–532, Sep 1996.

[8] J.H. Ghosh and P.K. Sen. On the asymptotic performance of the loglikelihood ratio statistic for the mixture model and related results. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, volume 2, pages 789–806, 1985.

[9] Bruno Goffinet, Patrice Loisel, and Beatrice Laurent. Testing in normal mixture models when the proportions are known. *Biometrika*, 79(4):842–846, Dec. 1992.

[10] D. Gordon, S. J. Finch, M. Nothnagel, and J. Ott. Power and sample size calculations for case-control genetic association tests when errors are present: Application to single nucleotide polymorphisms. *Human Heredity*, 54(1):22–33, 2002.

[11] J. A. Hartigan. A failure of likelihood asymptotics for normal mixtures. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, volume 2, pages 807–810, 1985.

[12] H. S. Kang, Z. H. S. Qin, T. H. Niu, and J. S. Liu. Incorporating genotyping uncertainty in haplotype inference for single-nucleotide polymorphisms. *American Journal of Human Genetics*, 74(3):495–510, Mar 2004.

[13] S. J. Kang, S. J. Finch, C. Haynes, and D. Gordon. Quantifying the percent increase in minimum sample size for snp genotyping errors in genetic model-based association studies. *Human Heredity*, 58(3-4):139–144, 2004.

[14] L. Kruglyak and E. S. Lander. Complete multipoint sib-pair analysis of qualitative and quantitative traits. *American Jounal of Human Genetics*, 57(2):439–454, Aug 1995.

[15] Solomon Kullback. *Information theory and statistics*. Dover Publications Inc., Mineola, NY, 1997. Reprint of the second (1968) edition.

[16] E. L. Lehman and George Casella. *Theory of pont estimation*. Springer-Velarg New York, Inc., 1998.

[17] Xin Liu, Cristian Pasarica, and Yongzhao Shao. Testing homogeneity in gamma mixture models. *Scand. J. Statist.*, 30(1):227–239, 2003.

[18] Xin Liu and Yongzhao Shao. Asymptotics for likelihood ratio tests under loss of identifiability. *Ann. Statist.*, 31(3):807–832, 2003.

[19] Xin Liu and Yongzhao Shao. Asymptotics for the likelihood ratio test in a two-component normal mixture model. *J. Statist. Plann. Inference*, 123(1):61–81, 2004.

[20] Patrice Loisel, Bruno Goffinet, Herve Monod, and Gustavo Montes De Oca. Detecting a major gene in an F2 population. *Biometrics*, 50(2):512–516, 1994.

[21] Michael Lynch and Bruce Walsh. *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, 1998.

[22] G. J. McLachlan. On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture. *Applied Statistics*, 36(3):318–324, 1987.

[23] Geoffrey McLachlan and David Peel. *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York, 2000.

[24] Sujit Kumar Mitra. On the limiting power function of the frequency chi-square test. *The Annals of Mathematical Statistics*, 29(4):1221–1233, 1958.

[25] Karl Pearson. Contributions to the mathematical theory of evolution. *Journal of the Royal Statistical Society*, 56(4):675–679, 1893.

[26] C. Radhakrishna Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):159–203, 1948.

[27] Richard Redner. Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *The Annals of Statistics*, 9(1):225–228, 1981.

[28] Steven G. Self and Kung-Yee Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc.*, 82(398):605–610, 1987.

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

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

[31] D.M. Titterington, A.F.M. Titterington, and U.E. Makov. *Statistical analysis of finite mixture distributions*. Wiley-Interscience, New York, 1985.

[32] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[33] Abraham Wald. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):595–601, 1949.

[34] C. F. Jeff Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.