Distribution Agreement

Signature:

______________________________ _______________

Mingrui Zhang Date

Hypothesis testing on the number of components in finite mixture models

By

Mingrui Zhang
Master of Science in Public Health

Department of Biostatistics and Bioinformatics

____

John Hanfelt, PhD
Thesis Advisor

____

Limin Peng, PhD
Reader

Hypothesis testing on the number of components in finite mixture models

By

Mingrui Zhang

B.S.
University of Science and Technology of China
2018

Thesis Committee Chair: John Hanfelt, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2020

Abstract

Hypothesis testing on the number of components in finite mixture models
By Mingrui Zhang

In this paper, we develop a mathematical framework for studying finite mixture models based on a quotient space, a parameter space viewing parameterizations corresponding to same probability distribution as same equivalence class. The quotient space is used to solve the issue of identifiability in finite mixture models, which makes the study of asymptotic properties of maximum likelihood estimation (MLE) possible. In the quotient space, we prove the consistency of MLE under some conditions and use simulation designs to show the performance of the point estimation of parameters by EM algorithm. Also, we propose a generalized Wald test based on resampling. By simulation studies, we show that our generalized Wald tests under two-component Gaussian mixture models may be more powerful than the likelihood ratio tests in many cases.

Hypothesis testing on the number of components in finite mixture models

By

Mingrui Zhang

B.S.
University of Science and Technology of China
2018

Thesis Committee Chair: John Hanfelt, PhD

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2020

1 Introduction

Finite mixture models, used to model data sampled from multiple underlying sources, have been widely applied to various fields; see McLachlan and Peel (2004) [1] for a general introduction to finite mixture models. From a practical prospective, Schlattmann (2009) [2] illustrated the idea of heterogeneity in medicine and introduced some medical applications of finite mixture models. A finite mixture model could be preferable over some other models on dealing with unobserved heterogeneity, since it views the total variability of the data as two parts: variability between latent groups and variability of individuals within each group [2].

The statistical issue of selecting the number of components in finite mixture models has received increasing attention over years. One classical approach of the problem is the likelihood ratio test. Suppose we are interested in testing the null hypothesis that there exists k_0 components against the alternative hypothesis that there exists k_1 components, for some $k_1 > k_0$. The likelihood ratio test statistics can be obtained by the unrestricted maximum likelihood estimation in the parameter space with k_1 components, and restricted maximum likelihood estimation in the parameter space with k_0 components. However, the likelihood ratio statistics fails to follow an asymptotic chi-squared distribution due to the violation of identifiability and the singularity of Fisher information [3-4]. Some researchers [5-7] have shown that the asymptotic distribution is related to the Gaussian Process by studying some specific distribution families. Hartigan (1985) [7] found that the likelihood ratio statistics could be asymptotically unbounded, which makes it hard to obtain the asymptotic distribution of the likelihood ratio statistics under the null hypothesis. Therefore, Chen et al. (2001) [8] suggested a modified likelihood ratio test by modifying the likelihood function. Also, bootstrap can be served as another solution to determine the rejection region, as discussed by McLachlan (1987) [9], for example. Recently, under some assumptions of the distribution family, some other hypothesis tests have been developed, including testings using measurement by weighted relative entropy [10], L^2 distance [11], and goodness-of-fit [12], moment-based tests [13], and local score tests [14]. Besides hypothesis testing, information criteria in model selection, such as Akaike's information criterion [15] (AIC) and Bayesian information criterion [16] (BIC), can be applied to this statistical problem. However, information criterion cannot consistently estimate the true number of components. For example, it has been studied that AIC may underestimate the order of a model in various statistical scenarios [17-20]

In this paper, we develop a generalized Wald test on the number of components of finite mixture models. Specifically, similar to Redner's previous work [21], we define a quotient topological space as the new parameter space. Since there are various parameterizations for finite mixture models, the mapping from the original parameter space to the model space is not one-to-one. By viewing parameterizations corresponding to same probability distribution as same equivalence class and defining the set of all equivalence classes as a quotient space, we can solve the issue of identifiability in finite mixture models. Also, we define a metric on the quotient space so that we can study the consistency of maximum likelihood estimation (MLE) under some conditions. Under the consistency of MLE, we construct the generalized Wald test by the use of resampling. Finally, by simulation studies, we compare our generalized Wald tests with the likelihood ratio tests under two-component Gaussian mixture models.

2 Methods

2.1 Finite mixture models

We are interested a general class of finite mixture models, which can be represented as

$$\mathcal{M}_k = \left\{ \lambda_1 f(x;\mu_1) + \lambda_2 f(x;\mu_2) + ... + \lambda_k f(x;\mu_k) \middle| \sum_{i=1}^{k} \lambda_i = 1, 0 \leq \lambda_i \leq 1, \mu_i \in \mathcal{C} \right\}$$

where $\{f(x;\mu) | \mu \in \mathcal{C}\}$ is a family of probability density (mass) function of interest, $\mathcal{C}$ is a compact set, and k is the number of components in the mixture model. Let

$$\Theta_k = \left\{ (\lambda_1, ...\lambda_k, \mu_1, ..., \mu_k) \middle| \sum_{i=1}^{k} \lambda_i = 1, 0 \leq \lambda_i \leq 1, \mu_i \in \mathcal{C} \right\}.$$

Due to the label-switching problem and various parameterizations for degenerate mixture models, the mapping from Θ_k to $\mathcal{M}_k$ is not one-to-one. Following Redner's previous work [21], to satisfy the identifiability condition, we define an equivalence relation $\sim$ on the parameter space Θ_k satisfying the following three properties

(1) $(\lambda_1, ..., \lambda_i, ..., \lambda_j, ..., \lambda_k, \mu_1, ..., \mu_i, ..., \mu_j, ..., \mu_k) \sim (\lambda_1, ..., \lambda_j, ..., \lambda_i, ..., \lambda_k, \mu_1, ..., \mu_j, ..., \mu_i, ..., \mu_k)$, for any $(\lambda_1, ..., \lambda_k, \mu_1, ..., \mu_k) \in \Theta_k$, and for any $1 \leq i, j \leq k$,

(2) $(\lambda_1, ..., \lambda_{i-1}, 0, \lambda_{i+1}, ..., \lambda_j, ..., \lambda_k, \mu_1, ..., \mu_i, ..., \mu_k) \sim (\lambda_1, ..., \lambda_{i-1}, \lambda'_i, \lambda_{i+1}, ..., \lambda'_j, ..., \lambda_k, \mu_1, ..., \mu_{i-1}, \mu_j, \mu_{i+1}, ..., \mu_k)$, for any $(\lambda_1, ..., \lambda_{i-1}, 0, \lambda_{i+1}, ..., \lambda_k, \mu_1, ..., \mu_k) \in \Theta_k$, and for any $1 \leq i, j \leq k$, $\lambda'_i, \lambda'_j \geq 0$ and $\lambda'_i + \lambda'_j = \lambda_j$,

(3) if $\theta_1 \sim \theta_2$ and $\theta_2 \sim \theta_3$ then $\theta_1 \sim \theta_3$ for all $\theta_1, \theta_2, \theta_3 \in \Theta_k$.

Consider the metric d on Θ_k such that

$$d((\lambda_1, ..., \lambda_k, \mu_1, ..., \mu_k), (\lambda'_1, ..., \lambda'_k, \mu'_1, ..., \mu'_k)) = \sum_{1 \leq i \leq k} (|\lambda_i - \lambda'_i| + ||\mu_i - \mu'_i||)$$

for every $(\lambda_1, ..., \lambda_k, \mu_1, ..., \mu_k) \in \Theta_k$ and $(\lambda'_1, ..., \lambda'_k, \mu'_1, ..., \mu'_k) \in \Theta_k$, where $|\cdot|$ denotes the Euclidean distance in $\mathbb{R}$ and $||\cdot||$ denotes the distance in $\mathcal{C}$. It is easy to check d is a valid metric on Θ_k. Then we focus on the quotient metric space $(\Theta_k/\sim, d/\sim)$ defined by

$$\Theta_k/\sim \doteq \left\{ [\theta] \middle| \theta \in \Theta \right\} = \left\{ \{\theta' \in \Theta | \theta' \sim \theta\} \middle| \theta \in \Theta \right\}$$

and

$$(d/\sim)([\theta_1],[\theta_2]) = \inf\left\{d(\theta'_{a_1},\theta'_{b_1})+...+d(\theta'_{a_m},\theta'_{b_m})\middle|\theta'_{a_1}\in[\theta_1],\theta'_{b_1}\sim\theta'_{a_2},...,\theta'_{b_{m-1}}\sim\theta'_{a_m},\theta'_{b_m}\in[\theta_2]\right\}$$

where $[\theta_1],[\theta_2]\in\Theta_k/\sim$. By the above definition, it is easy to show that there exists a natural bijective mapping Φ_k from $\Theta_k/\sim$ to $\mathcal{M}_k$. Also, we can prove that $(d/\sim)$ is a valid metric on $\Theta_k/\sim$ satisfying non-negativity, identity of indiscernible, symmetry, and subadditivity.

Proposition 1. *$(d/\sim)$ is a valid metric on $\Theta_k/\sim$.*

For different k, we can define the equivalence relation $\sim$ and quotient space similarly. Here, we use the same notation $\sim$, for simplicity, across different choice of k.

Since we have a metric on $\Theta_k/\sim$, we can measure the distance of two points in the quotient space, which makes the study of consistency possible. Although Redner [21] used the topology of quotient space to study consistency without a metric defined, the topology is not natural and it is hard for explanation, since the topology is not based on a metric. That's why we define a metric on the quotient space. It should be noted that the metric defined on the quotient space is the infimum of sum of distance of any finite routes rather than the simple the infimum of distance of representatives of two quotient sets, such that the subadditivity or the triangle inequality can be satisfied. The metric d on space Θ_k is not arbitrary as well. A bad metric on Θ_k may lead to the violation of identity of indiscernible for the induced metric on $\Theta_k/\sim$. Here our choice makes $(d/\sim)$ both valid as a metric and easy for calculation.

2.2 Hypothesis test

Suppose that $X_1,...,X_n$ are i.i.d. sampled from a distribution $p(x)\in\mathcal{M}_k$, with parameter $[\theta]=[\lambda_1,...,\lambda_k,\mu_1,...,\mu_k]$. First, we can prove the consistency of MLE under some conditions.

Proposition 2. *Let $X_1,...,X_n$ be i.i.d. sampled from the distribution $p(x;[\theta_0])=\lambda_{1,0}f(x;\mu_{1,0})+\lambda_{2,0}f(x;\mu_{2,0})+...+\lambda_{k,0}f(x;\mu_{k,0})$, where $p(x;[\theta_0])\in\mathcal{M}_k$. Denote $[\hat{\theta}]=[(\hat{\lambda}_1,...,\hat{\lambda}_k,\hat{\mu}_1,...,\hat{\mu}_k)]$ as the MLE in the quotient space $\Theta_k/\sim$. If the distribution family satisfies*

(1) $f(x;\mu)$ is continuous in μ for all $\mu\in\mathcal{C}$ and all $x\in\mathcal{X}$;

(2) there exists a function $d(x)$ such that $|\log p(x;[\theta])|\le d(x)$ for all $[\theta]\in\Theta_k/\sim$ and all $x\in\mathcal{X}$, and $\mathrm{E}_{[\theta_0]}[d(X)]<\infty$;

(3) $Q_0([\theta])=\mathrm{E}_{[\theta_0]}[\log p(X;[\theta])]$ is uniquely maximized at $[\theta_0]$;

Then

$$[(\hat{\lambda}_1,...,\hat{\lambda}_k,\hat{\mu}_1,...,\hat{\mu}_k)]\xrightarrow{p}[(\lambda_{1,0},...,\lambda_{k,0},\mu_{1,0},...,\mu_{k,0})]$$

with respect to $(d/\sim)$.

Then we want to test the null hypothesis test $H_0 : k = k_0$ against the alternative hypothesis $H_A : k = k_0 + 1$. To conduct the hypothesis test, we focus on the parameter space $\Theta_{k_0+1}/\sim$ and $\Theta_{k_0}/\sim$. Let

$$l_n([\lambda_1, ..., \lambda_{k_0+1}, \mu_1, ..., \mu_{k_0+1}]) = \sum_{i=1}^{n} \log \left(\lambda_1 f(X_i; \mu_1) + ... + \lambda_{k_0+1} f(X_i; \mu_{k_0+1}) \right)$$

where $[\lambda_1, ..., \lambda_{k_0+1}, \mu_1, ..., \mu_{k_0+1}] \in \Theta_{k_0+1}/\sim$ and

$$l_n([\lambda_1, ..., \lambda_{k_0}, \mu_1, ..., \mu_{k_0}]) = \sum_{i=1}^{n} \log \left(\lambda_1 f(X_i; \mu_1) + ... + \lambda_{k_0} f(X_i; \mu_{k_0}) \right)$$

where $[\lambda_1, ..., \lambda_{k_0}, \mu_1, ..., \mu_{k_0+1}] \in \Theta_{k_0}/\sim$. Then the likelihood ratio test statistic is

$$LR = 2\left(\sup_{[\theta] \in \Theta_{k_0+1}/\sim} l_n([\theta]) - \sup_{[\theta] \in \Theta_{k_0}/\sim} l_n([\theta]) \right).$$

If the unrestricted MLE and restricted MLE can be expressed as

$$[\hat{\theta}] = [(\hat{\lambda}_1, ..., \hat{\lambda}_{k_0+1}, \hat{\mu}_1, ..., \hat{\mu}_{k_0+1})] = \operatorname*{arg\,max}_{[\theta] \in \Theta_{k_0+1}/\sim} l_n([\theta])$$

and

$$[\tilde{\theta}] = [(\tilde{\lambda}_1, ..., \tilde{\lambda}_{k_0}, \tilde{\mu}_0, ..., \tilde{\mu}_{k_0})] = \operatorname*{arg\,max}_{[\theta] \in \Theta_{k_0}/\sim} l_n([\theta]),$$

then the likelihood ratio test can be written as

$$LR = 2\left(l_n([\hat{\theta}]) - l_n([\tilde{\theta}]) \right).$$

Our proposed generalized Wald test statistic has the following form

$$g([\hat{\theta}]) = \left(\min_{1 \le i \le k_0+1} \hat{\lambda}_i \right) \left(\min_{1 \le i < j \le k_0+1} |\hat{\mu}_i - \hat{\mu}_j| \right)^{\alpha}$$

where α is a free positive real number. It can be shown easily that $g([\theta])$ is a valid function that does not depend on the choice of $\theta \in [\theta]$. In fact, the functional forms of the test statistic are not unique, and there is no guarantee that one form of the test statistic is uniformly powerful than others. Our generalized Wald test statistic is only one of the possible choices that are simple and reasonable.

The idea of our proposed generalized Wald test is that we are more likely to reject the null hypothesis for larger distance from unrestricted MLE to the restricted parameter space $\Theta_{k_0}/\sim$. However, the metric $d/\sim$ has some drawbacks. First, the distance contributed by the component of $(\lambda_1, ..., \lambda_{k_0+1})$ and the component of $(\mu_1, ..., \mu_{k_0+1})$ are not comparable, thus it is not reasonable to view them equally. Second, the distance to the restricted parameter space is not smooth. For example, in the model of two-component mixture model, the distance from $(\hat{\lambda}_1, \hat{\lambda}_2, \hat{\mu}_1, \hat{\mu}_2) = (0.9, 0.1, 0, t)$ to the restricted parameter space (homogeneous model) is $\min\{0.1, |t|\}$. Therefore, we consider a function of unrestricted MLE as the test statistic which could measure the distance instead. To illustrate the advantages of the generalized Wald test over the likelihood ratio test or score test, we simulate data of size 100 from a two-component Gaussian mixture model $0.3N(0,1) + 0.7N(1,1)$, and plot the likelihood function as Figure 1. We can learn from the plot that the surface near the global maximum could be very flat. In fact, although the difference between unrestricted MLE and restricted MLE may be large, their log likelihood values are very close in most cases. Therefore, we would like to conduct a Wald-based test rather than the likelihood ratio test. Additionally, the restricted MLE could be a saddle point of the log likelihood function, which makes the score test powerless as well.

2.3 Computation

The computation burden of both the generalized Wald test and the likelihood ratio test is large, because there is not an analytical solution to the MLE of finite mixture model, and the asymptotic distribution of test statistics fails to be regular. First, we review the expectation-maximization (EM) algorithm [22], an iterative method to find the MLE in the presence of hidden variables. The classical application of EM algorithm is on the finite mixture model, especially the Gaussian mixture models. Consider the latent random variable Z that takes values $1, ..., k$ following the multinomial distribution

$$Z \sim \text{Multi}(\lambda_1, ..., \lambda_k).$$

Then, if the conditional random variable $X|Z$ follows the distribution of interest

$$X|Z \sim f(x; \mu_Z),$$

the marginal distribution of X follows the distribution

$$X \sim f(x; \lambda_1, ..., \lambda_k, \mu_1, ..., \mu_k) = \lambda_1 f(x; \mu_1) + \lambda_2 f(x; \mu_2) + ... + \lambda_k f(x; \mu_k)$$

which gives an explanation of the finite mixture model. Consider the following complete likelihood function

$$L(\theta; X, Z) = \prod_{i=1}^{n} f(X_i, Z_i; \theta).$$

The EM algorithm is to locate MLE as following [23]:

- Expectation step (E step): Define the Q function as the conditional expectation of complete log likelihood function $l_n(\theta) = \sum_{i=1}^{n} \log f(X_i, Z_i; \theta)$, given all the X_i and current estimates of parameters $\theta^{(i)}$:

$$Q(\theta; \theta^{(i)}) = \text{E}[\log f(X, Z; \theta)|X, \theta^{(i)}]$$

- Maximization step (M step): Find the parameters that maximize the quantity:

$$\theta^{(i+1)} = \arg\max_{\theta} Q(\theta;\theta^{(i)})$$

Some works [24] have shown that the algorithm will converge to a local maximum, and the global maximum can be elsewhere. However, EM algorithm is still the most popular way for finding MLE in finite mixture models.

Next, there are various resampling approaches to determine the rejection region of the likelihood ratio test and the generalized Wald test. In this paper, we will apply the following Monte-Carlo simulation-based procedure [25] in the part of simulation study:

- Obtain the restricted MLE by EM algorithm.
- Repeat simulating data by restricted MLE for N times, obtain its unrestricted and restricted MLE by EM algorithm, and calculate the test statistics.
- Determine the rejection region by the 0.95 quantile of N simulated statistics, N selected as 1000 for the following study.

3 Study on Gaussian mixture models

In this section, we focus on the two-component Gaussian mixture models with known variance. We write the model as

$$\mathcal{M} = \left\{ \lambda_1 N(\mu_1, 1) + \lambda_2 N(\mu_2, 1) \middle| \lambda_1 + \lambda_2 = 1, 0 \le \lambda_1, \lambda_2 \le 1, \mu_1, \mu_2 \in \mathbb{R} \right\}$$

3.1 Evaluation of point estimation by EM algorithm

Since it is not guaranteed that the EM algorithm will converge to the global maximum of the likelihood function, i.e. the MLE, we use simulation study to evaluate the estimate of parameters by EM algorithm. In this special case of two-component Gaussian mixture model, the EM algorithm can be expressed in the following iteration [23].

- Initialize the means and mixing coefficients $\hat{\lambda}_1^{(0)}$, $\hat{\lambda}_2^{(0)}$, $\hat{\mu}_1^{(0)}$, $\hat{\mu}_2^{(0)}$.
- (E step) Evaluate the responsibilities using current parameters $\hat{\lambda}_1^{(i)}$, $\hat{\lambda}_2^{(i)}$, $\hat{\mu}_1^{(i)}$, $\hat{\mu}_2^{(i)}$.

  $$\hat{\gamma}_{jk}^{(i+1)} = \frac{\hat{\lambda}_k^{(i)}\phi(X_j;\hat{\mu}_k^{(i)})}{\hat{\lambda}_1^{(i)}\phi(X_j;\hat{\mu}_1^{(i)}) + \hat{\lambda}_2^{(i)}\phi(X_j;\hat{\mu}_2^{(i)})} \quad 1 \le j \le n, 1 \le k \le 2$$

  where $\phi(x;\mu)$ is the density function of normal distribution with mean μ and variance 1.
- (M step) Evaluate the parameters using current responsibilities

  $$\hat{\lambda}_k^{(i+1)} = \frac{\sum_{j=1}^{n} \hat{\gamma}_{jk}^{(i+1)}}{n} \quad 1 \le k \le 2$$

$$\hat{\mu}_k^{(i+1)} = \frac{\sum_{j=1}^{n} \hat{\gamma}_{jk}^{(i+1)} X_j}{\sum_{j=1}^{n} \hat{\gamma}_{jk}^{(i+1)}} \quad 1 \leq k \leq 2$$

- Repeat E step and M step, until convergence of parameters.

We conduct the Monte-Carlo simulation. Consider the true model $0.3N(0,1) + 0.7N(\mu,1)$, where μ is chosen as 0.5, 1, and 1.5 respectively. Repeat sampling data with size M from the true model for 1000 times and obtaining the estimate of parameters $[\hat{\theta}] = [\hat{\lambda}_1, \hat{\lambda}_2, \hat{\mu}_1, \hat{\mu}_2]$, where M is chosen as 100, 1000 and 5000 respectively.

The mean squared error (MSE) of the point estimation are shown in Table 1. It should be noted that in Table 1, μ_1^* refers to the smaller value between μ_1 and μ_2 in $[\lambda_1, \lambda_2, \mu_1, \mu_2]$, and λ_1^* refers to the mixing coefficient corresponding to μ_1^*; while μ_2^* refers to the larger value between μ_1 and μ_2 in $[\lambda_1, \lambda_2, \mu_1, \mu_2]$, and λ_2^* refers to the mixing coefficient corresponding to μ_2^*. From Table 1, we can see that the MSE of point estimation of parameters decreases as μ increases, which implies that EM algorithm performs better when the heterogeneity is more significant. Also, sample size is an important factor that influences the point estimation.

3.2 Simulation study on the power of hypothesis testing

In this subsection, we conduct the Monte-Carlo simulation study to compare the power of the likelihood ratio test and the generalized Wald test. Consider the true model $\lambda N(0,1) + (1-\lambda)N(\mu,1)$, where μ is chosen as $0.1, 0.2, ..., 2.0$, and λ is chosen as 0.1, 0.3 and 0.5 respectively. Repeat sampling data with size 1000 from the true model for 10000 times and conduct the generalized Wald test and the likelihood ratio test.

The results are shown in Table 2-4. From the results, we can learn that for $\alpha < 1$, the curve of power versus the change of μ is not always increasing especially when $\lambda = 0.1$, which suggests that $\alpha < 1$ is not a good choice for general testing. By comparing the performance of $\alpha = 1, 2, 3, 4, 5$, although $\alpha = 3$ is better than $\alpha = 2$ when $\lambda = 0.1$ and $\mu > 1$, $\alpha = 2$ is a generally good choice for most cases, including when $\lambda = 0.1$ and $0 < \mu \leq 1$, $\lambda = 0.3$, and $\lambda = 0.5$. The Figure 2-4 show the change of power for the generalized Wald test when $\alpha = 2$ and the likelihood ratio test. We can see that when $\lambda = 0.3$ and $\lambda = 0.5$, the curve of our test is almost completely above the curve of the likelihood ratio test. In fact, by two-proportion z test, our test is significantly powerful than the likelihood ratio test when $\lambda = 0.3$ and $\mu = 0.3, 0.4, ..., 0.8$, with average power gain of 0.028, and when $\lambda = 0.5$ and $\mu = 0.4, 0.5, ..., 0.9$, with average power gain of 0.034. However, when $\lambda = 0.1$, our test is only significantly powerful than the likelihood ratio test when $\mu = 0.7$, with a power gain of 0.0175.

4 Discussion

In this paper, we develop a mathematical framework for studying finite mixture models based on the quotient space. To study the consistency of MLE in the quotient space, a distance function should be defined. We define the distance naturally induced by the equivalence relation. However, the property of the distance function is not good enough, which makes it complicated to study the topological property of the quotient space. In future studies, we can work on designing a metric with better topological property. In the study of consistency of MLE, we assume that $\mathcal{C}$ is compact. However, when the parameter space is not compact, the MLE may not be consistent unless more strong uniform consistency is satisfied.

In the proposed quotient space, a generalized Wald test is developed. For the hypothesis testing, we are interested in the alternative hypothesis $H_A : k = k_0 + 1$, because it is straightforward to construct a generalized Wald test statistic as a measure of distance from the unrestricted MLE in $\Theta_{k_0+1}/\sim$ to the restricted parameter space $\Theta_{k_0}/\sim$. In practice, we can start with a small k_0, say 1, and test upward. Actually, we can also directly test against the hypothesis $H_A : k = k_1$ where $k_1 > k_0$, as long as we can construct a similar test statistic to measures the distance from the unrestricted MLE in $\Theta_{k_1}/\sim$ to the restricted parameter space $\Theta_{k_0}/\sim$. Obviously, the same test statistic does not work, and we need to find another reasonable test statistic. For the computation of determining the rejection region, there are some other nonparametric approaches. Future studies may work on the choice of computational approaches to conduct the tests.

From the simulation studies, for some cases, our generalized Wald test could be significantly powerful than the likelihood ratio tests. However, there are several issues for discussion. First, the choice of the functional form of the generalized Wald test is not unique. The functional form that we propose is simple, but it may not be the best. Also, $\alpha = 2$ is not the uniformly most powerful choice. For example, in our simulation study when $\lambda = 0.1$, $a = 3$ works better when μ is large. Therefore, the choice of α needs further studies. Second, the computational burden of both the generalized Wald test and the likelihood ratio test is large, due to the large time complexity of both EM algorithm and resampling. Thus, we still need to study the asymptotic property of the MLE and the test statistics. Finally, we need to study more generalized Gaussian mixture models by theoretical approaches and simulation studies based on the quotient space.

References

[1] McLachlan, G. J., & Peel, D. (2004). *Finite mixture models.* John Wiley & Sons.

[2] Schlattmann, P. (2009). *Medical applications of finite mixture models.* Berlin: Springer.

[3] Aitkin, M., & Rubin, D. B. (1985). Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society: Series B (Methodological), 47*(1), 67-75.

[4] Quinn, B. G., McLachlan, G. J., & Hjort, N. L. (1987). A note on the Aitkin-Rubin approach to hypothesis testing in mixture models. *Journal of the Royal Statistical Society: Series B (Methodological), 49*(3), 311-314.

[5] Ghosh, J. K., & Sen, P. K. (1984). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results.

[6] Titterington, D. M., Smith, A. F., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions.* Wiley,.

[7] Hartigan, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proc. Barkeley Conference in Honor of J. Neyman and J. Kiefer* (Vol. 2, pp. 807-810).

[8] Chen, H., Chen, J., & Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63*(1), 19-29.

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

[10] Li, Y., & Wang, L. (2008). Testing for homogeneity in mixture using weighted relative entropy. *Communications in Statistics—Simulation and Computation®, 37*(10), 1981-1995.

[11] Charnigo, R., & Sun, J. (2004). Testing homogeneity in a mixture distribution via the L^2 distance between competing models. *Journal of the American Statistical Association, 99*(466), 488-498.

[12] Wichitchan, S., Yao, W., & Yang, G. (2019). Hypothesis testing for finite mixture models. *Computational Statistics & Data Analysis, 132,* 180-189.

[13] Ning, W., Gupta, A. K., Yu, C., & Zhang, S. (2009). A moment-based test for homogeneity in finite mixture models. *Communications in Statistics-Theory and Methods, 38*(9), 1371-1382.

[14] Wu, Y., & Gupta, A. K. (2003). Local score tests in mixture exponential family. *Journal of statistical planning and inference, 116*(2), 421-435.

[15] Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control, 19*(6), 716-723.

[16] Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics, 6*(2), 461-464.

[17] Geweke, J., & Meese, R. (1981). Estimating regression models of finite but unknown order. *International Economic Review,* 55-70.

[18] Katz, R. W. (1981). On some criteria for estimating the order of a Markov chain. *Technometrics, 23*(3), 243-249.

[19] Koehler, A. B., & Murphree, E. S. (1988). A comparison of the Akaike and Schwarz criteria for selecting model order. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 37*(2), 187-195.

[20] Cubaynes, S., Lavergne, C., Marboutin, E., & Gimenez, O. (2012). Assessing individual heterogeneity using model selection criteria: how many mixture components in capture–recapture models?. *Methods in Ecology and Evolution, 3*(3), 564-573.

[21] Redner, R. (1981). Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *The Annals of Statistics, 9*(1), 225-228.

[22] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1-22.

[23] Bilmes, J. A. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute, 4*(510), 126.

[24] Wu, C. J. (1983). On the convergence properties of the EM algorithm. *The Annals of statistics*, 95-103.

[25] Fishman, G. (2013). *Monte Carlo: concepts, algorithms, and applications*. Springer Science & Business Media.

[26] Frangakis C. (2007). *Advanced Statistical Theory I-II* [PDF Document]. Retrieved from http://www.biostat.jhsph.edu/bstcourse/bio771/sec8a.pdf

A Proof of Proposition 1-2

We first define the standard quotient representation and its order for elements in the quotient space $\Theta_k/\sim$.

Definition 1. *For each $[\theta] \in \Theta_k/\sim$, there must exist unique $k_0 \leq k$ and $\theta' = (\lambda'_1, ..., \lambda'_k, \mu'_1, ..., \mu'_k) \in [\theta]$ such that $\mu'_1 < \mu'_2 < ... < \mu'_{k_0}$, $\mu'_{k_0+1} = ... = \mu'_k = 0$, $\lambda'_1\lambda'_2...\lambda'_{k_0} \neq 0$, and $\lambda'_{k_0+1} = ...\lambda'_k = 0$. Then, θ' is defined as the standard quotient representation of $[\theta]$ and k_0 is defined as the order of $[\theta]$.*

It is easy to show that existence and uniqueness of the standard quotient representation and its order. Before proving Proposition 1, we prove the following lemmas.

Lemma 1. *Consider $[\theta_1], [\theta_2] \in \Theta_k/\sim$, and their standard quotient representations $\theta' = (\lambda'_1, ..., \lambda'_k, \mu'_1, ..., \mu'_k) \in [\theta_1]$ with order k_1 and $\theta'' = (\lambda''_1, ..., \lambda''_k, \mu''_1, ..., \mu''_k) \in [\theta_2]$ with order k_2. Then, we have*

$$\inf_{\theta^* \in [\theta_1], \theta^{**} \in [\theta_2]} d(\theta^*, \theta^{**}) \geq \sum_{1 \leq i \leq k_1} S_{1i}$$

where

$$S_{1i} = \begin{cases} \min\{\max\{0, \lambda'_i - \lambda''_j\}, \min_{1 \leq l \leq k_2, l \neq j} ||\mu''_l - \mu'_i||\} & \textit{if } \mu''_j = \mu'_i \textit{ for some } 1 \leq j \leq k_2 \\ \min\{\lambda'_i, \min_{1 \leq l \leq k_2} ||\mu''_l - \mu'_i||\} & \textit{if } \mu''_j \neq \mu'_i \textit{ for any } 1 \leq j \leq k_2 \end{cases}$$

and similarly

$$\inf_{\theta^* \in [\theta_1], \theta^{**} \in [\theta_2]} d(\theta^*, \theta^{**}) \geq \sum_{1 \leq i \leq k_2} S_{2i}$$

where

$$S_{2i} = \begin{cases} \min\{\max\{0, \lambda''_i - \lambda'_j\}, \min_{1 \leq l \leq k_1, l \neq j} ||\mu'_l - \mu''_i||\} & \textit{if } \mu'_j = \mu''_i \textit{ for some } 1 \leq j \leq k_1 \\ \min\{\lambda''_i, \min_{1 \leq l \leq k_1} ||\mu'_l - \mu''_i||\} & \textit{if } \mu'_j \neq \mu''_i \textit{ for any } 1 \leq j \leq k_1 \end{cases}$$

Remark 1. *Define the minimum over an empty set as infinity.*

Proof. We just prove the first part of the lemma and the proof of the second part is very similar. For any $\theta^* = (\lambda^*_1, ..., \lambda^*_k, \mu^*_1, ..., \mu^*_k) \in [\theta_1]$ and any $\theta^{**} = (\lambda^{**}_1, ..., \lambda^{**}_k, \mu^{**}_1, ..., \mu^{**}_k) \in [\theta_2]$, we are interested in $d(\theta^*, \theta^{**})$. We introduce the index set $\mathcal{T}_i = \{1 \leq l \leq k | \mu^*_l = \mu'_i\}$ for $1 \leq i \leq k_1$. Then,

$$d(\theta^*, \theta^{**}) \geq \sum_{1 \leq i \leq k_1} \sum_{l \in \mathcal{T}_i} (|\lambda^*_l - \lambda^{**}_l| + ||\mu^*_l - \mu^{**}_l||)$$

For any $1 \leq i \leq k_1$, if $\mu''_j = \mu'_i$ and $\lambda''_j \geq \lambda'_i$ for some $1 \leq j \leq k_2$, then obviously

$$\sum_{l \in \mathcal{T}_i} (|\lambda^*_l - \lambda^{**}_l| + ||\mu^*_l - \mu^{**}_l||) \geq 0.$$

If $\mu''_j = \mu'_i$ and $\lambda''_j < \lambda'_i$ for some $1 \leq j \leq k_2$, then we write $\mathcal{T}_i = \mathcal{T}_{i,1} \cup \mathcal{T}_{i,2}$, where

$$\mathcal{T}_{i,1} = \{l \in \mathcal{T}_i | \mu_l^{**} = \mu_i' \text{ or } \lambda_l^{**} = 0\}$$

and

$$\mathcal{T}_{i,2} = \{l \in \mathcal{T}_i | \mu_l^{**} \neq \mu_i' \text{ and } \lambda_l^{**} \neq 0\},$$

thus we have

$$\sum_{l \in \mathcal{T}_{i,1}} (|\lambda_l^* - \lambda_l^{**}| + ||\mu_l^* - \mu_l^{**}||) \geq \lambda_i' - \lambda_j''$$

and

$$\sum_{l \in \mathcal{T}_{i,2}} (|\lambda_l^* - \lambda_l^{**}| + ||\mu_l^* - \mu_l^{**}||) \geq \min_{1 \leq l \leq k_2, l \neq j} ||\mu_l'' - \mu_i'||\}$$

which implies

$$\sum_{l \in \mathcal{T}_i} (|\lambda_l^* - \lambda_l^{**}| + ||\mu_l^* - \mu_l^{**}||) \geq \min\{\lambda_i' - \lambda_j'', \min_{1 \leq l \leq k_2, l \neq j} ||\mu_l'' - \mu_i'||\}.$$

If $\mu_j'' \neq \mu_i'$ for any $1 \leq j \leq k_2$, similarly we have

$$\sum_{l \in \mathcal{T}_i} (|\lambda_l^* - \lambda_l^{**}| + ||\mu_l^* - \mu_l^{**}||) \geq \min\{\lambda_i', \min_{1 \leq l \leq k_2} ||\mu_l'' - \mu_i'||\}.$$

Therefore,

$$d(\theta^*, \theta^{**}) \geq \sum_{1 \leq i \leq k_1} S_{1i}$$

and by taking infimum on the left side,

$$\inf_{\theta^* \in [\theta_1], \theta^{**} \in [\theta_2]} d(\theta^*, \theta^{**}) \geq \sum_{1 \leq i \leq k_1} S_{1i}.$$

□

The following result is a direct corollary of Lemma 1.

Corollary 1. *Consider $[\theta_1], [\theta_2] \in \Theta_k/\sim$, and their standard quotient representations $\theta' = (\lambda_1', ..., \lambda_k', \mu_1', ..., \mu_k') \in [\theta_1]$ with order k_1 and $\theta'' = (\lambda_1'', ..., \lambda_k'', \mu_1'', ..., \mu_k'') \in [\theta_2]$ with order k_2. Then, we have*

$$\inf_{\theta^* \in [\theta_1], \theta^{**} \in [\theta_2]} d(\theta^*, \theta^{**}) \geq \min\{A, B\}$$

where

$$A = \sum_{1 \leq i \leq k_1} A_i$$

$$A_i = \begin{cases} \max\{0, \lambda_i' - \lambda_j''\} & \textit{if } \mu_j'' = \mu_i' \textit{ for some } 1 \leq j \leq k_2 \\ \lambda_i' & \textit{if } \mu_j'' \neq \mu_i' \textit{ for any } 1 \leq j \leq k_2 \end{cases}$$

and

$$B = \min_{1 \le i \le k_1} B_i$$

$$B_i = \begin{cases} \min_{1 \le l \le k_2, l \ne j} ||\mu''_l - \mu'_i|| & \textit{if } \mu''_j = \mu'_i \textit{ for some } 1 \le j \le k_2 \\ \min_{1 \le l \le k_2} ||\mu''_l - \mu'_i|| & \textit{if } \mu''_j \ne \mu'_i \textit{ for any } 1 \le j \le k_2 \end{cases}$$

Lemma 2. *Consider* $[\theta_1], [\theta_2] \in \Theta_k/\sim$*, and their standard quotient representations* $\theta' = (\lambda'_1, ..., \lambda'_k, \mu'_1, ..., \mu'_k) \in [\theta_1]$ *with order* k_1 *and* $\theta'' = (\lambda''_1, ..., \lambda''_k, \mu''_1, ..., \mu''_k) \in [\theta_2]$ *with order* k_2*. Then, we have*

$$(d/\sim)([\theta_1], [\theta_2]) \ge \min\{A, B\}$$

where

$$A = \sum_{1 \le i \le k_1} A_i$$

$$A_i = \begin{cases} \max\{0, \lambda'_i - \lambda''_j\} & \textit{if } \mu''_j = \mu'_i \textit{ for some } 1 \le j \le k_2 \\ \lambda'_i & \textit{if } \mu''_j \ne \mu'_i \textit{ for any } 1 \le j \le k_2 \end{cases}$$

and

$$B = \min_{1 \le i \le k_1} B_i$$

$$B_i = \begin{cases} \min_{1 \le l \le k_2, l \ne j} ||\mu''_l - \mu'_i|| & \textit{if } \mu''_j = \mu'_i \textit{ for some } 1 \le j \le k_2 \\ \min_{1 \le l \le k_2} ||\mu''_l - \mu'_i|| & \textit{if } \mu''_j \ne \mu'_i \textit{ for any } 1 \le j \le k_2 \end{cases}$$

Proof. Consider any finite route from $[\theta_1]$ to $[\theta_2]$: $[\theta_1] \to [\theta_{a_1}] \to ... \to [\theta_{a_m}] \to [\theta_2]$ where $[\theta_{a_i}] \in \Theta_k/\sim$ for $1 \le i \le m$. Let their standard quotient representations be $\theta^{(a_i)} = (\lambda_1^{(a_i)}, ..., \lambda_k^{(a_i)}, \mu_1^{(a_i)}, ..., \mu_k^{(a_i)}) \in [\theta_{a_i}]$ with order k_{a_i} where $1 \le i \le m$. We prove the conclusion by mathematical induction. First, when $m = 0$, according to the Corollary 1, the conclusion is correct. Then, if the conclusion is correct for $m = 0, 1, 2, .., m_0$, when it comes to $m = m_0 + 1$, we analyze the route $[\theta_1] \to [\theta_{a_1}]$ and $[\theta_{a_1}] \to ... \to [\theta_{a_m}] \to [\theta_2]$ separately. Consider any $\theta^* = (\lambda^*_1, ..., \lambda^*_k, \mu^*_1, ..., \mu^*_k) \in [\theta_1]$ and any $\theta^{**} = (\lambda^{**}_1, ..., \lambda^{**}_k, \mu^{**}_1, ..., \mu^{**}_k) \in [\theta_{a_1}]$. We introduce the index set $\mathcal{T}_i = \{1 \le l \le k | \mu^*_l = \mu'_i\}$ for $1 \le i \le k_1$. We write $\mathcal{T}_i = \mathcal{T}_{i,1} \cup \mathcal{T}_{i,2} \cup \mathcal{T}_{i,3}$, where

$$\mathcal{T}_{i,1} = \{l \in \mathcal{T}_i | \lambda^{**}_l = 0\}$$

$$\mathcal{T}_{i,2} = \{l \in \mathcal{T}_i | \mu^{**}_l = \mu'_i, \lambda^{**}_l \ne 0\}$$

and

$$\mathcal{T}_{i,3} = \{l \in \mathcal{T}_i | \mu^{**}_l \ne \mu'_i \text{ and } \lambda^{**}_l \ne 0\}.$$

Then, along the route $[\theta_1] \to [\theta_{a_1}]$,

$$\sum_{l\in\mathcal{T}_i}(|\lambda_l^*-\lambda_l^{**}|+||\mu_l^*-\mu_l^{**}||)\geq\sum_{l\in\mathcal{T}_{i,1}}\lambda_l^*+\sum_{l\in\mathcal{T}_{i,2}}|\lambda_l^*-\lambda_l^{**}|+\sum_{l\in\mathcal{T}_{i,3}}(|\lambda_l^*-\lambda_l^{**}|+||\mu_i'-\mu_l^{**}||).$$

We consider the following two cases:

(1) If there exists $l\in\mathcal{T}_{i,3}$ such that $||\mu_l^{**}-\mu_i'||\geq\min_{1\leq p\leq k_2}||\mu_i'-\mu_p''||$, then the sum of distance along $[\theta_1]\to[\theta_{a_1}]\to...\to[\theta_{a_m}]\to[\theta_2]$ is no less than $\min\{A,B\}$.

(2) If $||\mu_l^{**}-\mu_i'||<\min_{1\leq p\leq k_2}||\mu_i'-\mu_p''||$ for any $l\in\mathcal{T}_{i,3}$, then for any $l\in\mathcal{T}_{i,3}$, $\mu_l^{**}\notin\{\mu_1'',..,\mu_2''\}$, and by triangle inequality,

$$\min_{1\leq p\leq k_2}||\mu_l^{**}-\mu_p''||+||\mu_i'-\mu_l^{**}||\geq\min_{1\leq p\leq k_2}||\mu_i'-\mu_p''||.$$

Then, we need to consider the sum of distance along the route $[\theta_{a_1}]\to...\to[\theta_{a_m}]\to[\theta_2]$. By the induction, we have

$$(d/\sim)([\theta_{a_1}],[\theta_2])\geq\min\{A',B'\}$$

where

$$A'=\sum_{1\leq i\leq k_{a_1}}A_i'$$

$$A_i'=\begin{cases}\max\{0,\lambda_i^{(a_1)}-\lambda_j''\} & \text{if } \mu_j''=\mu_i^{(a_1)} \text{ for some } 1\leq j\leq k_2\\ \lambda_i^{(a_1)} & \text{if } \mu_j''\neq\mu_i^{(a_1)} \text{ for any } 1\leq j\leq k_2\end{cases}$$

and

$$B'=\min_{1\leq i\leq k_{a_1}}B_i'$$

$$B_i'=\begin{cases}\min_{1\leq l\leq k_2,l\neq j}||\mu_l''-\mu_i^{(a_1)}|| & \text{if } \mu_j''=\mu_i^{(a_1)} \text{ for some } 1\leq j\leq k_2\\ \min_{1\leq l\leq k_2}||\mu_l''-\mu_i^{(a_1)}|| & \text{if } \mu_j''\neq\mu_i^{(a_1)} \text{ for any } 1\leq j\leq k_2\end{cases}$$

Therefore, we have

$$\begin{aligned}&\sum_{1\leq i\leq k_1}\sum_{l\in\mathcal{T}_i}(|\lambda_l^*-\lambda_l^{**}|+||\mu_l^*-\mu_l^{**}||)+(d/\sim)([\theta_{a_1}],[\theta_2])\\ &\geq\sum_{1\leq i\leq k_1}\left(\sum_{l\in\mathcal{T}_{i,1}}\lambda_l^*+\sum_{l\in\mathcal{T}_{i,2}}|\lambda_l^*-\lambda_l^{**}|+\sum_{l\in\mathcal{T}_{i,3}}(|\lambda_l^*-\lambda_l^{**}|+||\mu_i'-\mu_l^{**}||)\right)+\min\{A',B'\}\\ &\geq\min\{\sum_{1\leq i\leq k_1}\left(\sum_{l\in\mathcal{T}_{i,1}}\lambda_l^*+\sum_{l\in\mathcal{T}_{i,2}}|\lambda_l^*-\lambda_l^{**}|+\sum_{l\in\mathcal{T}_{i,3}}|\lambda_l^*-\lambda_l^{**}|\right)+A',\sum_{1\leq i\leq k_1}\sum_{l\in\mathcal{T}_{i,3}}||\mu_i'-\mu_l^{**}||+B'\}\\ &\geq\min\{A,B\}\end{aligned}$$

The conclusion is correct for $m=m_0+1$. By mathematical induction, the conclusion is correct for any finite route. $\square$

Proposition 1. *$(d/\sim)$ is a valid metric on $\Theta_k/\sim$.*

Proof. First, the non-negativity

$$(d/\sim)([\theta_1],[\theta_2]) \geq 0, \text{ for any } [\theta_1],[\theta_2] \in \Theta_k/\sim$$

and the symmetry

$$(d/\sim)([\theta_1],[\theta_2]) = (d/\sim)([\theta_2],[\theta_1]), \text{ for any } [\theta_1],[\theta_2] \in \Theta_k/\sim$$

are obvious due to the non-negativity and symmetry of distance d.

Next, we prove the identity of indiscernible. For any $[\theta] \in \Theta_k/\sim$, there exists a $\theta' \in \Theta_k$, and

$$0 \leq (d/\sim)([\theta],[\theta]) \leq d(\theta',\theta') = 0,$$

which implies

$$(d/\sim)([\theta],[\theta]) = 0.$$

On the other hand, if there exists $[\theta_1],[\theta_2] \in \Theta_k/\sim$ such that

$$(d/\sim)([\theta_1],[\theta_2]) = 0,$$

consider their standard quotient representations $\theta' = (\lambda'_1,...,\lambda'_k,\mu'_1,...,\mu'_k) \in [\theta_1]$ with order k_1 and $\theta'' = (\lambda''_1,...,\lambda''_k,\mu''_1,...,\mu''_k) \in [\theta_2]$ with order k_2, then by Lemma 2, we have $\min\{A,B\} = 0$. Since B is always positive, we can know that $A = \sum_{1\leq i\leq k_1} A_i = 0$ which implies $A_i = 0$ for every $1 \leq i \leq k_1$. Thus, for every $1 \leq i \leq k_1$, there exists different j_i such that $\mu'_i = \mu''_{j_i}$ and $\lambda'_i \leq \lambda''_{j_i}$. Since $\sum_{1\leq i\leq k_1} \lambda'_i = 1$, we have $1 \leq \sum_{1\leq i\leq k_1} \lambda''_{j_i} \leq 1$, which implies $\lambda'_i = \lambda''_{j_i}$ for every $1 \leq i \leq k_1$. Therefore, $\theta' \sim \theta''$ and $[\theta_1] = [\theta_2]$.

Finally, we prove the subadditivity

$$(d/\sim)([\theta_1],[\theta_2]) \leq (d/\sim)([\theta_1],[\theta_3]) + (d/\sim)([\theta_2],[\theta_3])$$

for any $[\theta_1],[\theta_2],[\theta_3] \in \Theta_k/\sim$. For any fixed $\epsilon > 0$, there exists $\theta'_{a_1} \in [\theta_1], \theta'_{b_1} \sim \theta'_{a_2},...,\theta'_{b_{m_1-1}} \sim \theta'_{a_{m_1}}, \theta'_{b_{m_1}} \in [\theta_3]$ such that

$$(d/\sim)([\theta_1],[\theta_3]) + \frac{\epsilon}{2} \geq d(\theta'_{a_1},\theta'_{b_1}) + d(\theta'_{a_2},\theta'_{b_2}) + ... + d(\theta'_{a_{m_1}},\theta'_{b_{m_1}})$$

and there exists $\theta'_{a'_1} \in [\theta_2], \theta'_{b'_1} \sim \theta'_{a'_2},...,\theta'_{b'_{m_2-1}} \sim \theta'_{a'_{m_2}}, \theta'_{b'_{m_2}} \in [\theta_3]$ such that

$$(d/\sim)([\theta_2],[\theta_3]) + \frac{\epsilon}{2} \geq d(\theta'_{a_1},\theta'_{b_1}) + d(\theta'_{a_2},\theta'_{b_2}) + ... + d(\theta'_{a_{m_2}},\theta'_{b_{m_2}}).$$

Therefore,

$$(d/\sim)([\theta_1],[\theta_2]) \leq (d/\sim)([\theta_1],[\theta_3]) + (d/\sim)([\theta_2],[\theta_3]) + \epsilon.$$

By taking $\epsilon \to 0$, we have

$$(d/\sim)([\theta_1],[\theta_2]) \leq (d/\sim)([\theta_1],[\theta_3]) + (d/\sim)([\theta_2],[\theta_3])$$

□

Lemma 3. (Θ_k, d) *is a compact metric space.*

Proof. We just need to prove that if (X_1, d_1) and (X_2, d_2) are compact, then $(X_1 \times X_2, d)$ is compact, where

$$d((x_1,x_2),(x_1',x_2')) = d_1(x_1,x_1') + d_2(x_2,x_2').$$

For any sequence $\{(x_{1,n}, x_{2,n})\}_{n\in\mathbb{N}}$ in X, since (X_1, d_1) and (X_2, d_2) are compact, they are sequentially compact, which implies that there exists a convergent subsequence $\{x_{1,a_n}\}_{n\in\mathbb{N}}$ whose limit $x_{1,a_{b_\infty}}$ is in X_1, and there exists a convergent subsequence $\{x_{2,a_{b_n}}\}_{n\in\mathbb{N}}$ whose limit $x_{2,a_{b_\infty}}$ is in X_2. For any $\epsilon > 0$ there exists $N_1, N_2 \in \mathbb{N}$ such that for any $n > N_1$,

$$d_1(x_{1,a_{b_n}}, x_{1,a_{b_\infty}}) < \frac{\epsilon}{2}$$

and for any $n > N_2$,

$$d_2(x_{2,a_{b_n}}, x_{2,a_{b_\infty}}) < \frac{\epsilon}{2}$$

which implies for any $n > \max\{N_1, N_2\}$,

$$d((x_{1,a_{b_n}}, x_{2,a_{b_n}}),(x_{1,a_{b_\infty}}, x_{2,a_{b_\infty}})) < \epsilon$$

Thus, $\{(x_{1,a_{b_n}}, x_{2,a_{b_n}})\}_{n\in\mathbb{N}}$ is a convergent subsequence whose limit $(x_{1,a_{b_\infty}}, x_{2,a_{b_\infty}})$ is in X. Therefore, (Θ_k, d) is sequentially compact. By the equivalence of sequential compactness and compactness in metric space, (Θ_k, d) is compact.

□

Lemma 4. $(\Theta_k/\sim, (d/\sim))$ *is a compact metric space.*

Proof. Let $\{O_\omega\}_{\omega\in\Omega}$ be any open cover of $\Theta_k/\sim$. For any point $[\theta] \in \Theta_k/\sim$, there exists an $\omega([\theta]) \in \Omega$ such that $[\theta] \in O_{\omega([\theta])}$, and there exists an $r([\theta]) > 0$ such that the open ball $B([\theta], r([\theta])) \subset O_{\omega([\theta])}$ centering at $[\theta]$ with radius $r([\theta])$ in $\Theta_k/\sim$. Since for any $\theta \in [\theta]$, the open ball $B(\theta, r([\theta]))$ centering at θ with radius $r([\theta])$ in Θ_k satisfies

$$B(\theta, r([\theta])) \subset \left\{\theta^* \in [\theta^*] \middle| [\theta^*] \in B([\theta], r([\theta]))\right\} \subset \left\{\theta^* \in [\theta^*] \middle| [\theta^*] \in O_{\omega([\theta])}\right\}.$$

Therefore, $\{B(\theta, r([\theta]))\}_{\theta\in\Theta_k}$ is an open cover of Θ_k. By the compactness of (Θ_k, d), there exists a finite set $U \subset \Theta_k$ such that $\{B(\theta, r([\theta]))\}_{\theta\in U}$ is a subcover of Θ_k, which implies that $\{O_{\omega(\theta)}\}_{\theta\in U}$ is a finite subcover of Θ_k. Therefore, $(\Theta_k/\sim, (d/\sim))$ is compact. □

Proposition 2. *Let $X_1, ..., X_n$ be i.i.d. sampled from the distribution $p(x; [\theta_0]) = \lambda_{1,0}f(x; \mu_{1,0}) + \lambda_{2,0}f(x; \mu_{2,0}) + ... + \lambda_{k,0}f(x; \mu_{k,0})$, where $p(x; [\theta_0]) \in \mathcal{M}_k$. Denote $[\hat{\theta}] = [(\hat{\lambda}_1, ..., \hat{\lambda}_k, \hat{\mu}_1, ..., \hat{\mu}_k)]$ as the MLE in the quotient space $\Theta_k/\sim$. If the distribution family satisfies*

(1) $f(x; \mu)$ is continuous in μ for all $\mu \in \mathcal{C}$ and all $x \in \mathcal{X}$;

(2) there exists a function $d(x)$ such that $|\log p(x; [\theta])| \leq d(x)$ for all $[\theta] \in \Theta_k/\sim$ and all $x \in \mathcal{X}$, and $\mathrm{E}_{[\theta_0]}[d(X)] < \infty$;

(3) $Q_0([\theta]) = \mathrm{E}_{[\theta_0]}[\log p(X; [\theta])]$ is uniquely maximized at $[\theta_0]$;

Then

$$[(\hat{\lambda}_1, ..., \hat{\lambda}_k, \hat{\mu}_1, ..., \hat{\mu}_k)] \xrightarrow{p} [(\lambda_{1,0}, ..., \lambda_{k,0}, \mu_{1,0}, ..., \mu_{k,0})]$$

with respect to $(d/\sim)$.

The proof of Proposition 2 is easy by the following results.

Lemma 5. *If $X_1, ..., X_n$ are i.i.d. sampled from the distribution $p(x; \theta_0) \in \{p(x; \theta)|\theta \in \Theta\}$, where Θ is compact, $\log p(x; \theta)$ is continuous in θ for all $\theta \in \Theta$ and all $x \in \mathcal{X}$, and if there exists a function $d(x)$ such that $|\log p(x; \theta)| \leq d(x)$ for all $\theta \in \Theta$ and $x \in \mathcal{X}$, and $\mathrm{E}_{\theta_0}[d(X)] < \infty$, then*

(1) $Q_0(\theta) = \mathrm{E}_{\theta_0}[\log p(X; \theta)]$ is continuous in θ;

(2) $\sup_\theta |Q(\theta; X_n) - Q_0(\theta)| \xrightarrow{p} 0$

where $Q(\theta; X_n) = \frac{1}{n}\sum_{1\leq i\leq n} \log p(X_i; \theta)$.

Lemma 6. *Suppose $Q(\theta; X_n)$ is continuous in θ and there exists a function $Q_0(\theta)$ such that*

(1) $Q_0(\theta)$ is uniquely maximized at θ_0;

(2) Θ is compact;

(3) $Q_0(\theta)$ is continuous in θ;

(4) $Q(\theta; X_n)$ converges uniformly in probability to $Q_0(\theta)$;

then $\hat{\theta}(X_n)$ defined as the value of $\theta \in \Theta$ which maximizes $Q(\theta; X_n)$ satisfies $\hat{\theta}(X_n) \xrightarrow{p} \theta_0$.

The proof of the above two lemmas can be found from Frangakis' lecture notes [26]

B Tables and Figures

Table 1: MSE of point estimation of parameters by EM algorithm

	$\mu = 0.0$				$\mu = 0.5$			
	μ_1^*	μ_2^*	λ_1^*	λ_2^*	μ_1^*	μ_2^*	λ_1^*	λ_2^*
$M = 100$	0.502	0.492	0.083	0.083	0.527	0.478	0.092	0.092
$M = 1000$	0.319	0.344	0.085	0.085	0.449	0.440	0.128	0.128
$M = 5000$	0.142	0.213	0.055	0.055	0.212	0.135	0.100	0.100

Table 1 Continued: MSE of point estimation of parameters by EM algorithm

	$\mu = 1.0$				$\mu = 1.5$			
	μ_1^*	μ_2^*	λ_1^*	λ_2^*	μ_1^*	μ_2^*	λ_1^*	λ_2^*
$M = 100$	0.563	0.517	0.104	0.104	0.310	0.198	0.052	0.052
$M = 1000$	0.109	0.040	0.035	0.035	0.029	0.008	0.005	0.005
$M = 5000$	0.019	0.005	0.008	0.008	0.006	0.002	0.001	0.001

Table 2: Power of the generalized Wald test and the likelihood ratio test when $\lambda = 0.1$

	$\mu = 0.1$	$\mu = 0.2$	$\mu = 0.3$	$\mu = 0.4$	$\mu = 0.5$	$\mu = 0.6$	$\mu = 0.7$
$\alpha = \frac{1}{5}$	0.0519	0.0565	0.0604	0.0704	0.0857	0.1035	0.1310
$\alpha = \frac{1}{4}$	0.0504	0.0551	0.0606	0.0719	0.0878	0.1052	0.1362
$\alpha = \frac{1}{3}$	0.0515	0.0555	0.0593	0.0731	0.0906	0.1088	0.1380
$\alpha = \frac{1}{2}$	0.0522	0.0560	0.0605	0.0743	0.0958	0.1132	0.1486
$\alpha = 1$	0.0507	0.0554	0.0620	0.0808	0.1063	0.1273	0.1774
$\alpha = 2$	0.0462	0.0547	0.0689	0.0908	0.1196	0.1677	0.2351
$\alpha = 3$	0.0442	0.0502	0.0694	0.0816	0.1109	0.1564	0.2010
$\alpha = 4$	0.0482	0.0521	0.0629	0.0654	0.0846	0.1045	0.1243
$\alpha = 5$	0.0476	0.0538	0.0581	0.0633	0.0771	0.0908	0.1013
LRT	0.0425	0.0495	0.0715	0.0897	0.1147	0.1597	0.2176

Table 2 Continued: Power of the generalized Wald test and the likelihood ratio test when $\lambda = 0.1$

	$\mu = 0.8$	$\mu = 0.9$	$\mu = 1.0$	$\mu = 1.1$	$\mu = 1.2$	$\mu = 1.3$	$\mu = 1.4$
$\alpha = \frac{1}{5}$	0.1644	0.1717	0.2108	0.1987	0.1965	0.1477	0.1306
$\alpha = \frac{1}{4}$	0.1675	0.1777	0.2180	0.2060	0.2045	0.1563	0.1389
$\alpha = \frac{1}{3}$	0.1736	0.1895	0.2279	0.2202	0.2247	0.1798	0.1618
$\alpha = \frac{1}{2}$	0.1840	0.2052	0.2461	0.2535	0.2624	0.2245	0.2163
$\alpha = 1$	0.2261	0.2663	0.3307	0.3761	0.4220	0.4621	0.5073
$\alpha = 2$	0.3296	0.4341	0.5568	0.7017	0.8116	0.9151	0.9640
$\alpha = 3$	0.3070	0.3922	0.5496	0.7062	0.8362	0.9294	0.9768
$\alpha = 4$	0.1673	0.2088	0.3086	0.4345	0.6026	0.7467	0.8680
$\alpha = 5$	0.1237	0.1397	0.1964	0.2430	0.3617	0.4979	0.6416
LRT	0.3283	0.4226	0.5879	0.7351	0.8580	0.9416	0.9801

Table 2 Continued: Power of the generalized Wald test and the likelihood ratio test when $\lambda = 0.1$

	$\mu = 1.5$	$\mu = 1.6$	$\mu = 1.7$	$\mu = 1.8$	$\mu = 1.9$	$\mu = 2.0$
$\alpha = \frac{1}{5}$	0.0943	0.0686	0.0406	0.0300	0.0085	0.0075
$\alpha = \frac{1}{4}$	0.1057	0.0831	0.0492	0.0378	0.0124	0.0110
$\alpha = \frac{1}{3}$	0.1281	0.1153	0.0676	0.0611	0.0241	0.0278
$\alpha = \frac{1}{2}$	0.1956	0.1923	0.1341	0.1619	0.0994	0.1239
$\alpha = 1$	0.5877	0.6886	0.7067	0.8411	0.8580	0.9356
$\alpha = 2$	0.9892	0.9965	0.9992	0.9999	0.9999	1.0000
$\alpha = 3$	0.9943	0.9987	0.9998	1.0000	1.0000	1.0000
$\alpha = 4$	0.9539	0.9834	0.9967	0.9994	1.0000	1.0000
$\alpha = 5$	0.8239	0.8992	0.9732	0.9926	0.9991	0.9999
LRT	0.9961	0.9989	0.9999	1.0000	1.0000	1.0000

Table 3: Power of the generalized Wald test and the likelihood ratio test when $\lambda = 0.3$

	$\mu = 0.1$	$\mu = 0.2$	$\mu = 0.3$	$\mu = 0.4$	$\mu = 0.5$	$\mu = 0.6$	$\mu = 0.7$
$\alpha = \frac{1}{5}$	0.0496	0.0625	0.0769	0.1050	0.1579	0.2264	0.3013
$\alpha = \frac{1}{4}$	0.0496	0.0647	0.0771	0.1072	0.1626	0.2354	0.3149
$\alpha = \frac{1}{3}$	0.0495	0.0652	0.0819	0.1111	0.1714	0.2493	0.3345
$\alpha = \frac{1}{2}$	0.0494	0.0686	0.0860	0.1140	0.1886	0.2714	0.3791
$\alpha = 1$	0.0543	0.0711	0.0907	0.1396	0.2299	0.3534	0.4903
$\alpha = 2$	0.0553	0.0745	0.1072	0.1699	0.3005	0.4887	0.6792
$\alpha = 3$	0.0523	0.0630	0.0944	0.1376	0.2313	0.4059	0.5909
$\alpha = 4$	0.0523	0.0589	0.0722	0.0891	0.1036	0.1731	0.2541
$\alpha = 5$	0.0512	0.0611	0.0683	0.0803	0.0781	0.1053	0.1076
LRT	0.0558	0.0682	0.0985	0.1530	0.2631	0.4558	0.6324

Table 3 Continued: Power of the generalized Wald test and the likelihood ratio test when $\lambda = 0.3$

	$\mu = 0.8$	$\mu = 0.9$	$\mu = 1.0$	$\mu = 1.1$	$\mu = 1.2$	$\mu = 1.3$	$\mu = 1.4$
$\alpha = \frac{1}{5}$	0.4031	0.5024	0.5877	0.6854	0.7516	0.7970	0.8604
$\alpha = \frac{1}{4}$	0.4167	0.5260	0.6096	0.7098	0.7811	0.8320	0.8941
$\alpha = \frac{1}{3}$	0.4427	0.5601	0.6548	0.7608	0.8314	0.8810	0.9399
$\alpha = \frac{1}{2}$	0.4984	0.6327	0.7347	0.8439	0.9105	0.9538	0.9837
$\alpha = 1$	0.6546	0.7985	0.9144	0.9673	0.9944	0.9989	1.0000
$\alpha = 2$	0.8587	0.9543	0.9937	0.9994	1.0000	1.0000	1.0000
$\alpha = 3$	0.7897	0.9329	0.9861	0.9989	1.0000	1.0000	1.0000
$\alpha = 4$	0.4135	0.6529	0.8424	0.9676	0.9965	0.9998	1.0000
$\alpha = 5$	0.1337	0.2428	0.4356	0.7320	0.9291	0.9883	0.9995
LRT	0.8332	0.9509	0.9916	0.9994	1.0000	1.0000	1.0000

Table 3 Continued: Power of the generalized Wald test and the likelihood ratio test when $\lambda = 0.3$

	$\mu = 1.5$	$\mu = 1.6$	$\mu = 1.7$	$\mu = 1.8$	$\mu = 1.9$	$\mu = 2.0$
$\alpha = \frac{1}{5}$	0.8845	0.9533	0.9728	0.9931	0.9925	0.9988
$\alpha = \frac{1}{4}$	0.9274	0.9743	0.9879	0.9977	0.9976	0.9998
$\alpha = \frac{1}{3}$	0.9671	0.9913	0.9971	0.9998	0.9998	0.9999
$\alpha = \frac{1}{2}$	0.9962	0.9997	1.0000	1.0000	1.0000	1.0000
$\alpha = 1$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\alpha = 2$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\alpha = 3$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\alpha = 4$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\alpha = 5$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
LRT	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 4: Power of the generalized Wald test and the likelihood ratio test when $\lambda = 0.5$

	$\mu = 0.1$	$\mu = 0.2$	$\mu = 0.3$	$\mu = 0.4$	$\mu = 0.5$	$\mu = 0.6$	$\mu = 0.7$
$\alpha = \frac{1}{5}$	0.0550	0.0678	0.0836	0.1324	0.1786	0.2538	0.3632
$\alpha = \frac{1}{4}$	0.0528	0.0690	0.0860	0.1363	0.1864	0.2691	0.3843
$\alpha = \frac{1}{3}$	0.0543	0.0711	0.0864	0.1395	0.1965	0.2953	0.4181
$\alpha = \frac{1}{2}$	0.0553	0.0719	0.0889	0.1514	0.2183	0.3364	0.4785
$\alpha = 1$	0.0566	0.0737	0.0987	0.1816	0.2706	0.4448	0.6310
$\alpha = 2$	0.0581	0.0809	0.1131	0.2269	0.3675	0.5926	0.8098
$\alpha = 3$	0.0616	0.0771	0.1046	0.1736	0.2867	0.4819	0.7120
$\alpha = 4$	0.0601	0.0630	0.0853	0.1080	0.1433	0.1942	0.3234
$\alpha = 5$	0.0578	0.0598	0.0763	0.0921	0.0932	0.0965	0.1008
LRT	0.0616	0.0814	0.1088	0.1876	0.3234	0.5383	0.7697

Table 4 Continued: Power of the generalized Wald test and the likelihood ratio test when $\lambda = 0.5$

	$\mu = 0.8$	$\mu = 0.9$	$\mu = 1.0$	$\mu = 1.1$	$\mu = 1.2$	$\mu = 1.3$	$\mu = 1.4$
$\alpha = \frac{1}{5}$	0.5022	0.6581	0.8033	0.9225	0.9730	0.9937	0.9997
$\alpha = \frac{1}{4}$	0.5353	0.6892	0.8294	0.9364	0.9807	0.9956	0.9998
$\alpha = \frac{1}{3}$	0.5767	0.7306	0.8649	0.9567	0.9892	0.9979	0.9999
$\alpha = \frac{1}{2}$	0.6478	0.7990	0.9197	0.9787	0.9961	0.9996	1.0000
$\alpha = 1$	0.8074	0.9291	0.9820	0.9980	1.0000	1.0000	1.0000
$\alpha = 2$	0.9415	0.9905	0.9988	0.9998	1.0000	1.0000	1.0000
$\alpha = 3$	0.8897	0.9785	0.9968	0.9999	0.9999	1.0000	1.0000
$\alpha = 4$	0.5211	0.7905	0.9394	0.9935	0.9996	1.0000	1.0000
$\alpha = 5$	0.1490	0.3392	0.6033	0.8642	0.9895	0.9991	1.0000
LRT	0.9202	0.9854	0.9980	0.9999	1.0000	1.0000	1.0000

Table 4 Continued: Power of the generalized Wald test and the likelihood ratio test when $\lambda = 0.5$

	$\mu = 1.5$	$\mu = 1.6$	$\mu = 1.7$	$\mu = 1.8$	$\mu = 1.9$	$\mu = 2.0$
$\alpha = \frac{1}{5}$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\alpha = \frac{1}{4}$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\alpha = \frac{1}{3}$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\alpha = \frac{1}{2}$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\alpha = 1$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\alpha = 2$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\alpha = 3$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\alpha = 4$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\alpha = 5$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
LRT	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

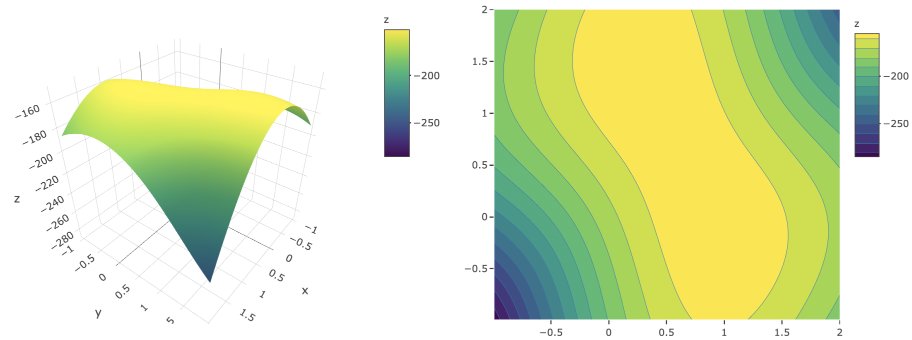

Figure 1: Log likelihood function of two-component Gaussian mixture model $\lambda N(\mu_1, 1) + (1-\lambda)N(\mu_2, 1)$. Left: x-axis and y-axis refer to μ_1 and μ_2 respectively, z-axis refers to the log likelihood, and λ is fixed as 0.3. Right: contour plot of the left 3D surface plot

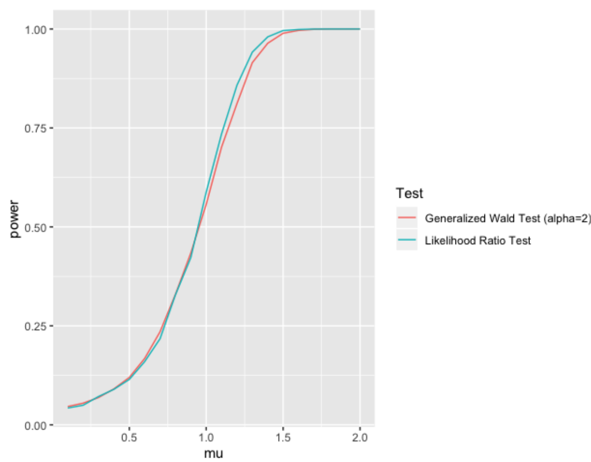

Figure 2: Power of the generalized Wald test and the likelihood ratio test when $\lambda = 0.1$

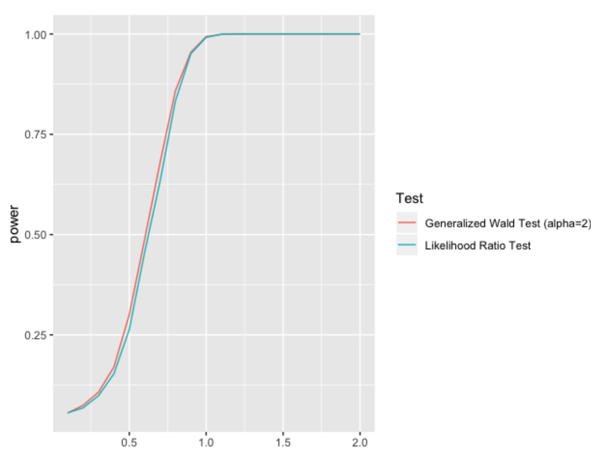

Figure 3: Power of the generalized Wald test and the likelihood ratio test when $\lambda = 0.3$

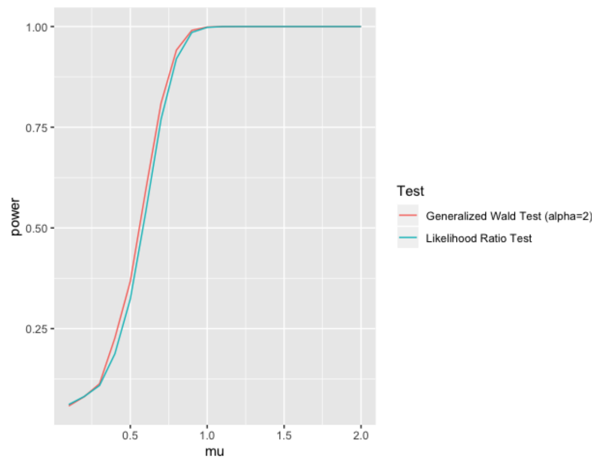

Figure 4: Power of the generalized Wald test and the likelihood ratio test when $\lambda = 0.5$