Computational Statistics and Data Analysis 127 (2018) 248–257

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

# Asymptotic null distribution of the modified likelihood ratio test for homogeneity in finite mixture models

Tony S.T. Wong [a,*], Kwok Fai Lam [b], Victoria X. Zhao [c]

[a] City University of Hong Kong, Hong Kong
[b] The University of Hong Kong, Hong Kong
[c] Harbin Institute of Technology, Shenzhen, China

ARTICLE INFO

*Article history:*
Received 12 December 2017
Received in revised form 15 May 2018
Accepted 15 May 2018
Available online 1 June 2018

*Keywords:*
Chi-bar-squared distribution
Degeneration
Generalized linear models
Linear independence
Negative definiteness

ABSTRACT

Likelihood-based methods play a central role in statistical inference for parametric models. Among these, the modified likelihood ratio test is preferred in testing for homogeneity in finite mixture models. The test statistic is related to the maximum of a quadratic function under general regularity conditions. Re-parameterization is shown to have overcome the difficulty when linear independence is not satisfied. Models with parameter constraints are also considered. The asymptotic null distribution of the test statistic is shown to have a chi-bar-squared distribution in both constrained and unconstrained cases. We extend the result to linear models and demonstrate that the chi-bar-squared distribution is also applicable. The general asymptotic result provides a much simpler testing procedure with an exact form of the asymptotic distribution compared to re-sampling approach in the literature. It also offers accurate $p$-value as shown in simulation. The results are checked by extensive simulation and are supplemented by a breast cancer data example.



## 1. Introduction

Finite mixture models offer a variety of techniques in statistical applications. They provide great flexibility in modeling data whose distribution is skewed, multi-modal or long-tailed. Statistical tests for mixture models identify clusters and verify assumptions such as normality. New applications of mixture models are growing in genetic imprinting (Li et al., 2015).

The ordinary likelihood ratio test is usually used in parametric testing problems because of the simple and usual chi-squared asymptotic null distribution. In finite mixture models, there is a nuisance parameter under the homogeneity hypothesis. The test has unusual asymptotic properties (Titterington et al., 1985; Ghosh and Sen, 1985; Bickel and Chernoff, 1993; Chernoff and Lander, 1995; Dacunha-Castelle and Gassiat, 1999; Chen and Chen, 2001) because of irregularity problems leading to inconsistent parameter estimators. When the parameter space is the set of all real numbers, the asymptotic null distribution of the test statistic involves the supremum of a Gaussian process (Ghosh and Sen, 1985; Chernoff and Lander, 1995; Chen and Chen, 2001). The Gaussian process has mean zero, variance one, and an autocorrelation function which depends on the parametric family. The distribution of its supremum is very complicated compared to the usual chi-squared distribution (Chen et al., 2001).

The modified likelihood ratio test was introduced by Chen et al. (2001) by adding a simple penalty to the mixing proportion. Consistent maximum likelihood estimator is restored in testing for homogeneity in finite mixture models. A number of interesting asymptotic null distributions is available in the literature. When the component probability density

* Corresponding author.
*E-mail address:* tonywong@cityu.edu.hk (T.S.T. Wong).

https://doi.org/10.1016/j.csda.2018.05.010
0167-9473/

function has exactly one parameter, the test statistic degenerates to zero or converges to a chi-squared distribution with one degree of freedom with equal weights (Chen et al., 2001); when the component probability density function is the two-parameter gamma distribution, the test statistic degenerates to zero or converges to a chi-squared distribution with two degrees of freedom. The weight for the degeneration depends on the parameters of the gamma distribution (Wong and Li, 2014). See also Zeng and Wong (2015) for a similar result in the beta distribution case. The above three examples do not provide clear clues about the general form of the asymptotic null distribution. Consequently, re-sampling methods (McLachlan, 1987; McLachlan and Khan, 2004) are used to circumvent the asymptotic theory. Zhu and Zhang (2004) studied mixture regression models and Niu et al. (2011) investigated multivariate mixture models, but the authors did not give the explicit form of the asymptotic null distribution. Re-sampling procedures were suggested to approximate the $p$-value. Kasahara and Shimotsu (2015) extended the EM test proposed by Li et al. (2009) to the modified EM test in normal mixture regression models. The asymptotic null distribution of the proposed test statistic is only obtained by simulation. Although different likelihood-based tests may involve different asymptotic null distributions, the general asymptotic null distribution of any one of the aforementioned tests remains mysterious. We fill this important research gap by showing that the general asymptotic null distribution of the modified likelihood ratio test is the chi-bar-squared distribution. This paper provides a clear guideline compared to the simulation approach in the literature.

In this work, a very general chi-bar-squared distribution for the asymptotic null distribution of the test statistic is obtained. Firstly, models with constraint in parameters are included. To our knowledge, investigation on hypothesis testing in mixture models subject to constraint in parameters is the first time. Besides, when linear independence is not satisfied, re-parameterization overcomes the difficulty. A stronger condition of strong identifiability introduced by Chen (1995) and imposed by Chen et al. (2001) is not essential. We demonstrate this by the normal mixture models and normal mixture linear regression models. The normal distribution is the only model we know that its elements are linearly dependent. Lastly, the theories will be extended to the generalized linear models. The paper is organized in a way that the asymptotic results are presented in Section 2. Section 3 shows the simulation results. An example of breast cancer data is demonstrated in Section 4. We conclude in Section 5. Proofs are given in a separate supplementary file entitled Proofs of Theorems and Lemmas.

## 2. Asymptotic results

### 2.1. Modified likelihood ratio test

Let $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ be a vector of independent random variables from a distribution with probability density function $f$. This work considers the test for homogeneity with hypotheses

$$H_0 : f = f(x; \boldsymbol{\theta}) \quad \text{versus} \quad H_1 : f = \pi f(x; \boldsymbol{\theta}_1) + (1-\pi) f(x; \boldsymbol{\theta}_2),$$

where $f(x; \boldsymbol{\theta})$ is the probability density function with a $d$-dimensional parameter vector $\boldsymbol{\theta}$ and $0 < \pi < 1$ is the mixing proportion. Unless otherwise specified, all parameters $\boldsymbol{\theta}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ and $\pi$ are to be estimated. Assume that the parameter space of $f$ is compact. Test for homogeneity is carried out by the following test statistic

$$LR = 2L\left(\hat{\pi}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2; \boldsymbol{X}\right) - 2L\left(\hat{\boldsymbol{\theta}}; \boldsymbol{X}\right), \tag{1}$$

where $\hat{\boldsymbol{\theta}}$ maximizes the log-likelihood function $L(\boldsymbol{\theta}; \boldsymbol{X}) = \sum_{i=1}^{n} \log f(X_i; \boldsymbol{\theta})$, and $\hat{\pi}$, $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ maximize the penalized log-likelihood function

$$L(\pi, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{X}) = \sum_{i=1}^{n} \log\{\pi f(X_i; \boldsymbol{\theta}_1) + (1-\pi) f(X_i; \boldsymbol{\theta}_2)\} + \log c \log\{4\pi(1-\pi)\}$$

and a value $c > \|\boldsymbol{\theta}_j\|$ for $j = 1, 2$ is chosen in the penalty function $\log c \log\{4\pi(1-\pi)\}$. About the penalty, Chen et al. (2001) proposed to use $c = 50$. An alternative penalty $1.5 \log(1 - |1 - 2\pi|)$ was suggested by Li et al. (2009).

Denote $|\boldsymbol{a}|$ be the absolute value applied element-wise to $\boldsymbol{a}$, $\|\boldsymbol{a}\|$ be the Euclidean norm for any vector $\boldsymbol{a}$, $\|\boldsymbol{A}\|_F$ be the Frobenius norm for any matrix $\boldsymbol{A}$, vech $(\boldsymbol{A})$ be the half-vectorization of matrix $\boldsymbol{A}$, $\boldsymbol{0}$ as the zero matrix, and $\boldsymbol{I}$ be the identity matrix. Any operation between vectors and matrices applies element-wise.

### 2.2. Conditions

Let $\boldsymbol{\theta}_0$ be the true value of $\boldsymbol{\theta}$ under $H_0$. Define the following random quantities

$$\boldsymbol{Y}_i = \frac{1}{f(X_i; \boldsymbol{\theta}_0)} \frac{\partial f(X_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}, \quad \boldsymbol{Z}_i = \frac{1}{f(X_i; \boldsymbol{\theta}_0)} \frac{\partial^2 f(X_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}, \tag{2}$$

$$V_i = \left\{\sum_{j=1}^{n} \boldsymbol{Y}_j^T\right\} \left\{\sum_{j=1}^{n} \boldsymbol{Y}_j \boldsymbol{Y}_j^T\right\}^{-1} \boldsymbol{Y}_i, \quad \boldsymbol{U}_i = \boldsymbol{Z}_i - V_i \boldsymbol{Z}_i. \tag{3}$$

Note that $\boldsymbol{Y}_i$ is a $d \times 1$ vector, $\boldsymbol{Z}_i$ is a $d \times d$ matrix and $V_i$ is scalar. Denote by $f'(X_i; \boldsymbol{\theta})$, $f''(X_i; \boldsymbol{\theta})$ and $f'''(X_i; \boldsymbol{\theta})$ as the first-order, second-order and third-order partial derivatives of $f(X_i; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.