Computational Statistics and Data Analysis 70 (2014) 127–137

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Test for homogeneity in gamma mixture models using likelihood ratio

Tony Siu Tung Wong *, Wai Keung Li

Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 24 August 2012
Received in revised form 9 August 2013
Accepted 3 September 2013
Available online 9 September 2013

Keywords:
Chi-bar-square distributions
Gamma mixture
Likelihood ratio
Maximum likelihood
Negative-definite

ABSTRACT

A testing problem of homogeneity in gamma mixture models is studied. It is found that there is a proportion of the penalized likelihood ratio test statistic that degenerates to zero. The limiting distribution of this statistic is found to be the chi-bar-square distributions. The degeneration is due to the negative-definiteness of a complicated random matrix, depending on the shape parameter under the null hypothesis. In light of this dependency, bounds on the distribution are introduced and a weighted average procedure is proposed. Simulation suggests that the results are accurate and consistent, and that the asymptotic result applies to the maximum likelihood estimator, obtained via an Expectation–Maximization algorithm.

1. Introduction

In recent years, gamma mixture models have seen a surge of applications in many fields. Craig and Strassels (2010) examined the out-of-pocket prices of analgesic medications using a two-component gamma mixture model. See also Mayrose et al. (2005) for applications in bioinformatics and the references in Liu et al. (2003). Due to their importance, developing effective and handy statistical procedures for gamma mixture models is an imperative task, in particular for the test of homogeneity. An obvious way of approaching the problem is to use the ordinary likelihood ratio test (LRT). One of the few results available is Liu et al. (2003). The authors showed that when the range of some parameters is unbounded, the LRT statistic diverges to infinity at a rate of $\log\log n$ and that its asymptotic behaviour is of extreme-value type through a highly complex piece of stochastic analysis. However, their simulation results suggested that the limiting distribution is far from converging to the extreme value distribution and that a possible solution is to simulate the finite-sample null distribution. The peculiar behaviour of the statistic arises because the maximum likelihood estimator (MLE) of some parameters may not be consistent. See, for example, the asymptotic result for $R_n\,(\varepsilon; I)$ in Chen and Chen (2001). Related problems in general mixture models were also addressed by Ghosh and Sen (1985), Dacunha-Castelle and Gassiat (1999), Chen and Chen (2001) and Liu and Shao (2003). In particular, Ghosh and Sen (1985) and Chen and Chen (2001) showed that the asymptotic distribution involves the supremum of a Gaussian process. See also Liu and Shao (2004) in normal mixture models. However, there are several shortfalls of the above results. Firstly, the results lose their appeal because the supremum of a Gaussian process is difficult to compute (Chen et al., 2001). Secondly, the divergence to infinity is so slow that it is not detected in simulation (Liu and Shao, 2004). The convergence of the test statistic, normalized by $\log\log n$, to the extreme value distribution is hardly detectable (Liu et al., 2003). Lastly, Hall and Stewart (2005) provided a theoretical analysis on the reduction of power against alternative hypotheses regarding the above results.

* Corresponding author. Tel.: +852 28592469; fax: +852 28589041.
E-mail addresses: wongtst@hotmail.com, wongtonyst@hku.hk (T.S.T. Wong), hrntlwk@hku.hk (W.K. Li).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.09.001

In light of the peculiar behaviour of LRT, a resampling approach (McLachlan, 1987; McLachlan and Peel, 2000; McLachlan and Khan, 2004) can be carried out. However, when some of the regularity conditions are restored, especially consistency of the estimator, it is of great theoretical significance to further investigate the likelihood ratio.

The consistency of the MLE in the test for homogeneity has not been solved until the introduction of a clever penalized procedure proposed by Chen et al. (2001). The authors innovated the modified likelihood ratio test (MLRT) by incorporating a penalty function. The MLRT was also developed by Chen and Kalbfleisch (2005) in normal mixture models and further extended to an EM-test by Li et al. (2009) and Chen and Li (2009). Exact theoretical results on the asymptotic null distribution have been obtained in some special cases. For densities with a single parameter of interest, the MLRT statistic has the limiting distribution $0.5\chi_0^2 + 0.5\chi_1^2$ (Chen et al., 2001; Li et al., 2009). For the normal mixture model, the statistic has χ_2^2 when the means and the variances are unequal and unknown (Chen and Li, 2009). Conceivably, the MLRT falls into the type II likelihood ratio problem (Lindsay, 1995, Section 4.4) which generates the chi-bar-square distributions of which some are parameter-dependent limiting null distributions. The above result in the normal mixture models returns to the χ_2^2 distribution due to loss of strong identifiability (Chen and Li, 2009, Example 1). Qin and Smith (2006) investigated an extension of the MLRT in multivariate normal mixture models. The authors showed the asymptotic null distribution being a mixture of distributions and suggested it must be found using numerical methods. For models with multidimensional parameters, Zhu and Zhang (2000) analysed the asymptotic properties of LRT and Niu et al. (2011) considered an EM test. Although the problem of estimator consistency has been solved in MLRT and the EM-test, in many other mixture models, such as the gamma mixture models, the results $0.5\chi_0^2 + 0.5\chi_1^2$ or χ_2^2 cannot be applied directly without theoretical justifications. The general testing problem has not been completely solved and remains as a long-standing open problem. Charnigo and Sun (2004) acknowledged the generalization of the MLRT to higher dimensional problems and suggested that the null distribution can be obtained by simulation. However, the extension is not at all straightforward as presented in this paper and simulation of the null distribution in the absence of a closed-form expression should no longer be tolerated. A clear guideline has been long overdue for practitioners in the rejection or retention of the homogeneity assumption. The purposes of the paper are to fill this research gap in gamma mixture models and to explore how the limiting null distribution depends on the parameters.

Motivated by the above needs and the importance of the gamma mixture models, this paper aims at investigating the limiting distribution of the MLRT statistic. We obtain the condition under which the MLRT statistic degenerates to zero and determine the proportion of degeneration. Then, we can show that the asymptotic null distribution has parameter-dependent chi-bar-square distributions. This subsequently establishes a foundation for quick model selection using the χ_2^2 distribution in practice. Moreover, in light of the popular Expectation–Maximization (EM) algorithm for parameter estimation in finite mixture models, we demonstrate through intensive simulation studies that our results can be applied to the likelihood ratio statistic evaluated at the MLE obtained via the EM algorithm.

The article is organized as follows. In Section 2, we present the asymptotic results. Section 3 lists a number of considerations in the applications of the results. The asymptotic analysis is supplemented by simulation in Section 4. Section 5 presents two data examples and Section 6 gives a conclusion.

2. Asymptotic results

We consider a two-parameter gamma density function

$$f(x; \alpha, \beta) = \frac{1}{\Gamma(\alpha)} \beta^{\alpha} x^{\alpha-1} e^{-\beta x}, \quad x > 0,$$

where $\alpha > 1$ and $\beta > 0$ are shape and scale parameters, respectively. Given a set of independent and identically distributed data, we are interested in testing the homogeneity hypothesis H_0 against the alternative hypothesis of a two-component gamma mixture model H_1 where

$$H_0 : f(x) = f(x; \alpha, \beta);$$
$$H_1 : f(x) = \pi f(x; \alpha_1, \beta_1) + (1-\pi) f(x; \alpha_2, \beta_2),$$

and $0 < \pi < 1$ is a mixing proportion. In this paper, we study a very general testing problem that the parameters under the hypotheses are all unknown and unequal. This is completely different from the setting in Liu et al. (2003). For parametric hypothesis testing problems it is customary to use the ordinary LRT based on the statistic which is defined as

$$LR_n = 2\left\{L\left(\hat{\pi}, \hat{\alpha}_1, \hat{\beta}_1, \hat{\alpha}_2, \hat{\beta}_2\right) - L\left(0.5, \hat{\alpha}, \hat{\beta}, \hat{\alpha}, \hat{\beta}\right)\right\},$$

where

$$L(\pi, \alpha_1, \beta_1, \alpha_2, \beta_2) = \sum_{i=1}^{n} \log\left\{\pi f(x_i; \alpha_1, \beta_1) + (1-\pi) f(x_i; \alpha_2, \beta_2)\right\} \quad (1)$$

is the log-likelihood function and $\left(\hat{\pi}, \hat{\alpha}, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}, \hat{\beta}_1, \hat{\beta}_2\right)$ is the MLE of parameter $(\pi, \alpha, \alpha_1, \alpha_2, \beta, \beta_1, \beta_2)$. It is well known that the consistency of the MLE, obtained by maximizing (1) directly, is not guaranteed. See for example Ghosh and Sen