

Model Selection for Mixture Models Using Perfect Sample

Sadegh Fallahigilan[†], Abdolreza Sayyareh^{*,‡}

[†] Razi University

[‡] Toosi University of Technology

Received: 6/17/2018 Approved: 7/20/2019

Abstract. We have considered a perfect sample method for model selection of finite mixture models with either known (fixed) or unknown number of components which can be applied in the most general setting with assumptions on the relation between the rival models and the true distribution. It is, both, one or neither to be well-specified or mis-specified, they may be nested or non-nested. We consider mixture distribution as a complete-data (bivariate) distribution by prediction of missing data variable (unobserved variable) and show that this ideas is applicable to use Vuong's test for select optimum mixture model when number of components are known (fixed) or unknown. We have considered *AIC* and *BIC* based on the complete-data distribution. The performance of this method is evaluated by Monte-Carlo method and real data set, as Total Energy Production.

Keywords. finite mixture model; perfect sample; model selection; missing data variable; Vuong's test

MSC 2010: 62F03.

* Corresponding author

1 Introduction

The mixture models are one of the most important statistical models, and also the most technically challenging. See, Everitt and Hand (1981); Titterington et al. (1985) have provided the properties of minimum-distance estimators for mixture models, especially for estimation of the weights, McLachlan and Basford (1988); McLachlan and Peel (2004) introduced a lot of algorithms that have been published to fit mixture models, and Redner and Walker (1984), have extended the regular assumptions in Wald (1948) for well-specified mixture model using some assumptions. Model selection is an important step in most empirical work and, accordingly, there exists a vast literature devoted to this issue. Model selection for mixture model can be view in two situations. First is about determining the number of components by hypothesis testing when two rival mixture models are nested and second is selecting optimum model between two rival non-nested mixture models with equal number of components under which the rival models are well-specified or mis-specified. In many practical situations, the number of components in a mixture distribution is unknown. Determining the number of components in a mixture distribution is an important but difficult problem. It is well known that the likelihood ratio statistic for testing the number of components in a mixture model fails to have the classical chi-squared distribution, since under the null hypothesis, the mixing proportions lie on the boundary of the parameter space and the parameters are not identifiable under the model of null hypothesis. Ghosh and Sen (1985) obtained the limiting distribution under a separation condition. The separation condition turned out to be unnecessary, which was shown by Chernoff and Lander (1995) for binomial mixtures, and in general by Chen et al. (2001); Dacunha-Castelle and Gassiat (1999) and others. Chen et al. (2001); Chen and Kalbfleisch (2005) developed a modified likelihood approach which they showed to be simple to use and to have good performance in many situations. Feng and McCulloch (1994) performed a simulation study to demonstrate their claim that the simulated null distribution of the likelihood ratio statistic for testing a single normal against a mixture of two normal distributions depends on the choice of restrictions imposed on the component variances. Lo et al. (2001) extended the work of Vuong (1989) by applying the Vuong's test to allow comparisons between rival models with different numbers of components. They showed that the likelihood ratio statistic is asymptotically distributed as a weighted sum of independent chi-squared random variables with one de-

gree of freedom. It is important to note that Jeffries (2003) pointed out that the conditions for the Lo-Mendell-Rubin (LMR) likelihood ratio test Theorem are not generally satisfied in the mixture modeling context when the parameters are in boundary of parameter space. When the null hypothesis of LMR is true, the parameters of the additional component hypothesized under the alternative hypothesis does not exist. Therefore, they don't have a unique maximum in parameter space, which is a violation of one of the assumptions presented in Lo et al. (2001). Despite this flaw, LMR has been widely used and has been shown to be effective in recovering the number of underlying components, Nylund et al. (2008); Morgan (2015); Morgan et al. (2016). Lo (2005) performed a simulation study to normal mixture with unequal variances and he confirmed the results of Lo et al. (2001). Also, Sayyareh (2016) considers a finite mixture of the known criterion to the model selection problem to answer to the question, how could infinite set of all possible models that could have given rise to data, be narrowed down to a reasonable set of statistical models. He has proposed two types of coefficients for the mixture criterion, one based on the density and another one based on the risk function. Fallahigilan and Sayyareh (2016) have selected model for mixture distribution with the Vuong's test, AIC and BIC, based on simulation study. Wichitchan et al. (2018) apply the idea of goodness of fit (GOF) testing procedure to finite mixture models and investigate their performance when performing hypothesis testing of mixture models for different types of alternative hypotheses.

The aim of the current paper is to give a method to model selection in mixture issue using Vuong's statistic and model selection criteria. In Section 2, the idea is to altering mixture model distribution and its development it to a complete-data (bivariate) distribution. In section 3 we propose a test procedure based on an extension of the Theorem which have introduced by Vuong (1989). We extend the Vuong (1989) has likelihood ratio test for model selection, *AIC* and *BIC*. We use the Vuong's test to show when the rival models are nested, based on certain regularity assumptions, the limiting distribution of the likelihood ratio statistic is a weighted sum of independent χ^2 random variables, and a normal distribution when the rival models are non-nested. These results do not require that either of the rival models be well-specified. Section 4 consider a discussion on this approach in more details. In particular, using the results of Sections 2 and 3, we shall obtain very simple tests and information criteria for selecting among two rival models. Numerical results,compression between our method with LMR

likelihood ratio test and Kolmogorov-Smirnov test based on Wichitchan et al. (2018) and also data analysis for Total Energy Production data set are presented in Section 5. There is a discussion about total energy production for 10 top Total Energy Production Rankings, in United State of America. Our basic verification show that this data has a finite mixture model. So we extend the known statistical test to find an optimum model between some nested, non-nested and mis-specified rival models.

2 Development of the Mixture Models

Finite mixture distributions have been used widely in the modern statistics. Consider a population made up of m subgroups, mixed at random in proportion to the relative group sizes τ_1, \dots, τ_m . Assume that interest lies in some random feature X which is heterogeneous across and homogeneous within the subgroups. Due to heterogeneity, X has a different probability distribution in each group, usually assumed to arise from the same parametric family $h(x|\phi)$, however, with the parameter ϕ differing across the groups. The groups may be labelled through a discrete indicator variable Z taking values in the set $1, \dots, m$. When sampling randomly from such a population, we may record not only X , but also the group indicator Z . The probability of sampling from the group labelled Z is equal to τ_j , whereas conditional on knowing Z , X is a random variable following the distribution $h(x|\phi_j)$ with ϕ_j being the parameter in group Z . The joint density $h(x, z)$ is given by

$$h(x, z) = h(x|z)h(z).$$

A finite mixture distribution arises if it is not possible to record the group indicator Z ; what we observe is only the random variable X .

Let $X = (X_1, \dots, X_n)^T \in \mathfrak{R}^k$ denote a random vector with size n , where X_i is a k -dimensional random vector and is independent and identically distributed random vector with mixture distribution function $H(x|\Upsilon) = \int_{-\infty}^x h(t|\Upsilon)dt$ where

$$h(x_i|\Upsilon) = \sum_{j=1}^m \tau_j h_j(x_i|\phi_j), x_i \in \mathfrak{R}^k, \phi_j \in \Phi_j \subseteq \mathfrak{R}^{p_j}; j = 1, \dots, m \quad (1)$$

is probability density and m is the number of components of the mixture and each τ_j is non-negative quantities that sums to one. The quantities τ_1, \dots, τ_m

are called the mix proportions or weights. The function of any components,

$$h_1(x|\phi_1), \dots, h_m(x|\phi_m)$$

are the same parametric density functions by $\phi_j \in \Phi_j \subseteq \mathbb{R}^{p_j}$ where ϕ_j is the vector of unknown parameters in the postulated form for the j th component and p_j is the number of parameters in j 'th component. So we consider (1) as definition of m -component mixture (incomplete-data) density function with parameter space

$$\Delta = \{(\tau_1, \dots, \tau_m, \phi_1, \dots, \phi_m) : \sum_{j=1}^m \tau_j = 1, \tau_j \geq 0, \phi_j \in \Phi_j \text{ for } j = 1, \dots, m\}$$

and $\Upsilon = \{(\tau_1, \dots, \tau_{m-1}, \phi_1, \dots, \phi_m)\} = \{\xi_1, \dots, \xi_{\ell_0}\}$; $\Upsilon \in \Delta \subseteq \mathbb{R}^{\ell_0}$ where Υ is the vector of all unknown parameters in the mixture model and ℓ_0 is the dimension of Υ that is $\ell_0 = m - 1 + \sum_{j=1}^m p_j$.

It is well known that the mixture model can be seen as an incomplete-data structure model, where the complete-data is given by

$$(X, Z) = \{(X_1, Z_1), \dots, (X_n, Z_n)\}$$

and Z_i represent the missing data variable which can be thought of as the component label of the feature vector X_i . It is convenient to work with a m -dimensional-label vector Z_i in place of the single categorical variable Z_i , where the j th element of Z_i , Z_{ij} , is defined to be one or zero, according to whether the component of origin of X_i in the mixture is equal to j or not ($j = 1, \dots, m$). Thus Z_i is distributed according to a multinomial distribution consisting of one draw on m categories with probabilities τ_1, \dots, τ_m ; that is,

$$P(Z_i = z_i) = \tau_1^{z_{i1}} \dots \tau_m^{z_{im}}$$

or

$$Z_i \sim \text{Multinomial}(1, \tau_1, \dots, \tau_m).$$

Then one can write the mixture density in the complete-data form as,

$$h(x_i, z_i|\Upsilon) = \prod_{j=1}^m (\tau_j h_j(x_i|\phi_j))^{z_{ij}}. \quad (2)$$

According to Z , we have found an alternative method for mixture models based on a complete-data distribution of X and Z . So, we could write a mixture model in a fully categorized, complete-data observation (X, Z) , with a complete-data density function (2) in the family of $H_{\Upsilon} = \{h(x, z|\Upsilon) : \Upsilon \in \Delta\}$. The first advantage of this model is that we could write a closed and linear form for the log-likelihood function. The second advantage is about the identifiability of mixture models. In a family of finite mixture distributions $H'_{\Upsilon} = \{h(x|\Upsilon) : \Upsilon \in \Delta\}$ one has to distinguish among three types of non-identifiability. Non-identifiability due to invariance to relabeling the components of the mixture distribution, Redner and Walker (1984), and non-identifiability due to potential overfitting, Crawford (1994). The last type of non-identifiability is a generic property of a certain class of mixture distributions (like finite mixtures of uniform distributions), Teicher (1961). In the following, we will avoid the last type of models.

According to Yakowitz and Spragins (1968); Titterton et al. (1985) we cannot use the identifiability definition for a mixture model. We assume that all weights are positive, $0 < \tau_1 < \dots < \tau_m < 1$ and for each mixture the component parameters are distinct in the weak sense defined, $\phi_1 < \dots < \phi_m$ (this strong constraint rules out many interesting mixtures if ϕ_j contained all elements of parameters. e.g. in the two-component normal mixture these restrictions are $0 < \tau_1 < \tau_2 < 1$ and $\mu_1 < \mu_2$ or $0 < \tau_1 < \tau_2 < 1$ and $\sigma_1 < \sigma_2$). So, we modified these conditions for arbitrary elements of parameters. Then these assumptions ensure that $h(x|\Upsilon_1) = h(x|\Upsilon_2)$ implies $\Upsilon_1 = \Upsilon_2$.

The third advantage is that this model contains more information about the unknown parameters than observed data. Now, let σ_Y be the σ -finite measure on Y . The vector Y denotes a random vector which is partitioned to $Y = (X, Z)$. Let (X, σ_X) and (Z, σ_Z) be the measurable spaces associated with X and Z . Let H_Y^0 be the true distribution of Y . Based on regularity assumptions given in Vuong (1989) as Assumptions 1-6. We have explained some of these Assumptions as follows.

Assumption 1. The random vector Y is independent and identically distributed, *i.i.d.*, with common true underlying distribution H_Y^0 on (Y, σ_Y) with measurable Radon-Nikodym density $h(\cdot)$.

Now we consider two rival parametric families of mixture distributions with complete-data form: $F_{\Gamma} = \{f(y|\Gamma); \Gamma \in \Omega \in \mathfrak{R}^{\ell}\}$ and $G_{\Psi} = \{g(y|\Psi); \Psi \in \Lambda \in \mathfrak{R}^j\}$. These rival models can be nested or non-nested. Also, both, only one or neither can be well-specified. Each rival model must be satisfied

Vuong's (1989) Assumptions 2 to 5 that stated in terms of F_Γ . It is clear that similar Assumptions are made on G_Ψ .

According Vuong's Assumptions ensures the existence of the matrices:

$$\begin{aligned} A_f(\Gamma_*) &= \langle E_h(\nabla_\Gamma^2 \log f(Y|\Gamma_*)) \rangle \\ A_g(\Psi_*) &= \langle E_h(\nabla_\Psi^2 \log g(Y|\Psi_*)) \rangle \\ B_f(\Gamma_*) &= \langle E_h([\nabla_\Gamma(\log f(Y|\Gamma_*))][\nabla_\Gamma(\log f(Y|\Gamma_*))]^T) \rangle \\ B_g(\Psi_*) &= \langle E_h([\nabla_\Psi(\log g(Y|\Psi_*))][\nabla_\Psi(\log g(Y|\Psi_*))]^T) \rangle \\ B_{fg}(\Gamma_*, \Psi_*) &= B'_{gf}(\Psi_*, \Gamma_*) = \langle E_h([\nabla_\Gamma(\log f(Y|\Gamma_*))][\nabla_\Psi(\log g(Y|\Psi_*))] \rangle \end{aligned}$$

The expectation under $h(\dots)$, E_h , denote the expectation with respect to the true distribution of Y and the value Γ_* is called pseudo-true value of Γ for rival model F_Γ . Similarly, Ψ_* denotes the pseudo-true value of Ψ for G_Ψ .

2.1 Estimation of Missing Data Variable

The *EM* algorithm is a broadly applicable algorithm that provides an iterative procedure for computing *MLE*'s in situations where, but for the absence of some additional data, *MLE* would be straightforward. The *EM* algorithm approaches the problem of solving the incomplete-data log likelihood equation indirectly by proceeding iterative in terms of the complete-data log likelihood function.

One can write the log-likelihood

$$\begin{aligned} \log f(x|\Gamma) &= \log f(y|\Gamma) - \log \kappa_f(z_f|x; \Gamma) \Rightarrow \\ \log L(\Gamma) &= \log L_c(\Gamma) - \log \kappa_f(z_f|x; \Gamma), \end{aligned} \quad (3)$$

where $\log L(\Gamma)$ and $\log L_c(\Gamma)$ are incomplete-data and complete-data log likelihood functions. On taking the expectations of both sides of (3) with respect to the conditional distribution $\kappa_f(z_f|x; \Gamma^{(k)})$ using the fit $\Gamma^{(k)}$ for Γ , we have that

$$\begin{aligned} \log L(\Gamma) &= E_{\kappa_f(z_f|x; \Gamma^{(k)})}(\log L_c(\Gamma)|x) - E_{\kappa_f(z_f|x; \Gamma^{(k)})}(\log \kappa_f(Z_f|x; \Gamma)) \Rightarrow \\ &= Q(\Gamma; \Gamma^{(k)}) - H(\Gamma; \Gamma^{(k)}). \end{aligned} \quad (4)$$

More specifically, let $\Gamma^{(0)}$ be some initial value for Γ . Then we have for estimating parameters using *EM* algorithm as;

- E*-step: Calculation $Q(\Gamma; \Gamma^{(k)})$.
- M*-step: Choose $\Gamma^{(k+1)}$ to be any value of $\Gamma \in \Omega$ that maximizes $Q(\Gamma; \Gamma^{(k)})$.

The *E*- and *M*-steps are alternated repeatedly until the difference

$$L(\Gamma^{(k+1)}) - L(\Gamma^{(k)})$$

by an arbitrarily small amount in the case of convergence of the sequence of likelihood values $\{L(\Gamma^{(k)})\}$ where $k = 0, 1, \dots$. In reality, Z_f is unknown but yet we could find a prediction of the missing data variable Z_f . Kazakos (1977), Scott and Symons (1971) and Symons (1981) introduced many ways for predicate Z_f . We have selected the predictor of Z_f in last iteration of *EM* algorithm in terms of mean squared error. As McLachlan and Peel (2004) point out, the corresponding a classification distribution is

$$E_{\kappa_f(z_f|x;\Gamma)}(Z_{jf}|x) = P(Z_{jf}|x;\Gamma) = \frac{\alpha_j f(x|\theta_j)}{f(x|\Gamma)}. \quad (5)$$

After parameter estimation with *EM* algorithm, the predicted classifications are given by (5) in the last iterative. For convenience, write

$$\tilde{z}_{jf} = E_{\kappa_f(\tilde{z}_f|x;\hat{\Gamma}_n)}(Z_{jf}|x) = \frac{\hat{\alpha}_j f(x|\hat{\theta}_j)}{f(x|\hat{\Gamma}_n)}, \quad (6)$$

for $j = 1, \dots, m$. In some applications it is desirable to harden a posterior classifications and the most popular way to do this is to report maximum a posterior (MAP) classifications, i.e

$$MAP(\tilde{z}_{jf}) = \begin{cases} 1 & j = \underset{h}{\operatorname{argmax}}\{\tilde{z}_{hf}\} \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

3 Model Selection

In this section, based on Vuong (1989), we obtain the asymptotic distribution of the complete-data likelihood ratio (CLR) statistic under Vuong's Assumptions 1 to 5 and also we extend *AIC*, *BIC* for complete-data form of mixture distribution.

In the other hand, one can write *E*-step in the last iterative based on

previous section

$$\begin{aligned}\log L(\Gamma) &= E_{\kappa_f(\tilde{z}_f|x;\hat{\Gamma}_n)}(\log f(X, Z_f|\Gamma)|x) \Rightarrow \\ \log L(\hat{\Gamma}_n) &= E_{\kappa_f(\tilde{z}_f|x;\hat{\Gamma}_n)}(\log f(X, Z_f|\hat{\Gamma}_n)|x).\end{aligned}\quad (8)$$

Lemma 1. *Given Vuong's Assumptions 1 to 3 and Equation (8);*

$$\frac{1}{n} \sum_{i=1}^n E_{\kappa_f(\tilde{z}_{if}|x_i;\hat{\Gamma}_n)}(\log f(X_i, Z_{if}|\hat{\Gamma}_n)|x) \xrightarrow{a.s.} E_h(\log f(X, Z_{*f}|\Gamma_*)) \quad (9)$$

where z_{*f} is pseudo true value of z_f .

Similarly for G_Ψ

$$\begin{aligned}\log L(\Psi) &= E_{\kappa_g(z_g|x;\Psi^{(k)})}(\log L_c(\Psi)|x) - E_{\kappa_g(z_g|x;\Psi^k)}(\log \kappa_g(Z_g|x; \Psi)) \\ &= Q(\Gamma; \Gamma^{(k)}) - H(\Gamma; \Gamma^{(k)})\end{aligned}\quad (10)$$

and in the last iterative of E -step

$$\begin{aligned}\log L(\Psi) &= E_{\kappa_g(z_g|x;\hat{\Psi}_n)}(\log g(X, Z_g|\Psi)|x) \Rightarrow \\ \log L(\hat{\Psi}_n) &= E_{\kappa_g(\tilde{z}_g|x;\hat{\Psi}_n)}(\log g(X, Z_g|\hat{\Psi}_n)|x),\end{aligned}\quad (11)$$

where $\tilde{z}_{jg} = E_{\kappa_g(\tilde{z}_g|x;\hat{\Psi}_n)}(Z_{jg}|x) = \frac{\hat{\tau}_j g(x|\hat{\phi}_j)}{g(x|\hat{\Psi}_n)}$.

Corollary 1. *Given \tilde{z}_{jf} and \tilde{z}_{jg} we have;*

$$E_{\kappa_f(\tilde{z}_f|x;\hat{\Gamma}_n)}(\log f(X, Z_f|\hat{\Gamma}_n)|x) = \log f(x, \tilde{z}_f|\hat{\Gamma}_n) \quad (12)$$

$$E_{\kappa_g(\tilde{z}_g|x;\hat{\Psi}_n)}(\log g(X, Z_g|\hat{\Psi}_n)|x) = \log g(x, \tilde{z}_g|\hat{\Psi}_n). \quad (13)$$

Lemma 2. *Given Vuong's Assumptions 1 to 3 and Equations (8), (11), (12)*

and (13)

$$\begin{aligned} \frac{1}{n}CLR(\hat{\Gamma}_n, \hat{\Psi}_n) &= \frac{1}{n} \sum_{i=1}^n \{E_{\kappa_f(\tilde{z}_{if}|x_i; \hat{\Gamma}_n)}(\log f(X_i, Z_{if}|\hat{\Gamma}_n)|x) \\ &- E_{\kappa_g(\tilde{z}_{ig}|x_i; \hat{\Psi}_n)}(\log g(X_i, Z_{ig}|\hat{\Psi}_n)|x)\} = \frac{1}{n} \sum_{i=1}^n \{\log f(x_i, \tilde{z}_{if}|\hat{\Gamma}_n) \\ &- \log g(x_i, \tilde{z}_{ig}|\hat{\Psi}_n)\} \xrightarrow{a.s.} E_h\left(\log \frac{f(X, Z_{*f}|\Gamma_*)}{g(X, Z_{*g}|\Psi_*)}\right). \end{aligned} \quad (14)$$

Given all assumptions in Redner and Walker (1984), Vuong's Assumptions 1 to 5 and Lemma 1 and 2 we could supposed $\lim_{k \rightarrow \infty} \Gamma^{(k)} = \hat{\Gamma}_n$ and $\hat{\Gamma}_n \xrightarrow{a.s.} \Gamma_*$ as $n \rightarrow \infty$. Also, the *MLE* is consistent for Γ_* , and is asymptotically normally distributed with asymptotic covariance matrix $A_f^{-1}(\Gamma_*)B_f(\Gamma_*)A_f^{-1}(\Gamma_*)$. Similar properties hold for the *MLE* $\hat{\Psi}_n$ of Ψ_* . As a matter of fact, $\hat{\Gamma}_n$ and $\hat{\Psi}_n$ are jointly asymptotically normal with asymptotic covariance matrix that can be consistently estimated using the sample analogs of $A_l(\cdot)$, $B_l(\cdot)$ and $B_{fg}(\Gamma_*, \Psi_*)$ where $l = f, g$ are evaluated at $(\hat{\Gamma}_n, \hat{\Psi}_n)$, Vuong (1989).

Based on Vuong (1989), we obtain the asymptotic distribution of the complete-data likelihood ratio (CLR) statistic under null hypothesis $H_0^{fg} : E_h\{\log f(X, Z_{*f}|\Gamma_*)\} = E_h\{\log g(X, Z_{*g}|\Psi_*)\}$, that verify two family of mixture models are equivalent. against,

$$H_1^f : E_h\{\log f(X, Z_{*f}|\Gamma_*)\} > E_h\{\log g(X, Z_{*g}|\Psi_*)\},$$

which means that the rival model F_Γ is better than G_Ψ , or

$$H_1^g : E_h\{\log f(X, Z_{*f}|\Gamma_*)\} < E_h\{\log g(X, Z_{*g}|\Psi_*)\},$$

which means that the rival model G_Ψ is better than F_Γ .

The Vuong's Assumption 1 to 6 ensure that variance of $\log \frac{f(y|\Gamma_*)}{g(y|\Psi_*)}$ exists. Following, we restate Theorem 3.3 from Vuong (1989), based on the complete-data form of mixture density. Consider $Y = (X, Z_l)$ where $l = f, g$ having complete-data density. One difference between Vuong's test and this work is that the marginal density is parametric density. Based on Vuong (1989) we use following Theorem.

Theorem 1. Under Vuong's Assumptions 1 to 5

(a) If two rival models are nested, then

$$2CLR(\hat{\Gamma}_n, \hat{\Psi}_n) \xrightarrow{D} M_{\ell+j}(\cdot; \lambda_*) \quad (15)$$

where $M_{\ell+j}(\cdot; \lambda_*)$ is the weighted sum of Chi-square distribution and λ_* is the vector of $\ell + j$ eigenvalues of

$$W = \begin{bmatrix} -B_f(\Gamma_*)A_f^{-1}(\Gamma_*) & -B_{fg}(\Gamma_*, \Psi_*)A_g^{-1}(\Psi_*) \\ -B_{gf}(\Psi_*, \Gamma_*)A_f^{-1}(\Gamma_*) & -B_g(\Psi_*)A_g^{-1}(\Psi_*). \end{bmatrix} \quad (16)$$

(b) If two rival models are non-nested and Vuong's Assumption 6 holds, then

$$n^{-\frac{1}{2}}CLR(\hat{\Gamma}_n, \hat{\Psi}_n) - n^{\frac{1}{2}}E_h[\log \frac{f(Y|\Gamma_*)}{g(Y|\Psi_*)}] \xrightarrow{D} N(0, \omega_*^2) \quad (17)$$

where $\omega_*^2 = Var_h(\log \frac{f(Y|\Gamma_*)}{g(Y|\Psi_*)}) = Var_h(\log \frac{f(X, Z_{*f}|\Gamma_*)}{g(X, Z_{*g}|\Psi_*)})$.

In practice, λ_* can consistently estimated by $\hat{\lambda}_n$, that is estimated by replacing the matrices (16) by sample averages evaluated at $\hat{\Gamma}_n$ and $\hat{\Psi}_n$; for example

$$A_f^n(\hat{\Gamma}_n) = \langle n^{-1} \sum_{i=1}^n \nabla_{\Gamma}^2 \log f(x_i, \tilde{z}_{if}|\Gamma) |_{\Gamma=\hat{\Gamma}_n} \rangle \xrightarrow{a.s.} A_f(\Gamma_*)$$

and so on.

Lemma 3. Given Vuong's Assumptions 1 to 3 and 6 and Corollary 1

$$\hat{\omega}_n^2 \xrightarrow{a.s.} \omega_*^2.$$

So, $\hat{\omega}_n^2$ is a strongly consistent estimator for ω_*^2 , that is

$$\hat{\omega}_n^2 = \frac{1}{n} \sum_{i=1}^n [\log \frac{f(x_i, \tilde{z}_{if}|\hat{\Gamma}_n)}{g(x_i, \tilde{z}_{ig}|\hat{\Psi}_n)}]^2 - [\frac{1}{n} \sum_{i=1}^n \log \frac{f(x_i, \tilde{z}_{if}|\hat{\Gamma}_n)}{g(x_i, \tilde{z}_{ig}|\hat{\Psi}_n)}]^2.$$

On the other hand, one of the basic tools to model selection is class of

information criteria. Akaike (1973) has defined his information criterion as;

$$AIC = -2 \log\text{-likelihood function} + 2(\text{number of parameters}).$$

In this context, the Akaike information criterion for mixture and complete-data distributions are respectively;

$$AIC = -2 \sum_{i=1}^n \log \sum_{j=1}^m \hat{\alpha}_j f(x_i | \hat{\theta}_j) + 2\ell$$

and

$$AIC_b = -2 \sum_{i=1}^n \sum_{j=1}^m \log \tilde{z}_{ijf} (\log \hat{\alpha}_j + \log f(x_i | \hat{\theta}_j)) + 2(\ell + 1)$$

Schwarz (1978) introduces his information criterion as a competitor to the *AIC*;

$$BIC = -2 \log\text{-likelihood function} + (\text{number of parameters})(\log n).$$

Similarly, the Schwarz information criterion for mixture and complete-data distributions are respectively;

$$BIC = -2 \sum_{i=1}^n \log \sum_{j=1}^m \hat{\alpha}_j f(x_i | \hat{\theta}_j) + \ell \log(n).$$

and

$$BIC_b = -2 \sum_{i=1}^n \sum_{j=1}^m \log \tilde{z}_{ijf} (\log \hat{\alpha}_j + \log f(x_i | \hat{\theta}_j)) + (\ell + 1) \log(n)$$

Obviously, because of the prediction of variable Z , we add the number of parameters ℓ in the *AIC_b* and *BIC_b*.

4 Discuss in More Detail about Vuong's Test

In Section 2, it is suggested using complete-data form of mixture models. In this section, we shall discuss this approach in more detail. In particular, using the results of Sections 2 and 3, we will extend Vuong's test for selecting among

two rival models whether they are nested (for testing number of component) or non-nested when both, only one or neither can be well-specified.

4.1 Nested Models: Testing Number of Components

In this subsection, we consider the case where the models F_Γ and G_Ψ are nested and may or may not contain the true underlying distribution. In this case, we have supposed all the component distributions in two rival models, F_Γ and G_Ψ are from the same parametric family. Furthermore, we assume that the parameter m is unknown. We postulate that the first rival model, G_Ψ have m_0 -component and second rival model, F_Γ have m_1 -component, where m_0 and m_1 are known constants with $m_0 < m_1$. Note that a m_0 -component mixture distribution is nested within a m_1 -component mixture distribution when two rival models come from the same parametric family. So, we can use Theorem 1(a) for this situation.

Assumption 2. There exists a function $\nu(\cdot)$ from Λ to $\Omega \in \mathfrak{R}^\ell$ such that, for almost all y , $g(y|\Psi) = f(y|\nu(\Psi))$ for Ψ in $\Lambda \in \mathfrak{R}^j$.

Assumption 2 states that any complete-data mixture density $g(y|\Psi)$ is also a complete-data mixture density $f(y|\Gamma)$ for some Γ in Ω . Since $\nu(\Psi)$ is included in Ω then $g(y|\Psi)$ is nested in $f(y|\Gamma)$. As a matter of fact, the alternative to the null hypothesis H_0^{fg} is H_1^f inasmuch H_1^g can never occur because the rival model $g(y|\Psi)$ can never be better than $f(y|\Gamma)$. Hence, our testing reduces to the smaller model is equivalent to or worse than the larger model, but in this situation when two rival models are equivalent we select the smaller model because of the number of parameters.

Theorem 2. *Given Vuong's Assumptions 1 to 5 and Assumption 1, under null hypothesis:*

(a) *If F_Γ is mis-specified*

$$2CLR(\hat{\Gamma}_n, \hat{\Psi}_n) \xrightarrow{D} M_{\ell+j}(\cdot; \hat{\lambda}_n), \quad (18)$$

(b) *If rival models are well-specified*

$$2CLR(\hat{\Gamma}_n, \hat{\Psi}_n) \xrightarrow{D} \chi_{\ell-j}^2(\cdot). \quad (19)$$

It is carried out by choosing a critical value from $M_{\ell+j}(\cdot; \hat{\lambda}_n)$ or $\chi_{\ell-j}^2(\cdot)$ when rival models are mis-specified or well-specified respectively and by rejecting the hypothesis that the models are equivalent if twice CLR statistic is greater than this critical value for choosing H_1^f in favour of H_0^{fg} .

4.2 Non-nested Models

In this subsection, we consider the case where the models F_Γ and G_Ψ are non-nested and may or may not contain the true underlying distribution. This situation can accrue when we know the number of components are equal and specified but the rival models are from the non-nested parametric families. Since the complete-data models F_Γ and G_Ψ do not need to more assumptions. So based on Theorem 1(b);

Theorem 3. *Given Vuong's Assumptions 1 to 5 and null hypothesis, if two rival models are non-nested, then*

$$\frac{n^{-\frac{1}{2}} CLR(\hat{\Gamma}_n, \hat{\Psi}_n)}{\hat{\omega}_n} \xrightarrow{D} N(0, 1). \quad (20)$$

Theorem 3 provide very simple tests for mixture model selection. Since, we can chosen a critical value c from the standard normal distribution. If the value of statistic (20) is higher than c then we rejects the null hypothesis and we say rival model F_Γ is better than rival model G_Ψ . Similarly, if it is smaller than $-c$ then we say rival model G_Ψ is better than rival model F_Γ .

5 Simulaion Study

In the remaining, we decide whether two rival mixture models are equivalent or which of them is optimum to estimate the true model. Simulation studies were conducted to investigate the finite sample properties of the test. The maximum likelihood estimates of the parameters and prediction of missing data variable were obtained by the EM algorithm. In this section, we do simulation study to show that, Vuong's test could select the optimum mixture model when mixture of density functions write in complete-data form. Also, we try to show different between AIC_b , BIC_b with AIC , BIC . We consider the data generating probabilities as mixtures of normal (MN_t), mixtures of log-normal (MLN_t), mixtures of Weibull (MW_t), mixture of Gamma (MG_t)

Table 1. Different situations for generating data set

True model	D_1	D_2	D_3
$M2N_t$	(0.3,0,1,1,1)	(0.3,0,1,2,1)	(0.3,0,1,3,1)
$M2E_t$	(0.3,1,4)	(0.3,0.5,7)	$(0.3, \frac{1}{5}, 5)$
$M3N_t$	(0.2,0.4,0,1,1,1,2,1)	(0.3,0.4,0,1,1,1,3,1)	(0.3,0.4,0,1,3,1,6,1)
$M4N_t$	(0.3,0.3,0.2,0,1,1,2,1,5,1)	(0.3,0.3,0.2,0,1,2,1,7,1,12,1)	(0.3,0.35,0.2,0,1,3,1,8,1,12,1)
$M4E_t$	$(0.3,0.3,0.2, \frac{1}{8}, 1, 4, 8)$	$(0.3,0.3,0.2, \frac{1}{6}, 0.5, 2, 6)$	$(0.3,0.3,0.2, \frac{1}{20}, \frac{1}{13}, \frac{1}{6}, 1)$
$M4G_t$	(0.2,0.3,0.3,10,20,13,20,17, 20,43,20)	(0.2,0.3,0.3,10,20,15,20,30, 20,60,20)	(0.2,0.3,0.3,10,20,20,20,40,20, 80,20)
$M4W_t$	(0.2,0.3,0.2,5,8,12,8,18,8,40, 8)	(0.2,0.3,0.2,5,0.8,12,6,18,6, 43,9)	(0.2,0.3,0.2,5,0.8,12,4,18,6, 43,9)

Note: $M2N_t$: mixture of two normal, $M2E_t$: mixture of two exponential, $M3N_t$: mixture of three normal, $M4N_t$: mixture of four normal, $M4E_t$: mixture of four exponential, $M3G_t$: mixture of three Gamma, $M4G_t$: mixture of four Gamma and $M4W_t$: mixture of four Weibull

distributions. We generate 10000 Monte-Carlo data sets of sample sizes $n = 50, 80, 200, 500$.

5.1 Simulations Study for Nested Models: Testing Number of Components

In this subsection, we do simulation study to show that how can select the mixture model with optimum component when the true underlying model is well-specified or mis-specified. For more competition, we calculated non-parametric test as Kolmogorov-Smirnov test, KS -test, AIC , BIC , AIC_b and BIC_b criteria. We used 12 configurations, based on four sample sizes, 50, 80, 200 and 500, and three values of D_i , $i = 1, 2, 3$ where D_i generate the different mixture model for each of true underlying models. Table 2 shows D_i for each underlying true mixture models.

Table 2 shows the simulated significance levels for Vuong's test and KS -test, at $\alpha = 0.05$ for 12 configurations, based on four sample sizes, 50, 80, 200 and 500, and three values of D_i , $i = 1, 2, 3$. In row 1, we have consid-

ered random sample which drawn from the two-component normal mixture distributions, $m = 2$ with $\alpha_1 < \alpha_2$, $\mu_1 < \mu_2$ and $\sigma_1 = \sigma_2$ in three values D_1, D_2 and D_3 . For example, the data is generated from parameters D_1 , Table 2. Although, data come from a mixture model with two-component but the density plot is look like normal, Figure 1, *a*. In D_2 and D_3 , the difference between mean of components has been increased, Figures 1, *b* and *c*, respectively. Therefore, we choose two competing models, mixture of two normal $M2N$ and mixture of three normal $M3N$, based on Figures 1, *a*, *b* and *c*. As we can see for any sample size, Under p-value of Vuong's test, we accept null hypothesis when the two-components are not well separated, that is D_1 or D_2 and the two-components, D_3 , are well separated. So, we say two rival models are equivalent but we select two-component normal mixture model as optimum model because of number of parameters. Next, in the similarly way, we considered random sample which has been drawn from the two-component exponential mixture distributions in three values D_1, D_2 and D_3 . So, we choose two competing models, mixture of two exponential $M2E$ and mixture of three exponential $M3E$. Under p-value of Vuong's test we accept null hypothesis when the two-components are not well separated, that is D_1 or D_2 and the two-components are well separated, D_3 , for any sample size. We next considered random sample which has been drawn from the three-component normal mixture distributions, $m = 3$ with $\alpha_1 < \alpha_2 < \alpha_3$, $\mu_1 < \mu_2 < \mu_3$ and $\sigma_1 = \sigma_2 = \sigma_3$ in three values D_1, D_2 and D_3 , as true underlying models. Based on Figures 2 *d*, *e* and *f*, we choose two competing models, mixture of two normal $M2N$ and mixture of three normal $M3N$. Here, Under p-values of Vuong's test we reject null hypothesis and select the mixture of three normal distributions at the 5 percent levels of significance respect to Theorem 2 when the three-components are not well separated, that is D_1 or D_2 and the three-components are well separated, D_3 for any sample size expect D_1 with $n = 50$, see Figure 1 *d*, *e* and *f*. Similarly, in rows 4 to 7, Vuong's test select the rival models that are equal to the true underlying mixture model according number of components. Furthermore, in row 7, the true mixture of four Weibull distribution is mis-specified. Comparisons of the actual level to the nominal level for each sample size indicate that the *KS*-test can not select the optimum models and almostly, all models are equivalent for any sample size. Furthermore, we have calculated AIC_b , BIC_b and AIC , BIC information criteria for more competitions, Tables 3 and 4 respectively. In rows 1 and 2, according true distributions and rival models and three value D_i $i = 1, 2, 3$, AIC and BIC criteria select the rival

model $M2N$ and $M2E$, as optimum models. In this situation, the results of AIC and BIC criteria equal to AIC_b and BIC_b . As we see in the Figure 1 and according to the true model, $M2N_t$, the shapes of the true models have one mode in each situations D_1 , D_2 and it has two modes for D_3 . As we have seen, the information criteria have selected the model with less components, $M2N$ and $M2E$. In other rows, Table 6, 7 the AIC and BIC are different from the AIC_b , BIC_b and the model selection test. In row 3, although the true model is three-components normal distribution, $M3N_t$, but the shapes of true model are one mode and two mode, for D_1 and D_2 respectively, see Figure 1, d and e . We see that the model selection test and AIC_b , BIC_b select the rival model, $M3N$, as optimum model. Tables 2 and 3, while AIC and BIC select the rival model, $M2N$. For D_3 , the true model has three modes, see Figure 1, d . In this situation, the results of AIC and BIC are equal to AIC_b , BIC_b and model selection test. Similarly, in rows 4 to 7, AIC and BIC have the same behavior. Furthermore, in row 7, the rival models are mis-specified according the true underlying model. In this situation AIC , BIC and AIC_b , BIC_b select the same rival model as $M4G$. So, we can say, when the true underlying model is well separated (D_3) the AIC , BIC , AIC_b , BIC_b and the model selection test make the same decision. In the other situations, D_1 , D_2 , the model selection test, AIC_b and BIC_b select the correct rival model than the AIC and BIC . We can see from Table 5 that the power of the test depends on the difference between component local parameters and sample sizes. Furthermore, we can see the test power increases uniformly as the difference between two local parameters increases (D_3) and is low when the components are not well separated (D_1). Indeed, the power of test are increasing when we reject null hypothesis, sample size increase and components are well separated.

5.2 Simulations Study for Non-nested Models

In previous subsection, we select the mixture model with optimum component when two rival models are nested. The second case will be worse when two rival models are non-nested. In this subsection, our goal is to pay attention to this problem. Table 6 summarizes the situation under which both rival mixture models have equal numbers of component and they are non-nested. In row 1, we considered four random sample which has drawn from the two-component normal mixture distribution with size 50, 80, 200 and 500; the mixing proportion α was set to 0.3 and $\mu_1, \mu_2, \sigma_1 = \sigma_2$ were set

Table 2. p-value of *KS*-test and Vuong's test.

True model	Rival model	test	50			80		
			D_1	D_2	D_3	D_1	D_2	D_3
$M2N_t$	$M3N$	KS-test	0.5487	0.9667	0.8693	0.8219	0.4383	0.8219
		Vuong's test	0.9686	0.9767	0.9780	0.9894	0.9895	0.9896
	$M2N$	KS-test	0.8693	0.5487	0.3959	0.4383	0.6953	0.6953
$M2E_t$	$M3E$	KS-test	0.7166	0.8693	0.7235	0.9794	0.7103	0.7501
		Vuong's test	1	1	1	1	1	1
	$M2E$	KS-test	0.7166	0.6221	0.4512	0.3307	0.6412	0.5231
$M3N_t$	$M3N$	KS-test	0.5487	0.8693	0.9667	0.922	0.8219	0.9794
		Vuong's test	1	0.0090	0.0097	0.0003	0.0001	0.0049
	$M2N$	KS-test	0.7166	0.9667	0.9667	0.9794	0.922	0.2424
$M4N_t$	$M4N$	KS-test	0.7166	0.7166	0.8693	0.922	0.9979	0.9794
		Vuong's test	0.0012	0.0006	0.0000	0.0035	0.0004	0.0000
	$M3N$	KS-test	0.8693	0.9667	0.9765	0.5625	0.8219	0.9752
$M4E_t$	$M4E$	KS-test	0.8693	0.5487	0.8695	0.6953	0.6215	0.6953
		Vuong's test	0.0062	0.0037	0.0001	0.0051	0.0000	0.0000
	$M3E$	KS-test	0.5487	0.8693	0.8658	0.4383	0.5871	0.2424
$M4G_t$	$M4G$	KS-test	0.9977	0.9977	0.3959	0.5625	0.922	0.3307
		Vuong's test	0.0002	0.0007	0.0122	0.0001	0.0091	.0033
	$M3G$	KS-test	0.3959	0.2719	0.8693	0.9220	0.6953	0.8219
$M4W_t$	$M4G$	KS-test	0.1124	0.8693	0.8693	0.9794	0.5625	0.8219
		Vuong's test	0.0027	0.0057	0.0062	0.0082	0.0490	0.0006
	$M3G$	KS-test	0.3959	0.1786	0.2719	0.922	0.8219	0.2424

Table 2. (continued)

True model	Rival model	test	200			500		
			D_1	D_2	D_3	D_1	D_2	D_3
$M2N_t$	$M3N$	KS-test	0.5945	0.843	0.5945	0.9600	0.7699	0.7184
		Vuong's test	0.9896	0.9991	0.9998	1	1	1
	$M2N$	KS-test	0.1565	0.9078	0.955	0.4595	0.3291	0.6654
$M2E_t$	$M3E$	KS-test	0.8211	0.8123	0.7631	0.9231	0.8610	0.7801
		Vuong's test	1	1	1	1	1	1
	$M2E$	KS-test	0.6102	0.7551	0.6201	0.7567	0.8212	0.6651
$M3N_t$	$M3N$	KS-test	0.1565	0.9078	0.955	0.7699	0.8632	0.9895
		Vuong's test	0.0001	0.0001	0.0219	0.0000	0.0000	0.0000
	$M2N$	KS-test	0.5107	0.7654	0.3613	0.8186	0.3696	0.7184
$M4N_t$	$M4N$	KS-test	0.3613	0.4324	0.9078	0.7699	0.6654	0.6654
		Vuong's test	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$M3N$	KS-test	0.6808	0.4324	0.3613	0.7184	0.1114	0.5085
$M4E_t$	$M4E$	KS-test	0.8643	0.7920	0.5441	0.6654	0.9959	0.6787
		Vuong's test	0.0044	0.0000	0.0000	0.0001	0.0000	0.0000
	$M3E$	KS-test	0.9228	0.2203	0.3275	0.5085	0.1294	0.7184
$M4G_t$	$M4G$	KS-test	0.5107	0.4324	0.9078	0.9987	0.8186	0.9347
		Vuong's test	0.0004	0.0007	0.0045	0.0000	0.0000	0.0000
	$M3G$	KS-test	0.0956	0.9550	0.4324	0.4131	0.08152	0.05873
$M4W_t$	$M4G$	KS-test	0.5107	0.8430	0.9550	0.4131	0.9600	0.7184
		Vuong's test	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	$M3G$	KS-test	0.5945	0.1565	0.1235	0.5085	0.9022	0.8544

Table 3. AIC_b and BIC_b information criteria based on complete-data variable

True model	Rival model	Information criterion	50			80		
			D_1	D_2	D_3	D_1	D_2	D_3
$M2N_t$	$M3N$	AIC_b	181.7636	193.3893	211.3481	287.8234	298.2834	330.2251
		BIC_b	197.0598	210.2381	224.8134	294.9810	327.7613	351.8778
	$M2N$	AIC_b	161.7345*	182.3258*	200.6571*	269.0923*	283.4005*	315.8756*
		BIC_b	182.1231*	190.8921*	210.4854*	271.5675*	301.5634*	340.8709*
$M2E_t$	$M3E$	AIC_b	70.5807	110.5993	129.3718	119.8848	161.4404	212.4416
		BIC_b	97.0123	112.0715	140.8750	134.1770	175.7336	226.7338
	$M2E$	AIC_b	58.2392*	60.5417*	120.9908*	95.8570*	110.9453*	203.7081*
		BIC_b	87.3512*	68.1898*	128.6389*	105.3852*	120.4734*	213.2362*
$M3N_t$	$M3N$	AIC_b	199.5634	201.5487*	250.8976*	304.5105*	312.8596*	390.9809*
		BIC_b	222.7452	223.1582*	271.7845*	323.5667*	329.2506*	399.8909*
	$M2N$	AIC_b	195.8989*	212.9808	260.9120	312.4523	334.0978	400.7809
		BIC_b	209.9097*	231.9089	287.0909	333.8987	356.8767	415.9812
$M4N_t$	$M4N$	AIC_b	230.7612*	240.1254*	251.3454*	445.6787*	450.6787*	489.7898*
		BIC_b	235.7843*	261.6534*	272.5675*	460.7898*	478.6787*	490.7898*
	$M3N$	AIC_b	239.6781	252.1248	267.3487	454.7676	462.7865	493.7612
		BIC_b	242.8798	271.5647	287.9886	469.6752	485.8909	500.1233
$M4E_t$	$M4E$	AIC_b	169.4921*	171.4113*	317.1173*	363.5249*	494.771*	512.0806*
		BIC_b	181.7882*	186.7075*	332.4135*	382.5811*	501.1201*	531.1368*
	$M3E$	AIC_b	171.8554	177.5833	324.2900	389.0058	499.4510	559.3427
		BIC_b	183.3275	189.0554	335.7622	402.2980	510.5567	573.6348
$M4G_t$	$M4G$	AIC_b	70.5674*	111.4563*	151.6743*	176.6575*	233.6574*	254.6574*
		BIC_b	66.6785*	119.7612*	176.6758*	210.8982*	261.4563*	282.6574*
	$M3G$	AIC_b	80.8977	121.7898	163.1239	184.7667	241.8098	269.9091
		BIC_b	84.8976	129.8798	183.7898	220.9696	274.1243	299.9231
$M4W_t$	$M4G$	AIC_b	163.8774*	175.7779*	181.5589*	265.0966*	271.8798*	290.8977*
		BIC_b	186.5665*	188.0008*	193.5676*	288.8881*	297.9008*	300.1112*
	$M3G$	AIC_b	176.9778	185.9088	195.9080	355.9094	358.0911	360.0871
		BIC_b	190.8932	205.9090	208.7676	364.1347	365.7850	378.0980

Note: * shows the select models.

Table 3. (continued)

True model	Rival model	Information criterion	200			500		
			D_1	D_2	D_3	D_1	D_2	D_3
$M2N_t$	$M3N$	AIC_b	711.7823	845.9098	858.8934	1634.1231	1780.7823	2100.1712
		BIC_b	719.7834	782.8912	861.8893	1658.8910	1801.8931	2158.9812
	$M2N$	AIC_b	701.3461*	823.0128*	825.8129*	1599.8912*	1771.8923*	1999.1273*
		BIC_b	703.7833*	776.3421*	841.7123*	1610.8983*	1782.7831*	2101.7822*
$M2E_t$	$M3E$	AIC_b	185.3385	233.2544	414.6072	400.3018	515.0554	733.5318
		BIC_b	205.1783	253.3658	434.3974	417.1620	540.3430	757.8118
	$M2E$	AIC_b	147.9722*	232.2145*	412.8297*	372.5810*	380.5277*	674.3376*
		BIC_b	161.1655*	245.7665*	426.0224*	310.5289*	388.8861*	691.1960*
$M3N_t$	$M3N$	AIC_b	840.8794*	910.4629*	1023.8933*	1945.8922*	1999.0192*	2080.8983*
		BIC_b	867.8941*	1002.8922*	1059.9022*	1959.0192*	2041.8291*	2371.9883*
	$M2N$	AIC_b	860.1389	989.9884	1100.9883	1989.2223	2020.8292	2110.1728
		BIC_b	910.8943	1042.2983	1110.8911	2000.1232	2092.9829	2450.3491
$M4N_t$	$M4N$	AIC_b	1002.7887*	1032.4568*	1101.9988*	2001.9093*	2080.9112*	2129.0091*
		BIC_b	1015.8784*	1047.8895*	1125.8895*	2025.3483*	2110.3450*	2140.0122*
	$M3N$	AIC_b	1081.1923	1121.8011	1180.8222	2091.8112	2118.7999	2160.1110
		BIC_b	1099.1922	1140.8849	1201.8811	2102.0091	2180.9988	2204.5886
$M4E_t$	$M4E$	AIC_b	673.4519*	733.1732*	1280.9720*	1590.9670*	1859.3920*	3016.9860*
		BIC_b	699.8384*	759.5598*	1307.3590*	1624.6845*	1893.1090*	3050.7030*
	$M3E$	AIC_b	680.2928	778.8488	1384.4470	1680.2810	2015.3360	3340.9160
		BIC_b	700.0827	798.6383	1404.2370	1715.5695	2040.6134	3366.2040
$M4G_t$	$M4G$	AIC_b	343.3806*	659.8998*	901.8894*	1101.8119*	1210.2110*	1382.8211*
		BIC_b	365.4128*	739.8888*	920.9983*	1158.8894*	1221.2883*	1402.8221*
	$M3G$	AIC_b	381.8819	710.1811	961.8812	1131.0112	1250.5534	1405.0081
		BIC_b	399.1110	761.1223	984.1210	1210.0911	1289.0111	1480.1120
$M4W_t$	$M4G$	AIC_b	645.8572*	658.8837*	668.9991*	1551.1128*	1567.8223*	1581.1282*
		BIC_b	681.0122*	688.9222*	701.8811*	1575.9891*	1578.6547*	1580.2321*
	$M3G$	AIC_b	830.1233	856.0911	880.1920	1977.8912	1989.8731	2031.8911
		BIC_b	855.2331	877.1110	909.0131	2010.1110	2031.4223	2051.0991

Note: * shows the select models.

Table 4. *AIC* and *BIC* information criteria

True model	Rival model	Information criterion	50			80		
			D_1	D_2	D_3	D_1	D_2	D_3
$M2N_t$	$M3N$	<i>AIC</i>	159.3868	173.5643	198.8734	262.2376	282.2352	306.5204
		<i>BIC</i>	174.6838	195.7831	214.1696	271.0726	301.2914	324.5766
	$M2N$	<i>AIC</i>	155.9500*	172.4763*	196.1822*	260.0726*	278.5705*	305.3156*
		<i>BIC</i>	165.5101*	186.2371*	205.7423*	260.9828*	290.4807*	318.2258*
$M2E_t$	$M3E$	<i>AIC</i>	32.6897	173.5643	198.8734	57.7254	282.2352	306.5204
		<i>BIC</i>	174.6838	195.7831	214.1696	69.6356	301.2914	324.5766
	$M2E$	<i>AIC</i>	26.8869*	172.4763*	196.1822*	51.78455*	278.5705*	305.3156*
		<i>BIC</i>	65.8873*	186.2371*	205.7423*	56.5486*	290.4807*	318.2258*
$M3N_t$	$M3N$	<i>AIC</i>	187.6528	201.5487	230.5847*	304.5105	312.8596	371.0245*
		<i>BIC</i>	202.9490	223.1582	246.8421*	323.5667	329.2506	386.4101*
	$M2N$	<i>AIC</i>	185.8852*	199.3485*	236.1284	299.9416*	310.1944*	374.5000
		<i>BIC</i>	195.4454*	220.6125*	254.3651	311.8518*	324.7698*	390.0807
$M4N_t$	$M4N$	<i>AIC</i>	225.3819	233.4283	242.1284*	433.1281	444.3952*	471.8251*
		<i>BIC</i>	230.2482	252.6306	269.3218*	455.6281	470.3825*	485.8240*
	$M3N$	<i>AIC</i>	220.8012*	219.8677*	252.5491	423.0240*	460.5010	479.6038
		<i>BIC</i>	226.3527*	245.9239*	279.5270	442.0012*	475.3069	492.3875
$M4E_t$	$M4E$	<i>AIC</i>	159.8805	170.2253	334.4294	277.1564	452.2310	548.4550
		<i>BIC</i>	173.2647	186.6094	347.8136	294.6306	422.6719	565.1292
	$M3E$	<i>AIC</i>	156.4845*	169.4060*	330.6524*	274.5718*	453.245*	544.4864*
		<i>BIC</i>	166.0446*	178.6961*	340.2125*	286.4819*	415.0519*	556.3965*
$M4G_t$	$M4G$	<i>AIC</i>	50.3806	90.0947	134.6775*	168.7139	213.0945	229.0890*
		<i>BIC</i>	65.4128	100.8149	155.7099*	193.8460	238.9456	255.2912*
	$M3G$	<i>AIC</i>	32.4336*	88.0248*	141.6136	151.5092*	206.9051*	251.6605
		<i>BIC</i>	53.7298*	93.6491*	156.9098	184.8031*	227.8361*	270.7167
$M4W_t$	$M4G$	<i>AIC</i>	163.9204*	165.2389*	166.1110*	247.3486*	250.0946*	259.7270*
		<i>BIC</i>	184.9526*	185.0912*	187.1433*	273.5509*	280.5498*	285.9293*
	$M3G$	<i>AIC</i>	196.0858	210.6743	219.4907	345.1415	346.3481	348.2700
		<i>BIC</i>	211.3820	223.8015	234.7869	364.1977	375.1092	383.3264

Note: * shows the select models.

Table 4. (continued)

True model	Rival model	Information criterion	200			500		
			D_1	D_2	D_3	D_1	D_2	D_3
$M2N_t$	$M3N$	<i>AIC</i>	624.6085	767.9033	769.0981	1498.2150	1725.2860	1986.1584
		<i>BIC</i>	650.9951	699.0816	795.4846	1531.9320	1759.0030	1947.6871
	$M2N$	<i>AIC</i>	619.3370*	716.9855*	763.4512*	1493.7790*	1713.4900*	1974.1587*
		<i>BIC</i>	635.8285*	683.1862*	779.9428*	1514.8520*	1734.5690*	1929.1758*
$M2E_t$	$M3E$	<i>AIC</i>	87.4421	767.9033	769.0981	189.4558	1725.2860	1986.1584
		<i>BIC</i>	103.9338	699.0816	795.4846	210.5289	1759.0030	1947.6871
	$M2E$	<i>AIC</i>	81.5124*	716.9855*	763.4512*	183.4539*	1713.4900*	1974.1587*
		<i>BIC</i>	88.1090*	683.1862*	779.9428*	191.8831*	1734.5690*	1929.1758*
$M3N_t$	$M3N$	<i>AIC</i>	730.2270	886.3458	937.2738*	1629.2514	1701.1000	1796.1282*
		<i>BIC</i>	756.6134	912.1452	957.3823*	1658.1520	1742.5128	1801.0025*
	$M2N$	<i>AIC</i>	697.1364*	875.5240*	940.8907	1621.8380*	1692.5419*	1802.2000
		<i>BIC</i>	713.6280*	902.3650*	963.6603	1642.6110*	1731.2561*	1815.0096
$M4N_t$	$M4N$	<i>AIC</i>	825.3819	833.4283	742.1284*	1833.1281	1844.3952*	1871.8251*
		<i>BIC</i>	830.2482	852.6306	769.3218*	1855.6281	1870.3825*	1885.8240*
	$M3N$	<i>AIC</i>	720.8012*	719.8677*	854.5491	1423.0240*	1860.5010	1879.6038
		<i>BIC</i>	726.3527*	745.9239*	875.5270	1842.0012*	1875.3069	1892.3875
$M4E_t$	$M4E$	<i>AIC</i>	595.8842	693.1851	1367.8790	1474.0090	1726.6490	3330.7320
		<i>BIC</i>	618.9724	716.2733	1390.9670	1503.5060	1756.1510	3360.2430
	$M3E$	<i>AIC</i>	593.5536*	689.8887*	1364.1120*	1472.9670*	1723.1480*	3337.0060*
		<i>BIC</i>	610.0505*	706.3803*	1380.6030*	1494.0900*	1744.2210*	3348.5790*
$M4G_t$	$M4G$	<i>AIC</i>	343.3806	585.8901	802.8193*	972.3012	1101.8010	1228.4370*
		<i>BIC</i>	365.4128	647.8903	816.0923*	949.6619	1116.0128	1274.7970*
	$M3G$	<i>AIC</i>	332.4336*	572.9851*	813.9289	960.1420*	1051.9273*	1395.7460
		<i>BIC</i>	353.7298*	628.1749*	889.8230	930.8609*	1080.8173*	1429.4630
$M4W_t$	$M4G$	<i>AIC</i>	640.2799*	653.8098*	663.5338*	1547.6260*	1564.7801*	1579.6650*
		<i>BIC</i>	676.5614*	682.8532*	699.8153*	1573.9870*	1574.1204*	1575.7380*
	$M3G$	<i>AIC</i>	826.2116	853.8902	875.6029	1972.9450	1986.9025	2011.3690
		<i>BIC</i>	852.5981	872.8395	901.9894	2006.6670	2026.3098	2045.0850

Note: * shows the select models.

Table 5. Power of Vuong's test

Sample size	Different situations	α	$M2N_t$	$M2E_t$	$M3N_t$	$M4N_t$	$M4E_t$	$M4G_t$	$M4W_t$
50	D_1	0.01	0.1694	0.1756	0.1832	0.7861	0.7825	0.7531	0.7547
		0.05	0.2372	0.2896	0.2569	0.7925	0.7999	0.7546	0.7658
		0.10	0.2711	0.3201	0.3011	0.8014	0.8023	0.7890	0.7891
	D_2	0.01	0.1523	0.1725	0.7631	0.7516	0.8601	0.8131	0.8201
		0.05	0.2261	0.2745	0.7768	0.7798	0.8721	0.8473	0.8402
		0.10	0.2631	0.2864	0.7899	0.8051	0.8950	0.8248	0.8654
	D_3	0.01	0.1379	0.1562	0.8512	0.8624	0.9021	0.8976	0.9096
		0.05	0.2586	0.2678	0.8690	0.8644	0.9154	0.9051	0.9167
		0.10	0.2951	0.3025	0.8899	0.8983	0.9498	0.9342	0.9378
80	D_1	0.01	0.1732	0.1968	0.7231	0.7391	0.7652	0.8001	0.7720
		0.05	0.2391	0.2548	0.7571	0.7452	0.7898	0.7741	0.7891
		0.10	0.2821	0.3047	0.7686	0.7496	0.7966	0.7939	0.8061
	D_2	0.01	0.1651	0.1968	0.7701	0.7631	0.8064	0.8259	0.8169
		0.05	0.2112	0.2354	0.7886	0.7959	0.8458	0.8361	0.8429
		0.10	0.2772	0.3021	0.8090	0.8135	0.8507	0.8501	0.8586
	D_3	0.01	0.1412	0.1602	0.7539	0.9672	0.9698	0.8981	0.9074
		0.05	0.2601	0.3012	0.7651	0.9732	0.9899	0.9136	0.9208
		0.10	0.3061	0.3201	0.7948	0.9801	0.9915	0.9351	0.9312
200	D_1	0.01	0.1821	0.2031	0.8459	0.7540	0.7698	0.7839	0.7736
		0.05	0.2457	0.2632	0.8642	0.7590	0.7782	0.7940	0.7879
		0.10	0.3039	0.3214	0.8967	0.8021	0.8325	0.8039	0.8140
	D_2	0.01	0.1769	0.1870	0.9349	0.8539	0.8741	0.8353	0.8369
		0.05	0.2323	0.2504	0.9561	0.8851	0.8999	0.8447	0.8518
		0.10	0.2911	0.3102	0.9710	0.9824	0.9901	0.9095	0.9041
	D_3	0.01	0.1475	0.1608	0.8112	0.9836	0.9840	0.9252	0.9368
		0.05	0.2786	0.3087	0.8396	0.8905	0.9877	0.9331	0.9460
		0.10	0.3275	0.3501	0.8501	0.9836	0.9898	0.9460	0.9553
500	D_1	0.01	0.2032	0.2402	0.8061	0.8071	0.8366	0.8052	0.8096
		0.05	0.2611	0.3088	0.8520	0.8766	0.8442	0.8596	0.8499
		0.10	0.3301	0.3677	0.8970	0.8496	0.8677	0.8731	0.8728
	D_2	0.01	0.1964	0.2278	0.9016	0.8642	0.8971	0.9057	0.9236
		0.05	0.2516	0.2891	0.8520	0.8938	0.9401	0.9531	0.9749
		0.10	0.3291	0.3587	0.8970	0.9536	0.9748	0.9747	0.9874
	D_3	0.01	0.1630	0.2079	0.9016	0.9931	0.9931	0.9870	0.9888
		0.05	0.3056	0.3478	0.9256	0.9973	0.9975	0.9889	0.9893
		0.10	0.3451	0.3877	0.9590	1	1	0.9931	0.9991

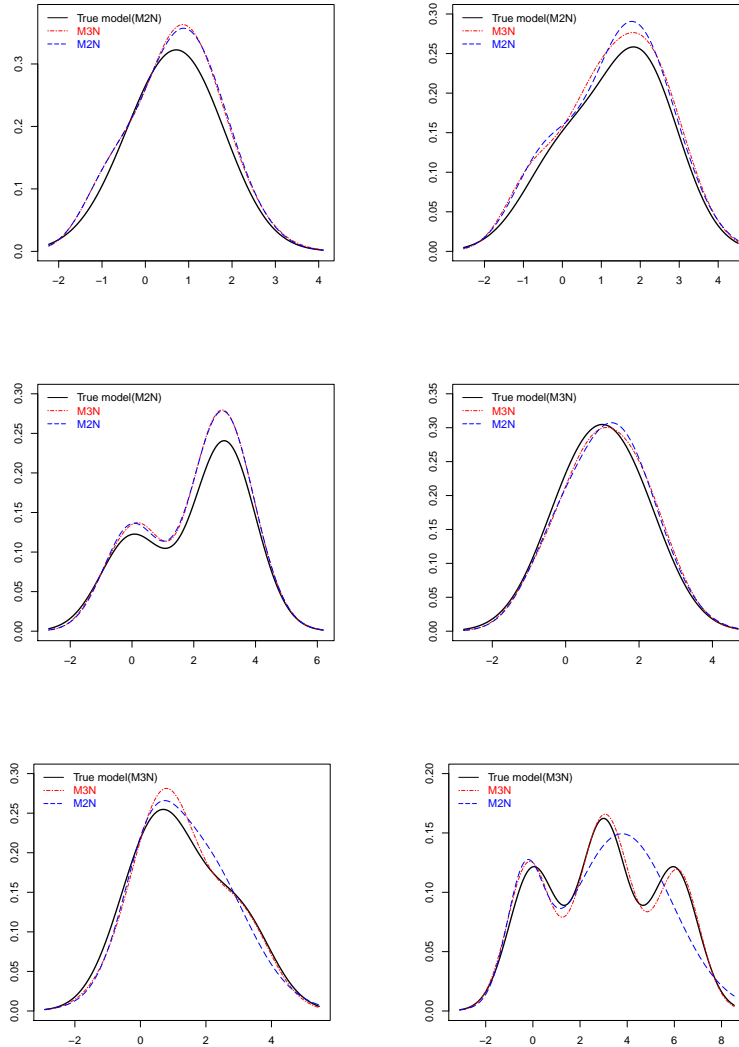


Figure 1. Results of Table 2 for rows 1, 3 and D_1 , D_2 , D_3 when sample size is 500.

-4, 3 and 1 respectively, $(0.3, -4, 1, 3, 1)$, as true underlying model. Therefore, we choose two competing models, mixture of two normal $M2N$ and mixture of two Cauchy $M3C$, based on Figures 2, *a*, *b* and *c*. Under p-value

of Vuong's test, we reject null hypothesis at the 1 percent significance level and based on Vuong's statistic, we select the mixture of two normal distributions as optimum model for all sample sizes, AIC_b , BIC_b and AIC , BIC confirm this result, Tables 7 and 8 respectively. Next, we considered random sample which has drawn from the mixture of two log-normal and mixture of two Weibull distributions as true underlying models; $(0.2, 1, 0.1, 2, 0.2)$ and $(0.4, 3, 0.5, 9, 3)$ respectively. In each rows, we choose two competing models, mixture of two Gamma $M2G$ and mixture of two log-normal $M3LN$ when data come from mixture of two log-normal $M3LN_t$. To compare the rival models, we see that in well-specified cases (row 2), based on p-value at the 1 percent significance level, Vuong's test indicates that two-components log-normal mixture as optimum model, AIC_b , BIC_b and AIC , BIC confirm this result for any sample size. Similarly, the two-components Gamma distributions are optimum models when the true models is mis-specified (row 3), $M2W_t$, but AIC and BIC select the $M2LN$ as optimum model when sample size is small. However AIC_b , BIC_b and model selection test select the two-components Gamma distributions. In this situation, the comparison between AIC_b , BIC_b and AIC , BIC when the rival models are non-nested, have the same result. Comparisons of the p-values for each sample size indicate that the KS -test can not select a optimum model because all models are equivalent. In this case the power of test growing when sample size increase for each rows when $\alpha = 0.05$.

5.3 Comparisons with LMR Tests and Proposed Test

We performed simulation studies to compare the performances of the proposed test and the Lo-Mendell-Rubin (LMR) likelihood ratio test in terms of the observed significance level. The observed LMR p-value and the observed CLR p-value were estimated from 10^4 replications for each random sample. Following Vuong (1989), Lo et al. (2001) showed that the likelihood ratio statistic is asymptotically distributed as a weighted sum of independent chi-squared random variables with one degree of freedom. They conducted simulation studies for the case of a single normal versus a two-component normal mixture where samples were generated from the standard normal distribution and the case of a two-component normal mixture versus a three-component normal mixture where samples were generated from the two-component normal mixture. Their simulation results showed that the test works well for testing the number of components in a homoscedastic normal mixture with

Table 6. p-value of KS test and Vuong’s test and power of Vuong’s test

True model	Rival model	Statistical test	50	80	200	500	
$M2N_t$	$M2N^*$	KS-test	0.9667	0.9794	0.3613	0.9022	
		Vuong’s test	5.0747	7.2997	7.4336	12.4652	
		p-value	0.0027	0.0004	0.0000	0.0000	
			Power of test	0.9719	0.9741	0.9827	1
	$M2C$	KS-test	0.5487	0.5625	0.5107	0.2262	
		$M2G$	KS-test	0.7166	0.6953	0.3613	0.8632
Vuong’s test			-2.996	-3.8995	-4.5533	-9.3653	
p-value	0.0081		0.0012	0.0000	0.0000		
		Power of test	0.9613	0.9854	1	1	
$M2LN_t$	$M2LN^*$	KS-test	0.9667	0.9794	0.5107	0.8632	
		$M2G^*$	KS-test	0.9667	0.9220	0.9550	0.4595
			Vuong’s test	2.9723	3.6068	4.5466	7.4894
p-value	0.0085		0.0003	0.0000	0.0000		
		Power of test	0.9480	0.9855	.9941	1	
$M2W_t$	$M2LN$	KS-test	0.7166	0.1202	0.0322	0.0007	

Note: $M2N$: mixture of two normal, $M2C$: mixture of two Cauchy, $M2LN$: mixture of two log-normal, $M2G$: mixture of two Gamma and $M2W$: mixture of two Weibull. * Shows the select models based on Vuong’s test.

Table 7. AIC_b and BIC_b information criteria

True model	Rival model	Criteria information	50	80	200	500
$M2N_t$	$M2N$	AIC_b	219.1923*	359.9992*	850.3431*	2105.2231*
		BIC_b	230.0122*	377.8123*	878.9123*	2135.5772*
	$M2C$	AIC_b	242.9982	439.7878	911.9134	2260.3455
		BIC_b	259.1398	448.5222	929.5611	2276.9911
$M2LN_t$	$M2G$	AIC_b	211.9988	356.8877	804.9881	2001.3343
		BIC_b	220.3421	374.8871	824.7774	2040.8899
	$M2LN$	AIC_b	200.6544*	348.9912*	788.2333*	1990.8844*
		BIC_b	211.4588*	359.5774*	809.2884*	2010.3342*
$M2W_t$	$M2G$	AIC_b	83.3443*	148.7784*	359.9999*	881.7781*
		BIC_b	95.8988*	159.9911*	388.6556*	900.7784*
	$M2LN$	AIC_b	89.8827	160.3121	370.9744	890.8883
		BIC_b	100.1220	170.1133	398.3443	910.9912

Note: * Shows the select models.

Table 8. AIC and BIC information criteria

True model	Rival model	Criteria information	50	80	200	500
$M2N_t$	$M2N$	AIC	211.1814*	352.4871*	845.3184*	2095.3130*
		BIC	226.4775*	371.5433*	871.7050*	2129.0300*
	$M2C$	AIC	236.0998	431.5767	907.1646	2252.0890
		BIC	245.6599	443.4856	923.6565	2273.1590
$M2LN_t$	$M2G$	AIC	205.9072	348.2098	792.5340	1992.7330
		BIC	221.2034	367.2657	818.9225	2026.4500
	$M2LN$	AIC	199.9322*	340.0101*	782.7042*	1985.7260*
		BIC	209.4924*	351.9202*	799.1958*	2006.7990*
$M2W_t$	$M2G$	AIC	84.7001	148.5464	366.9290*	870.1877*
		BIC	99.9968	167.6027	392.3079	903.9046*
	$M2LN$	AIC	82.1281*	147.6537*	368.9641	888.7999
		BIC	91.6881*	159.5638*	385.4557*	909.8729

Note: * Shows the select models.

suggested adjustment to the likelihood ratio statistic. Lo et al. (2001) showed that, under some regularity conditions, the likelihood ratio test statistic denoted by $2LR$ for testing $H_0^{fg} : E_h\{\log f(X|\Gamma_*)\} = E_h\{\log g(X|\Psi_*)\}$ against $H_1^f : E_h\{\log f(X|\Gamma_*)\} > E_h\{\log g(X|\Psi_*)\}$ is asymptotically distributed as a weighted sum of $\ell + j$ independent χ_1^2 random variables under the null hypothesis, where ℓ and j are unknown parameters of F_Γ and G_Ψ , respectively. They also show that the convergence rate of $2LR$ to the limiting distribution is very slow and suggested using the modified likelihood ratio statistic, LMR, $2LR^*(\hat{\Gamma}_n, \hat{\Psi}_n) = \frac{2LR(\hat{\Gamma}_n, \hat{\Psi}_n)}{1 + \{(\ell - j) \log n\}^{-1}}$ where $2LR(\hat{\Gamma}_n, \hat{\Psi}_n) = \sum_{i=1}^n \log \frac{f(x_i|\hat{\Gamma}_n)}{g(x_i|\hat{\Psi}_n)}$ to achieve reasonable accuracy. Table 9 gives the simulated significance levels for the proposed test and the LMR test for testing a single normal versus a mixture of two normal when the true distribution is standard normal. The results show that the significance levels accepted the null hypothesis for both tests when $\alpha = 0.05$. Table 10 give the simulated significance levels for the proposed test and the LMR test for testing a mixture of two normal versus a mixture of three normal when the true distribution is mixture of two normal where samples were generated from the two-component normal mixture base on Table 1. The results show that the significance levels accepted the null hypothesis for both tests when $\alpha = 0.05$.

Table 9. Simulated significance levels for the proposed and LMR tests for testing a single normal versus a mixture of two normal based on 10^4 replications for each sample size at $\alpha = 0.05$

Test procedure	50	75	100	200
proposed test	1	1	1	1
LMR test	0.8808	0.9081	0.9403	0.9690

Table 10. Simulated significance levels for testing a proposed and LMR testes for testing a mixture of two normal versus a mixture of three normal based on 10^4 replications for each sample size at $\alpha = 0.05$

Test procedure	50			75		
	D_1	D_2	D_3	D_1	D_2	D_3
proposed test	0.9999	0.9999	0.9611	0.9998	1	0.7231
LMR test	0.9999	0.9992	0.9999	1	1	0.9875

Continued on next Table.

Table 10. (continued)

Test procedure	100			200		
	D_1	D_2	D_3	D_1	D_2	D_3
proposed test	0.9547	1	1	1	1	1
LMR test	0.4882	0.9998	0.9978	0.9999	0.9880	0.9450

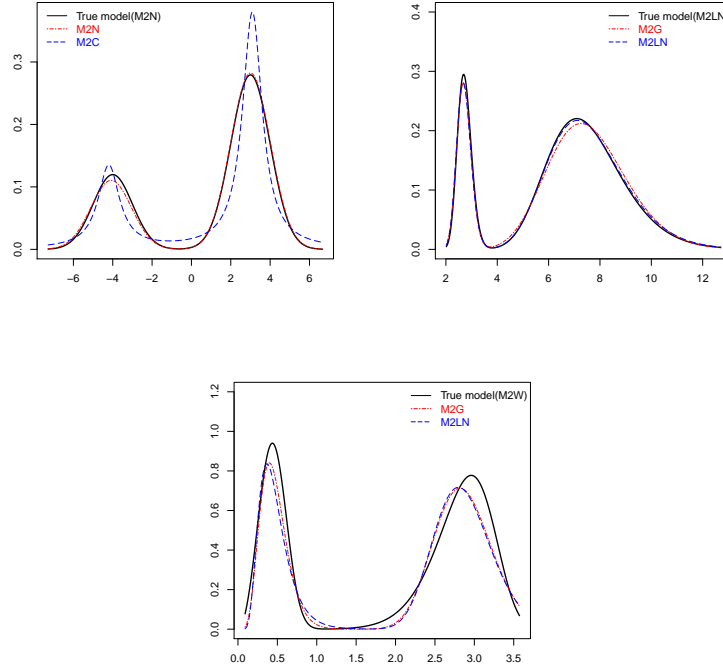


Figure 2. Results of Table 6 for each rows when sample size is 500

5.4 Total Energy Production Analysis

Here, we present a data analysis of Total Energy Production obtained by the U.S. Energy Information Administration's (EIA) State Energy Data System (SEDS) for 10 top Total Energy Production Rankings, in United State of America. Energy sources are also measured by their heat content, generally expressed in Billion British thermal units (Billion Btu). For illustrative purpose, we will be considering the data of total energy production of Table 11 from 1960 to 2013, ($n = 540$). Summary of data is given in Table 11 and Figure 3. This plot shows a strong relationship supporting the appropriateness of the some mixture distribution. So for more comparison, first we purpose adaptation normal mixture versus Cauchy mixture and log-normal mixture versus folded normal mixture models (MFN) as competing models, rows 1 and 5, Table 12. First, we estimate parameters of distributions and missing

Table 11. Summary of data

Rankings	State	Sample size	Mean	Sd	Median	Min	Max
1	Texas	54	12720000	2240779	12390000	9180000	17400000
2	Wyoming	54	4685000	3310105	3738000	1018000	10890000
3	Pennsylvania	54	2662000	712108.7	2596000	430800	5880000
4	West Virginia	54	1209000	207204.8	1208000	824400	1649000
5	Oklahoma	54	2798000	372107.9	2882000	2073000	3386000
6	North Dakota	54	633400	439662	669600	197500	2632000
7	Colorado	54	1168000	767796.7	831600	405600	2921000
8	Louisiana	54	6676000	3622799	6676000	2313000	14140000
9	Illinois	54	1925000	248559.7	1899000	1506000	2520000
10	New Mexico	54	2165000	440682.3	2156000	239100	2780000

data variable based on EM algorithm and then calculate our statistic. Based on Subsection 4.2, we reject null hypothesis at the 1 percent significance level and based on Vuong's statistic, we select the mixture of eight normal distributions and mixture of eight folded normal distributions as optimum model for rows 1 and 6 respectively, AIC_b , BIC_b and AIC , BIC confirm these results. Also, for more deliberation about number of component consider the different situations, in rows 2, 3, 4 and 5 about number of component in normal mixture distributions and in rows 7, 8, 9, 10 about number of component in folded normal mixture distributions. Based on Subsection 4.1, Under p-values for Vuong's test we reject null hypothesis and select the mixture of seven normal distributions, in row 2, and mixture of eight normal distributions at the 5 percent significance level respect to our test, in rows 3, 4 and 5 respectively. Similarly in row 7, we select the mixture of seven folded normal distributions and mixture of eight folded normal distributions at the 5 percent levels of significance, in rows 8, 9 and 10 respectively. On the other hand, KS -test result select both rival models in each test and instead of we use AIC_b and BIC_b criteria that they confirm our results. Therefore, we can select mixture of eight normal and mixture of eight folded normal distributions as optimum model based on our test with respect to support \mathfrak{R} and \mathfrak{R}^+ respectively. Obviously, if we want to select one model between two above models, we can used AIC_b and BIC_b , see Table , and select mixture of eight folded normal distribution because it has smaller information criteria.

Table 12. Values of Vuong's statistic, p-value of Vuong's statistic and *KS*-test, AIC and BIC.

Row	Competing model	Vuong statistic	p-value of Vuong's test	p-value of <i>KS</i> -test	AIC_b	BIC_b	AIC	BIC
1	<i>M8N*</i>	3.9764	0.0001	0.9685	17319.08*	17417.79*	17213.06*	17311.76*
	<i>M8C</i>			0.5478	18458.75	18157.15	18378.99	18066.22
2	<i>M7N*</i>	92.2824	0.0000	0.8569	17579.22*	17665.05*	17207.08	17292.91
	<i>M6N</i>			0.7469	17665.50	17738.46	17204.55*	17277.51*
3	<i>M8N*</i>	266.1404	0.0000	0.9685	17319.08*	17417.79*	17213.06	17311.76
	<i>M7N</i>			0.8569	17579.22	17665.05	17207.08*	17292.91*
4	<i>M9N</i>	-97.9347	1	0.6587	17429.01	17553.47	17218.51	17330.09
	<i>M8N*</i>			0.9685	17319.08*	17417.79*	17213.06*	17311.76*
5	<i>M10N</i>	-97.9347	1	0.5981	17429.01	17553.47	17214.57	17339.03
	<i>M8N*</i>			0.9685	17319.08*	17417.79*	17213.06*	17311.76*
6	<i>M8LN</i>	-4.4963	0.0000	0.6894	17431.08	17529.79	17375.42	17474.13
	<i>M8FN*</i>			0.9881	16899.07*	16997.78*	17155.33*	17254.04*
7	<i>M7FN*</i>	250.8745	0.0000	0.8426	17235.47*	17321.30*	17140.54	17189.11
	<i>M6FN</i>			0.6829	18547.51	17985.98	17103.67*	17031.65*
8	<i>M8FN*</i>	342.3121	0.0000	0.9881	16899.07*	16997.78*	17155.33	17254.04
	<i>M7FN</i>			0.8426	17235.47	17321.30	17140.54*	17189.11*
9	<i>M9FN</i>	-147.6431	1	0.8594	17052.80	17164.38	17182.66	17279.76
	<i>M8FN*</i>			0.9881	16899.07*	16997.78*	17155.33*	17254.04*
10	<i>M10FN</i>	-187.0711	1	0.8125	17098.23	17222.68	17250.79	17275.24
	<i>M8FN*</i>			0.9881	16899.07*	16997.78*	17155.33*	17254.04*

*Note:** Shows the select models.

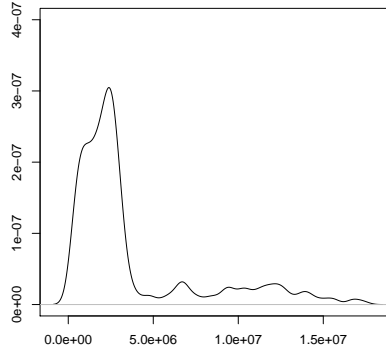


Figure 3. density plot for real data

6 Conclusion

We have proposed an altering mixture model distribution to a complete-data distribution using missing data variable and have shown that our idea is applicable to use Vuong's test for select optimum mixture models when number of components are known (fixed) or unknown. Indeed, this form of complete-data distribution have some privileges than mixture distribution; including, closed and linear form for log-likelihood function, identifiability of parameter space of mixture models, the complete-data set (X, Z) contains more information about the unknown parameter than observed data. We can to easy check all assumptions in White (1982) and Vuong (1989). Vuong's statistic is simpler than other tests in model selection theory. This explains why Vuong's test is a popular test in model selection. One different between Vuong (1989) test and this work is that the marginal density in this work is parametric density. Also, we see that AIC_b and BIC_b select the better model than AIC and BIC . In part of data analysis, although real data have positive value but maybe we select adapted normal mixture as optimum model.

Acknowledgment

I would like to thank Editor and referees for their valuable comments and suggestions that have greatly improved the presentation of the paper.

References

- Akaike, H. (1973). Information Theory and an Extension of Maximum Likelihood Principle. Second International Symposium on Information Theory, *Akademia Kiado*, 267-281.
- Chen, H., Chen, J. and Kalbfleisch, J.D. (2001). A Modified Likelihood Ratio Test for Homogeneity in the Finite Mixture Models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **63**, 19-29.
- Chen, J. and Kalbfleisch, J.D. (2005). Modified Likelihood Ratio Test in Finite Mixture Models with a Structural Parameter. *Journal of statistical Planning and Inference*, **129**, 93-107.
- Chernoff, H. and Lander, E. (1995). Asymptotic Distribution of the Likelihood Ratio Test that a Mixture of Two Binomials is a Single Binomial. *Journal of Statistical Planning and Inference*, **43**, 19-40.
- Crawford, S.L. (1994). An Application of the Laplace Method to Finite Mixture Distributions. *Journal of the American Statistical Association*, **89**, 259-267.
- Dacunha-Castelle, D. and Gassiat, E. (1999). Testing the Order of a Model using Locally Conic Parametrization: Population Mixtures and Stationary ARMA Processes. *The Annals of Statistics*, **27**, 1178-1209.
- Everitt, B.S. and David, J. Hand (1981). *Finite Mixture Distributions*. Monographs on Applied Probability and Statistics. Chapman and Hall, London, New York.
- Fallahigilan, S. and Sayyareh, A. (2016). Finite Mixture Model Selection for Total Energy Consumption. *International Journal of Energy and Statistics*, **4**.
- Feng, Z.D. and McCulloch, C.E. (1994). On the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture with Unequal Variances. *Biometrics*, 1158-1162.
- Ghosh, J.H. and Sen, P.K. (1985). *On the Asymptotic Performance of the Log Likelihood Ratio Statistic for the Mixture Model and Related Results*. In: Le Cam, L.M., Olshen, R.A. (Eds.) Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, vol. II. Wadsworth, Monterey, pp. 789-806.
- Jeffries, N.O. (2003). A Note on Testing the Number of Components in a Normal Mixture. *Biometrika*, **90**, 991-994.
- Kazakos, D. (1977). *Recursive Estimation of Prior Probabilities using a Mixture*. IEEE Transactions on Information Theory, **23**, 203-211.
- Lo, Y. (2005). Likelihood Ratio Tests of the Number of Components in a Normal Mixture with Unequal Variances. *Statistics & probability letters*, **71**, 225-235.

- Lo, Y., Mendell, N.R. and Rubin, D.B. (2001). Testing the Number of Components in a Normal Mixture. *Biometrika*, **88**, 767-778.
- McLachlan, G.J. and Basford, K.E. (1988). *Mixture Models: Inference and Applications to Clustering*, **84**,.
- McLachlan, G. and Peel, D. (2004). *Finite Mixture Models*. John Wiley & Sons.
- Morgan, G.B. (2015). Mixed Mode Latent Class Analysis: An Examination of Fit Index Performance for Classification. *Structural Equation Modeling: A Multidisciplinary Journal*, **22**, 76-86.
- Morgan, G.B., Hodge, K.J. and Baggett, A.R. (2016). Latent Profile Analysis with Nonnormal Mixtures: A Monte Carlo Examination of Model Selection using Fit Indices. *Computational Statistics & Data Analysis*, **93**, 146-161.
- Nylund, K.L., Asparouhov, T. and Muthen, B.O. (2008). Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study: Erratum.
- Redner, R.A. and Walker, H.F. (1984). *Mixture Densities, Maximum Likelihood and the EM Algorithm*. SIAM review, **26**, 195-239.
- Sayyareh, A. (2016). Admissible Set of Rival Models based on the Mixture of Kullback-Leibler Risks. *JSRI*. **13**, 59-88
- Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, **6**, 461-464.
- Scott, A.J. and Symons, M.J. (1971). Clustering Methods based on Likelihood Ratio Criteria. *Biometrics*, 387-397.
- Symons, M.J. (1981). Clustering Criteria and Multivariate Normal Mixtures. *Biometrics*, 35-43.
- Teicher, H. (1961). Identifiability of Mixtures. *The Annals of Mathematical Statistics*, **32**, 244-248.
- Titterton, D.M., Smith, A.F.M. and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.
- Vuong, Q.H. (1989). Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses. *Econometrica: Journal of the Econometric Society*, 307-333.
- Wald, A. (1948). Estimation of a Parameter when the Number of Unknown Parameters Increases Indefinitely with the Number of Observations. *The Annals of Mathematical Statistics*, 220-227.

White, H. (1982). Maximum Likelihood Estimation of Mis-specified Models. *Econometrica: Journal of the Econometric Society*, 1-25.

Wichitchan, S., Yao, W. and Yang, G. (2018). Hypothesis Testing for Finite Mixture Models. *Computational Statistics & Data Analysis*.

Yakowitz, S.J. and Spragins, J.D. (1968). On the Identifiability of Finite Mixtures. *The Annals of Mathematical Statistics*, 209-214.

Appendix

Proof of Lemma 1: Given Voung's Assumptions 1 to 3 and the strong law of large numbers Theorem,

$$\frac{1}{n} \sum_{i=1}^n E_{\kappa_f(z_{if}|x_i; \hat{\Gamma}_n)}(\log f(X_i, Z_{if}|\hat{\Gamma}_n)|x) \xrightarrow{\text{a.s.}} E_h(E_{\kappa_f(z_f|x; \Gamma_*)}(\log f(X, Z_f|\Gamma_*)|x))$$

Since by (2)

$$\begin{aligned} & E_h(E_{\kappa_f(z_f|x; \Gamma_*)}(\log f(X, Z_f|\Gamma_*)|x)) \\ &= E_h\left(\sum_{j=1}^m (\log \alpha_{j*} + \log f(X|\theta_{j*})) E_{\kappa_f(z_f|x; \Gamma_*)}(Z_{jf}|x)\right) \end{aligned} \quad (21)$$

On the other hand

$$\begin{aligned} E_{\kappa_f(z_f|x; \Gamma_*)}(Z_{jf}|x) &= \sum_{z_{jf}=0 \text{ or } 1} z_{jf} \kappa_f(z_f|x; \Gamma_*) \\ &= \frac{\alpha_{j*} f(x|\theta_{j*})}{f(x|\Gamma_*)} = z_{j*f} \end{aligned}$$

since (21)

$$\begin{aligned} & E_h(E_{\kappa_f(z_f|x; \Gamma_*)}(\log f(X, Z_f|\Gamma_*)|x)) \\ &= E_h\left(\sum_{j=1}^m (\log \alpha_{j*} + \log f(X|\theta_{j*})) E_{\kappa_f(z_f|x; \Gamma_*)}(Z_{jf}|x)\right) \\ &= E_h\left(\sum_{j=1}^m (\log \alpha_{j*} + \log f(X|\theta_{j*})) Z_{j*f}\right) \\ &= E_h(\log f(X, Z_{*f}|\Gamma_*)). \end{aligned}$$

Proof of Lemma 2: From the proof of Lemma 1 and the strong law of large numbers Theorem, we can write similarly;

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n E_{\kappa_g(\tilde{z}_{ig}|x_i; \hat{\Psi}_n)}(\log g(X_i, Z_{ig}|\hat{\Psi}_n)|x) &= \frac{1}{n} \sum_{i=1}^n \log g(x_i, \tilde{z}_{ig}|\hat{\Psi}_n) \\ &\xrightarrow{\text{a.s.}} E_h(\log g(X, Z_{*g}|\Psi_*)). \end{aligned}$$

Similarly, other parts are easy.

Proof of Theorem 1: From a Taylor expansion of $\log L_c(\Gamma_*)$ around $\hat{\Gamma}_n$, we obtain:

$$\begin{aligned} \log L_c(\Gamma_*) &= \log f(x, z_{*f}|\Gamma_*) = \log f(x, \tilde{z}_f|\hat{\Gamma}_n) + \nabla_{\Gamma} \log f(x, \tilde{z}_f|\Gamma)|_{\Gamma=\hat{\Gamma}_n} (\Gamma - \hat{\Gamma}_n) \\ &\quad + \frac{1}{2} (\Gamma - \hat{\Gamma}_n)' \nabla_{\Gamma}^2 \log f(x, \tilde{z}_f|\Gamma)|_{\Gamma=\hat{\Gamma}_n} (\Gamma - \hat{\Gamma}_n) + o_p(1) \end{aligned} \quad (22)$$

according to the strong law of large numbers Theorem and Young's Assumption 4, we have

$$\frac{1}{n} \sum_{i=1}^n \nabla_{\Gamma}^2 \log f(x, \tilde{z}_f|\Gamma)|_{\Gamma=\hat{\Gamma}_n} \xrightarrow{\text{a.s.}} E_h\{\nabla_{\Gamma}^2 \log f(x, z_{*f}|\Gamma_*)\} = A_f(\Gamma_*)$$

so

$$\log L_c(\Gamma_*) = \log f(x, \tilde{z}_f|\hat{\Gamma}_n) + \frac{n}{2} (\Gamma - \hat{\Gamma}_n)' A_f(\Gamma_*) (\Gamma - \hat{\Gamma}_n) + o_p(1). \quad (23)$$

Similarly, we have;

$$\log L_c(\Psi_*) = \log f(x, \tilde{z}_g|\hat{\Psi}_n) + \frac{n}{2} (\Psi - \hat{\Psi}_n)' A_g(\Psi_*) (\Psi - \hat{\Psi}_n) + o_p(1).$$

Since $CLR(\Gamma_*, \Psi_*) = \log L_c(\Gamma_*) - \log L_c(\Psi_*)$, we obtain;

$$\begin{aligned} CLR(\hat{\Gamma}_n, \hat{\Psi}_n) &= CLR(\Gamma_*, \Psi_*) - \frac{n}{2} (\Gamma - \hat{\Gamma}_n)' A_f(\Gamma_*) (\Gamma - \hat{\Gamma}_n) \\ &\quad + \frac{n}{2} (\Psi - \hat{\Psi}_n)' A_g(\Psi_*) (\Psi - \hat{\Psi}_n) + o_p(1). \end{aligned} \quad (24)$$

To prove (a), we note that $CLR(\Gamma_*, \Psi_*) = 0$ if two rival models be nested. Part (a) by Vuong's Lemma A and Lemma 2 obtain;

$$\begin{aligned} 2CLR(\hat{\Gamma}_n, \hat{\Psi}_n) &= n(\Gamma - \hat{\Gamma}_n)' A_f(\Gamma_*)(\Gamma - \hat{\Gamma}_n) + n(\Psi - \hat{\Psi}_n)' \\ &\quad \times A_g(\Psi_*)(\Psi - \hat{\Psi}_n) + o_p(1) \\ &= \begin{bmatrix} \sqrt{n}(\Gamma - \hat{\Gamma}_n) & \sqrt{n}(\Psi - \hat{\Psi}_n) \end{bmatrix} \begin{bmatrix} -A_f(\Gamma_*) & 0 \\ 0 & A_g(\Psi_*) \end{bmatrix} \\ &\quad \times \begin{bmatrix} \sqrt{n}(\Gamma - \hat{\Gamma}_n) \\ \sqrt{n}(\Psi - \hat{\Psi}_n) \end{bmatrix} + o_p(1). \end{aligned}$$

Then one can check that

$$\begin{aligned} W &= \begin{bmatrix} -A_f(\Gamma_*) & 0 \\ 0 & A_g(\Psi_*) \end{bmatrix} \\ &\quad \times \begin{bmatrix} A_f^{-1}(\Gamma_*)B_f(\Gamma_*)A_f^{-1}(\Gamma_*) & A_f^{-1}(\Gamma_*)B_{fg}(\Gamma_*, \Psi_*)A_g^{-1}(\Psi_*) \\ A_g^{-1}(\Psi_*)B_{gf}(\Psi_*, \Gamma_*)A_f^{-1}(\Gamma_*) & A_g^{-1}(\Psi_*)B_g(\Psi_*)A_g^{-1}(\Psi_*) \end{bmatrix}. \end{aligned}$$

to prove (b) we note that $\sqrt{n}(\Gamma - \hat{\Gamma}_n)$ and $\sqrt{n}(\Psi - \hat{\Psi}_n)$ are $O_p(1)$. Thus (24) we obtain;

$$\begin{aligned} n^{-\frac{1}{2}}CLR(\hat{\Gamma}_n, \hat{\Psi}_n) - n^{\frac{1}{2}}E_h(\log \frac{f(X, Z_{*f}|\Gamma_*)}{g(X, Z_{*g}|\Psi_*)}) &= n^{\frac{1}{2}}(CLR(\Gamma_*, \Psi_*) \\ &\quad - E_h(\log \frac{f(X, Z_{*f}|\Gamma_*)}{g(X, X_{*g}|\Psi_*)})) + o_p(1). \end{aligned}$$

But from the multivariate Central Limit Theorem, the first term in the right-hand side converges in distribution to $N(0, \omega_*)$. This variance is finite given Vuong's Assumption 6 and the Cauchy-Schwartz inequality. Part (b) follows.

Proof of Lemma 3: We know that

$$\hat{\omega}_n^2 = \frac{1}{n} \sum_{i=1}^n [\log \frac{f(x_i, \tilde{z}_{if}|\hat{\Gamma}_n)}{g(x_i, \tilde{z}_{ig}|\hat{\Psi}_n)}]^2 - [\frac{1}{n} \sum_{i=1}^n \log \frac{f(x_i, \tilde{z}_{if}|\hat{\Gamma}_n)}{g(x_i, \tilde{z}_{ig}|\hat{\Psi}_n)}]^2$$

and

$$\omega_*^2 = \text{Var}_h\left(\log \frac{f(X, Z_{*f}|\Gamma_*)}{g(X, Z_{*g}|\Psi_*)}\right)$$

Given Young's Assumptions 1 to 3 and the strong law of large numbers Theorem and Corollary 1

$$\frac{1}{n} \sum_{i=1}^n \log f(x_i, \tilde{z}_{if}|\hat{\Gamma}_n) \xrightarrow{\text{a.s.}} E_h(\log f(X, Z_{*f}|\Gamma_*))$$

and

$$\frac{1}{n} \sum_{i=1}^n \log g(x_i, \tilde{z}_{ig}|\hat{\Psi}_n) \xrightarrow{\text{a.s.}} E_h(\log g(X, Z_{*g}|\Psi_*))$$

Since

$$\frac{1}{n} \sum_{i=1}^n \log \frac{f(x_i, \tilde{z}_{if}|\hat{\Gamma}_n)}{g(x_i, \tilde{z}_{ig}|\hat{\Psi}_n)} \xrightarrow{\text{a.s.}} E_h\left(\log \frac{f(X, Z_{*f}|\Gamma_*)}{g(X, Z_{*g}|\Psi_*)}\right)$$

Similarly

$$\frac{1}{n} \sum_{i=1}^n \left(\log \frac{f(x_i, \tilde{z}_{if}|\hat{\Gamma}_n)}{g(x_i, \tilde{z}_{ig}|\hat{\Psi}_n)}\right)^2 \xrightarrow{\text{a.s.}} E_h\left(\log \frac{f(X, Z_{*f}|\Gamma_*)}{g(X, Z_{*g}|\Psi_*)}\right)^2$$

Proof of Theorem 2 and 3: Straightforward from Theorem 1 (a) and (b), respectively.

Sadegh Fallahigilan

Department of Statistics,
Razi University,
Kermanshah, Iran.
email: *sadegh.falahi@yahoo.com*,

Abdolreza Sayyareh

Department of Computer Science and Statistics,
Faculty of Mathematics.
K. N. Toosi University of Technology,
Tehran, Iran.
email: *asayyareh@kntu.ac.ir*