

# SELECTED ESTIMATED MODELS WITH $\phi$ -DIVERGENCE STATISTICS

PAPA NGOM

*Laboratoire de Mathématiques appliquées (LMA)  
Université Cheikh Anta Diop - Dakar - Sénégal  
e-mail : pngom@ucad.sn*

ABSTRACT. When testing for discriminating between two competing models, a statistical method, usually, proceeds by evaluating the measure for discrepancy between the observed data and each parametric model. The parameter model with smaller value of measure statistic is generally chosen. This paper addresses the question of testing for choosing between two estimated models using some  $\phi$ -divergence type statistics. We choice for arbitrary  $\sqrt{n}$ -asymptotically normal estimators to be used for introducing these statistics. The results here are illustrated by a simulation study, then Large Sample theory and bootstrap methods are used to construct our  $\phi$ -divergence tests in parametric models.

## 1. Introduction

Cochran [6], Watson [34] and Moore[17][18] have provided comprehensive surveys on Pearson chi-square type statistics, i.e., quadratic forms in the cell frequencies. Recently, Andrews [2],[3] has extended the Pearson chi-square testing method to non-dynamic parametric models, i.e., to models with covariates. Because Pearson chi-square statistics provide natural measures for the discrepancy between the observed data and a specific parametric model, they have also been used for discriminating among competing models. Such a situation is frequent in Social Sciences where many competing models are proposed to fit a given sample. A well know difficulty is that each chi-square statistic tends to become large without an increase in its degrees of freedom as the sample size increases. As a consequence goodness-of-fit tests based on Pearson type chi-square statistics will generally reject the correct specification of every competing model.

To circumvent such a difficulty, a popular method for model selection, which is similar to use of Akaike [1] Information Criterion (AIC), consists in considering that the lower the chi-square statistic, the better is the model.

The preceding selection rule, however, is not entirely satisfactory. Since chi-square statistics depend on the sample and are therefore random, their actual values are subject to statistical variations, we shall propose some convenient asymptotically standard normal tests for model selection based on  $\phi$ -divergence type statistics. By analogy with the approach introduced by Vuong [32], our tests are testing the

---

*Key words and phrases.* Asymptotic distributions,  $\phi$ -Divergence statistics, bootstrap methods, testing statistical hypotheses, test goodness fit.

null hypothesis that the competing models are as close to the data generating process (DGP) where closeness of a model is measured according to the discrepancy implicit in the  $\phi$ -divergence type statistics.

Following Morales and Pardo [21], let  $P_\theta : \theta \in \Theta$  be a family of probability measures on a measurable space  $(\mathcal{X}, \beta_{\mathcal{X}})$  with open  $\Theta \subset \mathbb{R}^d$ ,  $d \geq 1$ . Measures  $P_\theta$  are described by probability density functions (p.d.f.)  $f_\theta(x) = \frac{dP_\theta}{d\mu}(x)$  with respect to a dominating  $\sigma$ -finite measure  $\mu$  on  $\mathcal{X}$ . Sample space,  $\mathcal{X}$ , is the support of  $\sigma$ -finite measure  $\mu$ . Statistical model,  $((\mathcal{X}, \beta_{\mathcal{X}}), f_\theta : \theta \in \Theta, \mu)$ , satisfies the regularity assumptions (R1) – (R3) appearing in pages 144-145 of Serfling [27] and the identifiability condition : (R4) if  $f_{\theta_1} = f_{\theta_2}$ , then  $\theta_1 = \theta_2$ .

If  $\theta_0$  is the true value of the parameter  $\theta$  and (R1) – (R4) holds, then there exist a strongly consistent sequence  $\hat{\theta}_n$  of roots of the likelihood equations such that

$$(1) \quad \sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{L} N(0, I_F(\theta_0)^{-1}),$$

where  $I_F(\theta_0)$  is the Fisher information matrix and  $\hat{\theta}_n$  is assumed to be the maximum likelihood estimator (MLE).

We consider testing procedures based on a sequence of observations

$X_n = (X_1, X_2, \dots, X_n)$  with independent components taken from a p.d.f of the family  $f_\theta : \theta \in \Theta$ .

Recently, in the literature, many papers appeared where divergence or type measures of information have been used in testing statistical hypothesis. We refer, among others, to Cressie and Read [8], Nayak [22], Zografos, Ferentinos and Papaioannou [33] Salicrù, Morales, Menéndez and Pardo [23], Bar-Hen and Daudin [5] and references therein. Salicrù et al. [26] introduced the divergence statistics  $S_{\phi,n} \equiv 2nC_\phi(\hat{\theta}_n, \theta_0)$  where

$$(2) \quad C_\phi(\theta_1, \theta_2) = \int_{\mathcal{X}} f_{\theta_2}(x) \phi\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right) d\mu(x)$$

is the  $\phi$ -divergence of density from the family  $f_\theta : \theta \in \Theta$  introduced by Csiszàr [9]. Liese and Vajda [15] have introduced a systematic theory of these divergences.

Morales et al.[16] have established that the asymptotic distribution of  $S_{\phi,n} \xrightarrow{L} \chi_d^2$ . An important problem is to propose some divergences statistics for procedure tests.

The asymptotic behavior of the statistics based on  $C_\phi(\hat{\theta}_n, \theta_0)$  is needed for choosing between two estimated models. In order to suggest a testing procedure, we present a new method in association with the divergence statistic.

The paper is organized as follows. Section 2 introduces the basic notations and defines a class of asymptotically normal estimators. In section 3, we investigate the model selection problem based on divergence type statistics. A large sample test is proposed. In section 4, Efron [10] bootstrap method is used to propose alternative and simpler testing procedures for model selection. Section 5, some simulation results are given. Section 6 concludes the paper and mentions some extensions.

## 2. Assumption and Asymptotic behavior of the divergence statistic

Assumption (A1) :

(i) The function  $\phi : [0, +\infty[ \rightarrow ]-\infty, +\infty[$  is convex and continuous. Its restriction on  $[0, +\infty[$  is finite, twice continuously differentiable, with  $\phi(1) = \phi'(1) = 0$  and

$\phi''(1) = 1$ ;

(ii) Each  $\theta_0 \in \Theta$  has an open neighborhood  $V(\theta_0)$  and  $1 \leq i, j \leq d$ , it holds :

$$\begin{aligned} \frac{\partial}{\partial \theta_i} \int_{\mathcal{X}} f_{\theta_0}(x) \phi\left(\frac{f_{\theta}(x)}{f_{\theta_0}(x)}\right) d\mu(x) &= \int_{\mathcal{X}} \frac{\partial}{\partial \theta_i} \left[ f_{\theta_0}(x) \phi\left(\frac{f_{\theta}(x)}{f_{\theta_0}(x)}\right) \right] d\mu(x) \\ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \int_{\mathcal{X}} f_{\theta_0}(x) \phi\left(\frac{f_{\theta}(x)}{f_{\theta_0}(x)}\right) d\mu(x) &= \int_{\mathcal{X}} \frac{\partial^2}{\partial \theta_i \partial \theta_j} \left[ f_{\theta_0}(x) \phi\left(\frac{f_{\theta}(x)}{f_{\theta_0}(x)}\right) \right] d\mu(x) \end{aligned}$$

condition(i) deals with properties of  $\phi$ -divergence (cf. Liese and Vajda [15]). Condition (ii) is needed to apply delta method for obtaining asymptotic distributions of  $\phi$ -statistics. Conditions sufficient for (ii) are presented in Morales et al. [19].

Assume that  $(R_1) - (R_4)$  and A1 hold. Under  $H_o : \theta \in \Theta_o \subset \Theta$ , we present the asymptotic distribution of  $C_{\phi}(\hat{\theta}_n, \theta_o)$ .

**Theorem 2.1.** *Let the model and  $\phi$  satisfy (R1)-(R4) and (A1) respectively. Let  $\theta$  be the true parameter, with  $\theta \neq \theta_o$ . Then we have*

$$\sqrt{n} [C_{\phi}(\hat{\theta}_n, \theta_o) - C_{\phi}(\theta, \theta_o)] \rightarrow N[0, \Sigma_{\phi}^2(\theta, \theta_o)]$$

where  $\Sigma_{\phi}^2(\theta, \theta_o) = AI_F(\theta)^{-1}A^t$  and  $A = \nabla C_{\phi}(\theta, \theta_o)$  with  $\nabla = (\frac{\partial}{\partial \theta_1}, \dots, \frac{\partial}{\partial \theta_d})$ .

*Proof.* A first order Taylor expansion gives

$$C_{\phi}(\hat{\theta}_n, \theta_o) = C_{\phi}(\theta, \theta_o) + A(\hat{\theta}_n - \theta_o)^t + o(\|\hat{\theta}_n - \theta_o\|)$$

As

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{L} N(0, I_F(\theta)^{-1}),$$

it is clear that the random variables,

$\sqrt{n}[C_{\phi}(\hat{\theta}_n, \theta_o) - C_{\phi}(\theta, \theta_o)]$  and  $A\sqrt{n}(\hat{\theta}_n - \theta)^t$  have the same asymptotic distribution, because

$$\sqrt{n} o(\|\hat{\theta}_n - \theta\|) = o_p(1)$$

□

### 3. Selecting estimated models

As we mentioned earlier, the type divergences statistics can be used to discriminate among alternative models.

Let  $h$  be the true probability density of the observations  $X_n = (X_1, \dots, X_n)$ . We consider a specified model  $F_{\theta} = \{F(\cdot|\theta); \theta \in \Theta \subset \mathbb{R}^k\}$  with  $f_{\theta}(x)$  as the probability density function. Therefore, we define the discrepancy between the observations and the model  $F_{\theta}$  as following :

$$D(\theta) = C_{\phi}(h, f_{\theta}) = \int_{\mathcal{X}} f_{\theta}(x) \phi\left(\frac{h(x)}{f_{\theta}(x)}\right) d\mu(x)$$

Of special interests to us is the situation in which a researcher has two competing parametric models  $F_{\theta}$  and  $G_{\gamma} = \{G(\cdot|\gamma); \gamma \in \Gamma \subset \mathbb{R}^k\}$ , select the better of the two models based on their general discrimination statistics  $D_n(\hat{\theta}_n) = C_{\phi}(h, f_{\hat{\theta}_n})$  and  $D_n(\hat{\gamma}_n) = C_{\phi}(h, g_{\hat{\gamma}_n})$  where  $\hat{\theta}_n$  and  $\hat{\gamma}_n$  are general estimators satisfying condition (1).

**Definition 3.1.** (*Equivalent, Better and Worse*)

Consider two competing models  $F_\theta$  and  $G_\gamma$  and some discrimination type statistics  $D_n(\hat{\theta}_n)$  and  $D_n(\hat{\gamma}_n)$  where  $\hat{\theta}_n$  and  $\hat{\gamma}_n$  are general estimators satisfying condition (1). Let  $D(\cdot)$  be the probability limit of  $\sqrt{n}D_n(\cdot)$ .

The hypotheses

$$\begin{aligned} H_o &: D(\theta_o) = D(\gamma_o) \\ H_f &: D(\theta_o) < D(\gamma_o) \\ H_g &: D(\theta_o) > D(\gamma_o) \end{aligned}$$

mean that the estimated models  $F(\cdot|\theta_o)$  and  $G(\cdot|\gamma_o)$  are equivalent, that  $F(\cdot|\theta_o)$  is better than  $G(\cdot|\gamma_o)$ , and that  $F(\cdot|\theta_o)$  is worse than  $G(\cdot|\gamma_o)$ , respectively.

Definition (3.1) calls for some remarks. First, it does not require that the same divergence type statistics be used in forming  $D_n(\theta_n)$  and  $D_n(\gamma_n)$ . Choosing, however, different discrepancies for evaluating competing models is hardly justified. Second and more importantly, it allows estimators other than the matching divergence estimators to be used.

In any case, since  $\hat{\theta}_n, \hat{\gamma}_n$  are consistent estimators of  $\theta_o$  and  $\gamma_o$  by condition (1), we can use, from theorem 3.1,  $\sqrt{n}\{C_\phi(h, f_{\hat{\theta}_n}) - C_\phi(h, g_{\hat{\gamma}_n})\}$  to consistently estimate the indicator  $C_\phi(h, f_{\theta_o}) - C_\phi(h, g_{\gamma_o})$  which will be zero under the null hypothesis  $H_o$ . Using a standard Taylor expansion, we can obtain the asymptotic distribution of  $\sqrt{n}\{C_\phi(h, f_{\hat{\theta}_n}) - C_\phi(h, g_{\hat{\gamma}_n})\}$ , which is normal with zero mean and variance  $\omega^2$  under  $H_o$ . The detailed derivation and the expression for  $\omega^2$  can be found in the proof of the theorem (3.2).

Hence we define the statistic

$$\begin{aligned} \text{DI}_n &= \frac{\sqrt{n}}{\hat{\omega}}(D_n(\hat{\theta}_n) - D_n(\hat{\gamma}_n)) \\ (3) \quad &= \sqrt{n}\left\{\frac{C_\phi(h, f_{\hat{\theta}_n}) - C_\phi(h, g_{\hat{\gamma}_n})}{\hat{\omega}}\right\} \end{aligned}$$

where  $\hat{\omega}^2$  is a consistent estimator of  $\omega^2$ . (DI stands for Divergence Indicator).

We have,

**Theorem 3.2.** (*Asymptotic Distribution of DI Statistic*)

Given H1-H4, then

- (i) under the null hypothesis  $H_o$ ,  $\text{DI}_n \rightarrow N(0, 1)$  in distribution
- (ii) under the alternative  $H_f$ ,  $\text{DI}_n \rightarrow -\infty$  in probability,
- (iii) under the alternative  $H_g$ ,  $\text{DI}_n \rightarrow +\infty$  in probability.

*Proof.*

$$\sqrt{n}C_\phi(\hat{\theta}_n, \theta_o) = \sqrt{n}C_\phi(\theta, \theta_o) + A\sqrt{n}(\hat{\theta}_n - \theta_o)^t + \sqrt{n} o(\|\hat{\theta}_n - \theta_o\|)$$

$$\sqrt{n}C_\phi(\hat{\gamma}_n, \gamma_o) = \sqrt{n}C_\phi(\gamma, \gamma_o) + B\sqrt{n}(\hat{\gamma}_n - \gamma_o)^t + \sqrt{n} o(\|\hat{\gamma}_n - \gamma_o\|)$$

By difference, it follows that :

$$\sqrt{n}\{C_\phi(\hat{\theta}_n, \theta_o) - C_\phi(\hat{\gamma}_n, \gamma_o)\} = \sqrt{n}\{C_\phi(\theta, \theta_o) - C_\phi(\gamma, \gamma_o)\} + (A, B)\sqrt{n}\begin{pmatrix} \hat{\theta}_n - \theta_o \\ \hat{\gamma}_n - \gamma_o \end{pmatrix} + o_p(1)$$

From the multivariate central limit theorem and assumption (A1), we can now immediately obtain the asymptotic distribution of

$$\sqrt{n}\{C_\phi(\hat{\theta}_n, \theta_o) - C_\phi(\hat{\gamma}_n, \gamma_o)\}$$

under the null hypothesis of equivalence  $H_o$ .

Define :

$$T = (A, B) ; \hat{\alpha}_n = \sqrt{n}(\hat{\theta}_n - \theta_o) ; \hat{\beta}_n = \sqrt{n}(\hat{\gamma}_n - \gamma_o) ; \Lambda = \begin{pmatrix} I_F^{-1}(\theta) & \mathbb{E}(\hat{\alpha}_n \hat{\beta}_n^t) \\ \mathbb{E}(\hat{\beta}_n \hat{\alpha}_n^t) & I_G^{-1}(\gamma) \end{pmatrix}$$

with  $A = \nabla C_\phi(\theta, \theta_o)$  and  $B = \nabla C_\phi(\gamma, \gamma_o)$

Let  $\omega^2 = T \Lambda T^t$ , we then have

$$\sqrt{n} \left\{ \frac{C_\phi(\hat{\theta}_n, \theta_o) - C_\phi(\hat{\gamma}_n, \gamma_o)}{\hat{\omega}} \right\} \xrightarrow{L} N(0, 1)$$

□

**Remark 3.3.** One can note that there are some important measures of divergence which can not be written as  $\phi$ -divergence ; for instance, the divergence measures given by Battacharyya, Sharma-Mittal and Rényi. However, such measures can be written in the following form :

$$C_{\phi, h}(\theta_1, \theta_2) = h(C_\phi(\theta_1, \theta_2))$$

where  $h$  is a differentiable increasing function mapping from  $[0, +\infty[$  onto  $[0, +\infty[$ , with  $h(0) = 0$  and  $h'(x) > 0$ .

we present these divergence measures, in the following table.

Divergence	h function	$\phi$ function
Battacharyya	$h_B(x) = -\ln(-x + 1)$	$\phi_B(x) = -x^{1/2} + \frac{1}{2}(x + 1)$
Sharma-Mittal	$h_S(x) = \frac{1}{s-1} \left[ (1 + r(r-1)x)^{\frac{s-1}{r-1}} - 1 \right]$	$\phi_S(x) = \frac{x^r - r(x-1) - 1}{r(r-1)}$
Rényi	$h_R(x) = \frac{1}{r(r-1)} \ln(r(r-1)x + 1)$	$\phi_R(x) = \frac{x^r - r(x-1) - 1}{r(r-1)}$

Table 1 :  $(h, \phi)$  - Divergences with  $r \neq 0, 1$

Theorem (3.2) is quite general and gives us a wide variety of asymptotic standard normal tests for model selection based on divergence type statistics. Part (ii) and (iii) also implies that the test is consistent. In the next section, we detail the testing procedures based on Theorem (3.2) by using bootstrap methods.

#### 4. Bootstrap methods

Implementation of the model selection procedure proposed in section 3 requires the following computations :

- (i) Estimation of the parameters  $\hat{\theta}_n$  and  $\hat{\gamma}_n$ ,
- (ii) Computation of the two divergences statistics  $D_n(\hat{\theta}_n)$  and  $D_n(\hat{\gamma}_n)$  and the difference  $\hat{\mathcal{B}}_n \equiv \sqrt{n}[D_n(\hat{\theta}_n) - D_n(\hat{\gamma}_n)]$ ,
- (iii) Computation of the variance  $\hat{\omega}^2$  of  $\hat{\mathcal{B}}_n$  and finally, computation of  $DI_n \equiv \hat{\mathcal{B}}_n / \hat{\omega}$

Specifically, we carry out the following steps :

1) Let  $F_n$  be the empirical probability distribution of the original data  $x_1, x_2, \dots, x_n$  i.e.,  $F_n$  : mass  $1/n$  at  $x_i$ , ( $i = 1, 2, \dots, n$ ):

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(X_i \leq x)$$

Then draw an i.i.d "bootstrap sample"  $x_1^*, x_2^*, \dots, x_n^*$  from  $F_n$ , i.e., draw  $x_i^*$  randomly with replacement from the observed values  $x_1, x_2, \dots, x_n$ ,

2) Using this bootstrap sample  $x_i^*$ , estimate the competing models to obtain  $\theta_n^*$  and  $\gamma_n^*$ . Then calculate the statistic

$$\hat{\mathcal{B}}_n^* \equiv \sqrt{n}[D_n(\hat{\theta}_n^*) - D_n(\hat{\gamma}_n^*)]$$

3) Independently repeat steps 1 and 2 a large number of times  $S$ , say  $S=1000$ . Obtain "bootstrap replications"  $\hat{\mathcal{B}}_n^{*1}, \hat{\mathcal{B}}_n^{*2}, \dots, \hat{\mathcal{B}}_n^{*S}$ , and compute the sample variance of  $\{\hat{\mathcal{B}}_n^{*j}, j = 1, \dots, S\}$  :

$$\hat{\omega}_*^2 = \frac{1}{S} \sum_{j=1}^S (\hat{\mathcal{B}}_n^{*j} - \bar{\mathcal{B}}^*)^2,$$

where  $\bar{\mathcal{B}} = \frac{1}{S} \sum_{j=1}^S \hat{\mathcal{B}}_n^{*j}$  is the average of "bootstrap replications".

Once the bootstrap variance  $\hat{\omega}_*^2$  is obtained, the test statistic  $DI_n$  is calculated easily using the initial estimates  $\hat{\theta}_n$  and  $\hat{\gamma}_n$ . Under suitable regularity conditions and for a large number of replications [10],  $\hat{\omega}_*^2$  is a consistent estimator of  $\omega^2$ .

Thus, from theorem 3.2, a testing procedure for model selection can be based on the comparison of the value of  $DI_n$  to critical values from a standard normal table. For example, at 5% significance level, we compare  $DI_n$  with -1.96 and 1.96. If  $DI_n$  falls between -1.96 and 1.96, we conclude that both estimated models fit the data equally well. If  $DI_n$  is less than -1.96 (or larger than 1.96), then we reject the null hypothesis in favor of the alternative hypothesis that the estimated model  $F(\cdot|\hat{\theta}_n)$  (or  $G(\cdot|\hat{\gamma}_n)$ ) is closer to the true distribution.

Although using the bootstrap method to obtain an estimate of  $\omega^2$ , the basic justification of the preceding testing comes from the asymptotic properties obtained in Theorem 3.2.

## 5. Numerical study

We present briefly the basic assumptions on the model and parameter estimators, and we define our general divergence type statistics.

Assumption (A2) : The observed data  $X_i, i = 1, \dots$ , are independent and are identically distributed (iid) with some common true distribution  $H$ .

The sample space  $X$  is partitioned into  $M$  mutually disjoint fixed cells  $C_1, C_2, \dots, C_M$ . Let  $n$  be the sample size. Corresponding to the partition  $C_1, C_2, \dots, C_M$  we can compute the vector of observed cell probabilities

$$(4) \quad f = (f_1, f_2, f \dots, f_M)^t \quad \text{where} \quad f_i = \frac{1}{n} \sum_{j=1}^n I_{C_i}(X_j), \quad \text{for} \quad i = 1, 2, \dots, M.$$

and  $I_{C_i}(X_j)$  is the indicator function :

$$I_{C_i}(X_j) = \begin{cases} 1 & \text{if } X_j \text{ falls in cell } C_i, \\ 0 & \text{otherwise.} \end{cases}$$

Let a specified model be  $H_\theta = \{H(\cdot|\theta), \theta \in \Theta \subset \mathbb{R}^d\}$  and denote the vector of its predicted cell probabilities by :

$$h(\theta) = (h_1(\theta), h_2(\theta), \dots, h_M(\theta))^t \quad \text{where} \quad h_i(\theta) = \int_{C_i} dH(x|\theta)$$

where  $H(\cdot|\theta)$  is joint distribution for  $X_j$ .

We suppose  $h_i(\theta) > 0$  and  $h_i(\theta)$  is continuously differentiable (Assumption A1) for every  $i = 1, 2, \dots, M$ .

To illustrate the model selection procedure in the preceding section, we consider an example. We need to define the competing models, and the divergence type statistic to measure the departure of each proposed parametric model from the data generating process.

Here, we choose an important measure of divergence given by Rényi [25] which can be written in following form :

$$R^\alpha(f_{\theta_1}, f_{\theta_2}) = \frac{1}{\alpha(\alpha - 1)} \ln \left( \int_{\mathcal{X}} f_{\theta_1}^\alpha(x) f_{\theta_2}^{1-\alpha}(x) d\mu(x) \right); \quad \alpha \neq 0, 1$$

and limiting cases for  $\alpha = 0$  and  $\alpha = 1$ . That is,

$$R^0(f_{\theta_1}, f_{\theta_2}) = \lim_{\alpha \rightarrow 0} R^\alpha(f_{\theta_1}, f_{\theta_2}) = \int_{\mathcal{X}} f_{\theta_2}(x) \ln \frac{f_{\theta_2}(x)}{f_{\theta_1}(x)} d\mu(x)$$

and

$$R^1(f_{\theta_1}, f_{\theta_2}) = \lim_{\alpha \rightarrow 1} R^\alpha(f_{\theta_1}, f_{\theta_2}) = \int_{\mathcal{X}} f_{\theta_1}(x) \ln \frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} d\mu(x) = R^0(f_{\theta_2}, f_{\theta_1})$$

which is Kullback-Leibler divergence.

In case that  $f_{\theta_1}$  and  $f_{\theta_2}$  are discrete probability distributions, their Rényi's divergence is

$$(5) \quad R^\alpha(f_{\theta_1}, f_{\theta_2}) = \frac{1}{\alpha(\alpha - 1)} \ln \left( \sum_{x \in \Omega} f_{\theta_1}^\alpha(x) f_{\theta_2}^{1-\alpha}(x) \right); \quad \alpha \neq 0, 1$$

In statistical literature, the problem of choosing between the family of log-normal distributions and the family of exponential distributions has a long history. See [7] and [4] among others.

The log-normal distribution is parameterized by  $r = (r_1, r_2)$  and has density

$$h(x|r_1, r_2) = \begin{cases} \frac{1}{x(2\pi)^{1/2}r_2} \exp\left(-\frac{(\log x - r_1)^2}{2r_2^2}\right) & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The exponential distribution with parameter  $\beta$  has density

$$g(x|\beta) = \frac{1}{\beta} \exp(-x/\beta) \quad \text{for } x > 0$$

and zero otherwise.

The estimator used for each competing model is the maximum likelihood estimator (MLE). Specifically, for the log-normal model,

$$\hat{r}_1 = \frac{1}{n} \sum_{i=1}^n \log x_i \quad \text{and} \quad \hat{r}_2 = \sum_{i=1}^n (\log x_i - \hat{r}_1)^2$$

For the exponential model, the MLE is the sample average, i.e.,

$$\hat{\beta} = \frac{1}{n} \sum_{i=1}^n x_i$$

Lastly, we use the Rényi's divergence measure (5) to evaluate the discrepancy of a proposed model from the true data generating process. We partition the real line into  $M$  intervals  $\{(a_{i-1}, a_i), i = 1, \dots, M\}$  where  $a_i$  is a real number. The choice of the cells is discussed below. The Rényi statistic for the log-normal and exponential models are :

$$R^\alpha(f, h(r)) = \frac{1}{\alpha(\alpha - 1)} \ln \left( \sum_{i=1}^M h_i^\alpha(r) f_i^{1-\alpha} \right)$$

and

$$R^\alpha(f, g(\beta)) = \frac{1}{\alpha(\alpha - 1)} \ln \left( \sum_{i=1}^M g_i^\alpha(\beta) f_i^{1-\alpha} \right)$$

where  $h_i(r)$  and  $g_i(\beta)$  are probabilities of the interval  $(a_{i-1}, a_i)$  under  $h(x|r)$  and  $g(x|\beta)$  respectively, and  $f$  is the vector of observed cell probabilities defined in (4).

In our Monte Carlo study, we consider various sets of experiments in which the data are generated from a mixture of an exponential distribution and a log-normal distribution. These two distributions are calibrated so that have the same population means and variances, namely one and one. Hence the data generating process has density

$$d(p) = p \text{ Exponential } (1) + (1 - p) \text{ Log-normal } (-0.047, 0.5)$$

where  $p$  is set to some specific value for each set of experiments. In each set of experiments, several random samples are drawn from this mixture of distributions. The sample size varies from 100 to 1,000, and each sample size the number of replications is 1,000.

Throughout, the chosen partition has, four cells defined by the values  $a_0 = 1.0, a_1 = 1.5, a_2 = 2.0, a_3 = 3.0,$  and  $a_4 = +\infty$ . Similarly to the minimum Chi-square methods, note that because the log-normal distribution has two parameters, hence four is the minimum number of cells for which a perfect fit is not always achieved. Note also that the shapes of the log-normal and exponential densities differ greatly around the origin. This motivates the choice of  $a_0 = 1.0$ . The value  $\alpha = 0.5$  in (5) corresponds, approximatively, to the common density function in  $[1, +\infty[$  under the null hypothesis  $H_o$  (see figure 1-c).

We choose five different values for  $p$  which are : 0.00, 0.25, 0.41, 0.75 and 1.00. Although our proposed model selection procedure does not require that the data generating process belong to either of the competing models, we consider the two limiting cases  $p = 0.00$  and  $p = 1.00$  for they correspond to the correctly specified cases.

The value  $p = 0.410$  is determined to be the value for which the estimated log-normal distribution and the estimated exponential distribution are approximately at equal distance from the mixture  $d(p)$  according to Rényi's divergence. Thus this set of experiments corresponds approximately to the null hypothesis of our proposed model selection test  $DI_n$ . The results of our four sets of experiments are

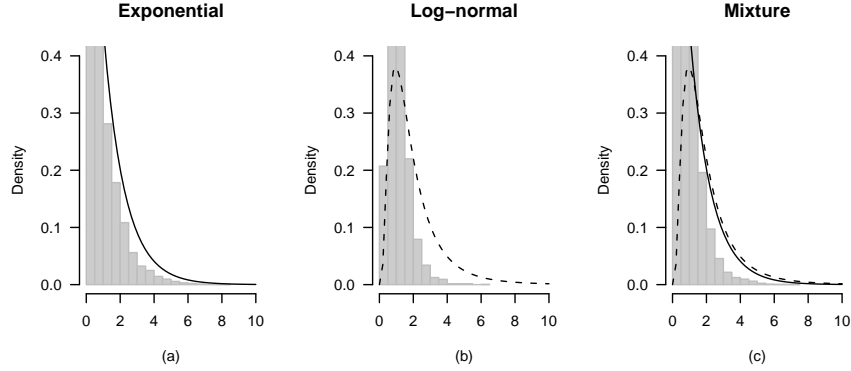


Figure 1 : Histogram and frequency of exponential, log-normal and mixture models with  $p = 0.41$  and  $\alpha = 0.5$

presented in Tables 1-5.

The first half of each table gives the average values of the ML estimators  $\hat{\beta}$ ,  $\hat{r}_1$ , and  $\hat{r}_2$ , the divergence goodness-of-fit statistics  $D_n(\hat{\theta}_n)$  and  $D_n(\hat{\gamma}_n)$ , and the model selection  $DI_n$  with its bootstrap estimated variance  $\hat{\omega}_*^2$ . The values in parentheses are standard errors. The second half of each table gives in percentage the number of times our proposed model selection procedures based on the method described in the previous section, favor the log-normal model, the exponential model, or are indecisive. The tests are conducted at the 5% nominal level.

In the two sets of experiments ( $p = 0$  and  $p = 1$ ) where one model is correctly specified, we use the labels "correct" and "incorrect" when a choice is made. This allows a comparison with the asymptotic  $N(0, 1)$  approximation under our null hypothesis of equivalence.

Data generating Process = Log-norm(-0.047, 0.5)				
n	100	300	600	1000
$\hat{\beta}$	0.927 (0.051)	0.925 (0.028)	0.924 (0.020)	0.925 (0.016)
$\hat{r}_1$	-0.046 (0.052)	-0.046 (0.021)	-0.047 (0.021)	-0.047 (0.016)
$\hat{r}_2$	0.497(0.035)	0.500 (0.021)	0.500 (0.014)	0.500 (0.011)
$\hat{\omega}_*$	1.413 (0.181)	1.383 (0.146)	1.325 (0.125)	1.303 (0.107)
$D_n(\hat{\theta}_n)$	3.699 (0.336)	3.636 (0.178)	3.617 (0.125)	3.616 (0.096)
$D_n(\hat{\gamma}_n)$	3.131(0.394)	3.103 (0.214)	3.087 (0.152)	3.092 (0.116)
$DI_n$	4.081 (1.187)	6.726 (1.082)	9.846 (1.121)	12.806 (1.286)
Incorrect	0 %	0 %	0 %	0 %
Indecisive	%	0 %	0 %	0 %
Correct	100 %	100 %	100 %	100 %

Table 1

	<b>DGP =0.25 Exp (1)+ 0.75 Log-norm (-0.047, 0.5)</b>			
n	100	300	600	1000
$\beta$	0.949 (0.061)	0.944 (0.038)	0.943 (0.026)	0.944 (0.019)
$r_1$	-0.181 (0.075)	-0,180 (0.046)	-0.179 (0.032)	-0.180 (0.042)
$r_2$	0.795(0.122)	0.804 (0.074)	0.802 (0.053)	0.806 (0.042)
$\omega$	1.349 (0.409)	1.265 (0.264)	1.233 (0.188)	1.240 (0.165)
$D_n(\hat{\theta})$	3.735 (0.336)	3.677 (0.198)	3.665 (0.135)	3.664 (0.103)
$D_n(\hat{\gamma})$	3.572 (0.385)	3.519 (0.225)	3.505 (0.154)	3.507 (0.117)
$DI_n$	1.295 (0.877)	2.609 (1.016)	3.250 (1.082)	4.091 (1.113)
Favor Exp	0 %	0 %	0 %	0 %
Indecisive	76 %	38 %	12 %	2 %
Favor Log-n	24 %	62 %	88 %	98 %

Table 2

	<b>DGP =0.445 Exp (1)+ 0.555 Log-norm (-0.047, 0.5)</b>			
n	100	300	600	1000
$\beta$	0.957 (0.069)	0.955 (0.040)	0.955 (0.028)	0.955 (0.021)
$r_1$	-0.263(0.090)	-0,263 (0.051)	-0.264 (0.035)	-0.265 (0.027)
$r_2$	0.932(0.131)	0.941 (0.074)	0.942 (0.055)	0.944 (0.042)
$\omega$	1.201 (0.343)	1.125 (0.198)	1.103 (0.162)	1.103 (0.132)
$D_n(\hat{\theta})$	3.775 (0.358)	3.712 (0.196)	3.710 (0.140)	3.706 (0.106)
$D_n(\hat{\gamma})$	3.771 (0.398)	3.711 (0.218)	3.711 (0.154)	3.708 (0.117)
$DI_n$	0.048 (0.896)	0.031 (0.921)	0.008 (0.908)	-0.044 (0.910)
Favor Exp	1%	1 %	1 %	1%
Indecisive	96 %	97 %	97 %	97 %
Favor Log-n	3 %	2%	2%	2%

Table 3

	<b>DGP =0.75 Exp (1)+ 0.25 Log-norm (-0.047, 0.5)</b>			
n	100	300	600	1000
$\beta$	0.986 (0.090)	0.981 (0.051)	0.981 (0.036)	0.980 (0.027)
$r_1$	-0.441(0.116)	-0,443 (0.067)	-0.445 (0.046)	-0.443 (0.036)
$r_2$	1.153(0.135)	1.158 (0.080)	1.159 (0.055)	1.158 (0.043)
$\omega$	0.947 (0.254)	0.853 (0.138)	0.840 (0.105)	0.835 (0.090)
$D_n(\hat{\theta})$	3.919 (0.416)	3.868 (0.229)	3.865 (0.105)	3.855 (0.121)
$D_n(\hat{\gamma})$	4.132 (0.436)	4.082 (0.239)	4.082 (0.164)	4.070 (0.127)
$DI_n$	-2.319 (0.902)	-4.426 (1.074)	-6.388 (1.076)	-8.238 (1.190)
Favor Exp	66 %	99 %	100%	100%
Indecisive	34 %	1 %	0 %	0%
Favor Log-n	0 %	0 %	0 %	0%

Table 4

<b>Data generating Process = Exponential(1)</b>				
n	100	300	600	1000
$\beta$	1.008 (0.105)	1.001 (0.059)	1.001 (0.040)	1.000 (0.031)
$r_1$	-0.570 (0.131)	-0.577 (0.076)	-0.576 (0.052)	-0.577 (0.040)
$r_2$	1.266(0.128)	1.284 (0.078)	1.280 (0.052)	1.282 (0.043)
$\omega$	0.840 (0.227)	0.757 (0.107)	0.738 (0.083)	0.735 (0.074)
$D_n(\hat{\theta})$	4.068 (0.466)	4.023 (0.250)	4.012 (0.179)	4.007 (0.138)
$D_n(\hat{\gamma})$	4.378 (0.465)	4.339 (0.250)	4.328 (0.178)	4.324 (0.137)
$DI_n$	-3.833 (1.040)	-7.300 (1.082)	-10.565 (1.195)	-13.745 (1.430)
Correct	100 %	100 %	100 %	100 %
Indecisive	0 %	0 %	0 %	0 %
Incorrect	0 %	0 %	0 %	0 %

Table 5

Tables 1 and 5, report the cases when one model is correctly specified. It is well-known that the MLE is consistent for the true parameter value under correct specification.

For example, in Table 1, the log-normal model is correctly specified, and the MLE of  $r = (r_1, r_2)$  approaches the true value  $r_o = (-0.047, 0.5)$  as the sample size increases from 100 to 1000. The bootstrap estimator of  $\omega$  also converges as the sample size becomes larger. The test statistic for model selection  $DI_n$  approximatively increases at a rate  $\sqrt{n}$ . In table 5, when the exponential model is correctly specified, one can observe similar results.

The second half of Table 1, summarizes the results for our model selection procedure. The method performs quite well and select the correct model almost 100% of the times, as expected.

For Tables 2, 3 and 4, the data was generated neither from the log-normal model nor from the exponential model, but from a mixture of these two models. Hence, the log-normal and the exponential model are both incorrectly specified.

In Table 3, the data generating process is chosen such that the log-normal model and the exponential model are approximatively equally close to it. The test statistic  $DI_n$  is expected to have a limiting standard normal  $N(0, 1)$ . This roughly confirmed in Table 3. For example, for  $n=1000$ ,  $DI_n$  has mean  $-0.044$  and standard error 0.910.

From our limited Monte Carlo study, one can observe that test statistic for model selection  $DI_n$  works relatively well, and fits equally well the data with a probability of around 95%.

## 6. Discusson

In summary, by analogy with the classical type chi-square statistics, we have introduced the divergence measures and propose some convenient asymptotically standard normal tests for model selection based on type divergence statistics that use estimators in a quite general class. The tests are designed to determine whether the estimated competing models are as close to the true distribution against the alternative hypothesis that one estimated model is closer, where closeness is measured according to discrepancy implicit in the divergence type statistic used. To determine the statistical divergence for the discrepancy between the observed data

and a specific parametric, computation has done by some numerical technique, by the help of Bootstrap methods, for evaluating the estimator of the asymptotic variance of our test statistic.

Several Monte Carlo experiments were conducted and showed that our procedure performs relatively well. Our work can be used to compare the power of tests statistics for model selection, based on some other type measures of information.

#### REFERENCES

- [1] Akaike H., (1973). Information theory and an Extension of the Likelihood Ratio Principe *Proceedings of the Second International Symposium of Information Theory, Ed. by Petrov, B.N. and Csaki, F. Budapest : Akademiai Kiado, 257-281.*
- [2] Andrews, D.W.K, (1967a) Chi-Square Diagnostic Tests for Econometric Models : Theory , *Econometrica*, **56** 1419-1453.
- [3] Andrews, D.W.K, (1988b) Chi-Square Diagnostic Tests for Econometric Models : Introduction and Applications, *Journal of Econometrics* **37** 135-156.
- [4] Atkinson A.C, (1970) A Method for Discriminating Between Models, *Journal of Royal Statistical Society, Series B* **32** 323-353.
- [5] Bar-Hen, A. and J.J Daudin (1995),Generalization of the Mahalanobis distance in the mixed case. *Journal of Multivariate Analysis*, **53**, 332-342.
- [6] Cochran, W. G. The  $\chi^2$  Test of goodness of fit, *Ann. Math. Statist.*, **23**, 315-345.
- [7] Cox, D.R ()(1962) Further Esults on Tests of Separate Families of Hypotheses, *Journal of the Royal Statistical Society, Series B*, **24** 406-421.
- [8] Cressie, N. and Read, T.R.C.(1984),Multinomial goodness of fit tests. *Journal of the Royal Statistical Society, Series B*, **46**, 440-464.
- [9] Csiszar I., (1967)Information-type measures of difference of probability distributions and indirect observations *Studia Sci. Math. Hung.*, 299-318.
- [10] Efron., (1982) The Jackknife, the bootstrap and Other Resampling Plans. *CBMS-NSF Regional Conference Series in Applied Mathematics* **38**.
- [11] Jeffrey H., (1946) Theory of probability, *Univ. Oxford, London*.
- [12] J. Burbea (1984) The Bose-Einstein entropy of degree  $\alpha$  and Jensen difference, *Utilitas Math.*, **26** 171-192
- [13] Kagan M. (1963) On the theory of Fisher's amount information, *Sov. Math. Dokl*,**4**, 99-993.
- [14] Kullback S., Leibler (1951) On the information and Sufficiency *Ann. Math. Statist.* **22** 79-86.
- [15] ,Liese, F. and Vajda I. (1987),Convex Statistical Distances. *Teubner, Leipzig.*,
- [16] Menéndez M. L., Morales, D. Pardo, and Salicrú, M. (1997) , Divergences measures between populations : applications in the exponential family. *Communications in Statistics (Theory and Methods)*, **25**, 1099-1117.
- [17] Moore D.S,(1977) , Generalized Inverses, Wald's Method and the Construction of Chi-Squared Tests of fit. *Journal of Statistical Association*, **7**, 131-137.
- [18] Moore D.S,(1984) , Measures of lack of fit from Tests of Chi-Squared Type. *Journal of Statistical Planning and Inference*, **7**, 131-137.
- [19] Morales D., Pardo L., and Vajda, I. (1997) , Some new statistics for testing hypotheses in parametric models. *Journal of Multivariate Analysis*, **10**, 151-166.
- [20] Morales D., Pardo L., and Zografos K.,(1998), Informational distances and related statistics in mixed continuous and categorical variables. *Journal of Statistical Planning and Inference*, **75**, 47-63.
- [21] Morales D., Pardo L., (2001), Some approximations to power functions of  $\phi$ -Divergences tests in parametric models. *Test* , **10**, 249-269.
- [22] Nayak, T. K. (1985) , On diversity measures based on entropy functions. *Communications in Statistics (Theory and Methods)*, **14**, 203-215.
- [23] Pardo, L. Salicrú, M. Menendez, M.L and Morales, D (1995) , Divergence mesures based on entropy functions and statistical inference . *Sankhya, Series B*, **57**, 315-337.
- [24] Pardo L., Morales D., Salicrú M., Menendez (1994). Asumptotic properties of divergence statistics in a stratified random sampling and its applications to test statistical hypotheses, *Journal of Statistical Planning and Inference*, **38** 201-222.

- [25] Rényi A. (1961): On measures of entropy and information. *Proc. 4<sup>slth</sup> Berkeley Symp. on Math. Statist. Univ. Calif. Press, Berkeley.* **1**, 547-561.
- [26] Salicrú, M. Menendez,Pardo, L. and Morales, D (1994) , On the applications of divergence type mesures in testing statistical hypoteses. *Journal of Multivariate Analysis*, **51**, 372-391.
- [27] Serfling, R. J. (1980). Approximations Theorems of Mathematical Statistics. *John Wiley, New York.*
- [28] Sharma B.D, D.P Mittal (1977) New nonadditive measures of entropy for discrete probability distributions, *J. Math. Sci.*, **10**, 28-40.
- [29] Taneja I. J.(1987) Statistical aspects of divergence measures, *Journal of Statistical Planning and Inference*, **16**, 136-145.
- [30] Taneja I. J.(1989) On generalized information measures and their applications, *Adv. Electron . Phys.*, **76**, 327-413.
- [31] Vadja I. (1973).  $\chi^2$ -divergence and generalized Fisher's information, *Trans. 6th Prague Conf. on Inform. Theory Statistical Decision Functions and Random Process, Prague*, 873-886.
- [32] Vuong, Q., Weiren, W., (1993) Selecting Estimated Models Using Chi-Square Statistics, *Annals D'Economie et de Statistique*, **30**, 144-164.
- [33] Zografos, K. Ferentinos, K. and Papaioannou, T. (1990).  $\phi$ -Divergence statistics : sampling properties and multinomial goodness of fit and divergence tests. *Communication in Statistics (Theory and Methods)*, **19**, 1785-1802.
- [34] Watson, G.S, (1959) Some Recent Results in Chi-Square Goodness-of-fit Tests, *Biometrics*, **15**, 440-468.