Far East Journal of Theoretical Statistics

http://www.pphmj.com
http://dx.doi.org/10.17654/TS054020189
Volume 54, Number 2, 2018, Pages 189-224 **ISSN: 0972-0863**

IMPROVED ESTIMATORS OF BREGMAN DIVERGENCE FOR MODEL SELECTION IN SMALL SAMPLES

Papa Ngom[a,b,*], Jean de Dieu Nkurunziza[b] and Carlos Ogouyandjou[b]

[a]LMA
Université Cheikh Anta Diop
Dakar, Senegal

[b]Institut de Mathématiques et de Sciences Physiques
Porto Novo, Benin

Abstract

Recently in [1, 2], Bromideh introduced the Kullback-Leibler Divergence (KLD) test statistic in discriminating between two models. It was found that the ratio minimized Kullback-Leibler divergence (RMKLD) works better than the ratio of maximized likelihood (RML) for small sample size. The aim of this paper is to generalize the works of Ali-Akbar Bromideh by proposing a hypothesis testing based on Bregman Divergence (BD) in order to improve the process of choice of the model. We investigate the problem of model choice and propose a unified method for model selection and estimation procedure with desired theoretical properties and computational convenience. After observing n data points of unknown density f; we firstly measure the closeness between the bias reduced kernel density estimator and the first estimated candidate model. Secondly between the bias reduced

Received: November 26, 2017; Accepted: December 20, 2017
2010 Mathematics Subject Classification: 62F10, 62F12, 62G07, 62G10, 62G20.
Keywords and phrases: a bias reduced kernel estimator, Bregman divergence, hypothesis test.
*Corresponding author

kernel density estimator and the second estimated candidate model. In these two cases, BD and the bias reduced kernel estimator [3] focuses on improving the convergence rates of kernel density estimators are used. We establish the asymptotic properties of BD estimator and approximations of the power functions are deduced. The multi-step MLE process will be used to estimate the parameters of the models. We explain the applicability of the BD by a real data set and by the data generating process (DGP). The Monte Carlo simulation and then the numerical analysis will be used to interpret the result.

1. Introduction

Bregman introduced for convex functions [4-7], the nonnegative measure of dissimilarity. His motivation was the problem of convex programming, but in the subsequent literature it became widely applied in many other problems under the name Bregman distance in spite of that it is not in general the usual metric distance (it is a pseudo-distance which is reflexive but neither symmetric nor satisfying the triangle inequality). In the last decade, BD has become an important tool in many research areas. For instance, several specific BD, such as Itakura-Saito Distance [8], Kullback-Leibler Divergence (KLD) [9], and Mahalanobis Distance [10] have been used in machine learning as the distortion functions (or loss functions) for clustering tasks. These divergences have been used in generalizations of principal component analysis to data with distributions belonging to the exponential family. Although, the goodness-of-fit and significance testing is initially used in selection of two probability densities; many models selection criteria have been proposed so far. Classical model selection criteria using least square error and log-likelihood include the C_p-criterion, cross-validation (CV), the Akaike Information Criterion (AIC) based on the well-known Kullback-Leibler divergence, Bayesian Information Criterion (BIC), a general class of criteria that also estimates the Kullback-Leibler Divergence (KLD). These criteria have been proposed by Mallows [11], Akaike [12], Schwarz [13] and Konishi and Kitagawa [14], Toma [15], respectively. Mohd Saat et al. [16] compared RML with Vuong's closeness test [17] to discriminant between Gamma and Weibull, in which they found both methods relatively similar.

Basu et al. [18] compared RML with Kolmogorov-Smirnov and chi-squared (with asymptotic properties) and some inconsistency among them are reported. Despite of significant amount of work on discrimination measures by different methods, in the study of the problem of model selection with divergence type statistics, it is convenient to use some convenient asymptotically standard tests based on Pearson chi-square statistics. A well known 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. In order to circumvent such a difficulty, a popular method for model selection, which is similar to use of Akaike Information Criterion (AIC), consists in considering that the lower the chi-square statistic, the better is the model. The preceding selection rule, however, does not take into account random variations inherent in the values of the statistics. The purpose of this paper is to generalize the works of Ali-Akbar [1, 2] by proposing a hypothesis testing based on Bregman divergence in order to improve the process of choice of the model. Our model selection approach differs from him. We propose here a procedure for taking into account the stochastic nature of these differences so as to assess their significance. The testing procedure for model selection will be based on the comparison of the value of Bregman type statistic to critical values from a standard normal table. The procedures considered here are testing the null hypothesis that the competing models are equally close to the observed data versus the alternative hypothesis that one model is closer to the real data, where closeness of a model is measured according to the discrepancy implicit in the Bregman divergence type statistic used.

The rest of the paper is organized as follows: We give some basic notation and general results in Section 2. In Section 3, theoretical properties of the Bregman divergence estimator are obtained. Applications for testing hypothesis are given in Section 4. Numerical studies are presented in Section 5 and finally the conclusion appears in Section 6.

2. Basic Notation and Some General Results

2.1. Definitions and notation

Let $(\mathcal{X}, \beta_{\mathcal{X}}, F)$ be the statistical space associated with the support $\mathcal{X} = \{1, 2, ..., M_0\}, \forall M_0 \geq 1$; $\beta_{\mathcal{X}}$ is the σ-algebra defined on $\mathcal{X}$ and $(\mathcal{X}, \beta_{\mathcal{X}})$, the measurable space.

Let

$$\Lambda_{M_0} = \left\{ F = (f_1, ..., f_{M_0})^T ; \forall x \in \mathbb{R}\ f_i(x) \geq 0, i = 1, ..., M_0 \text{ and } \sum_{i=1}^{M_0} f_i(x) = 1 \right\}$$

be the simplex of probability M_0-vectors. One can define the parametric family of models as follows:

$$\mathcal{F} = \{F_\theta = (f_1(\cdot, \theta), ..., f_{M_0}(\cdot, \theta))^T : \theta \in \Theta\},$$

where Θ is a compact subset of k-dimensional Euclidean space $(k < M_0 - 1, \forall M_0 \geq 1)$.

We assume that the probability distribution F_θ is absolutely continuous with respect to a σ-finite measure μ on $(\mathcal{X}, \beta_{\mathcal{X}})$. For simplicity μ is either the Lebesgue measure or a counting measure. The parametric family of models may or may not contain the true model. If $\mathcal{F}$ contains the true model, then there exists a $\theta_0 \in \Theta$ such that $F_{\theta_0} = F$ and the model F_θ is said to be c*orrectly specified*. We are interested in testing

$$H_0 : F = F_{\theta_0} \text{ versus } H_1 : F \neq F_{\theta_0}. \quad (1)$$

Note that $F(x) = (f_1(x), ..., f_{M_0}(x))^T$ can be estimated by a bias reduced kernel estimator based on a random sample of size n; $X_1, ..., X_{M_0}$. In this following section, we present the brief review of this estimator.

2.2. A bias reduced kernel estimator

Kernel density estimator was first introduced by Rosenblatt [19] and Parzen [20]. Suppose that $X_1, ..., X_n$ is a simple random sample from the unknown density function f. Let K be a function on real line, i.e., the "kernel", and let h be a positive value, i.e., the "bandwidth". Then the kernel density estimator of f is defined as:

$$\hat{f}_{n,h}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right). \tag{2}$$

To make the estimator meaningful, we introduce a measurable function $K(\cdot)$ that satisfies the following conditions:

(K.1) $K(\cdot)$ is of bounded variation on $\mathbb{R}$,

(K.2) $K(\cdot)$ is right continuous on $\mathbb{R}$,

(K.3) $\| K \|_\infty = \sup_{x\in\mathbb{R}} | K(x) | < \infty$,

(K.4) $\int_{\mathbb{R}} K(t)dt = 1$.

Under the regularity conditions on $K(\cdot)$, let f be twice continuously differentiable in a neighbourhood of x. Then

$$Bias(\hat{f}_{n,h}(x)) = E(\hat{f}_{n,h}(x)) - f(x) = \frac{h^2}{2} f''(x)\int u^2 K(u)du + o(h^2) \tag{3}$$

and

$$Var(\hat{f}_{n,h}(x)) = \frac{1}{nh} f(x)\int K^2(u)du + o((nh)^{-1}). \tag{4}$$

Then from (3) and (4), we have

$$\begin{aligned} MSE(\hat{f}_{n,h}(x)) &= Bias^2(\hat{f}_{n,h}(x)) + Var(\hat{f}_{n,h}(x)) \\ &= \frac{1}{4}(f''(x))^2 h^4\left[\int u^2 K(u)du\right]^2 \\ &\quad + \frac{1}{nh} f(x)\int K^2(u)du + o(h^4 + (nh)^{-1}). \end{aligned}$$

Devroye and Gyorfi [21] showed that the optimized bandwidth is $h \sim O(n^{-\frac{1}{5}})$ and then the optimal MSE is of the order $n^{-\frac{4}{5}}$. Xie and Wu [3] introduced a new type of density estimator in order to reduce bias, investigated and calculated its bias, variance and MSE which show some improvement over the ordinary kernel density estimator. Since the leading term of the bias is unavailable due to the unknown f, we can simply use its estimation to reduce the bias of the ordinary kernel density estimator, i.e.,

$$\hat{f}^{b}_{n,h}(x) = \hat{f}_{n,h}(x) - \widehat{Bias}(\hat{f}_{n,h}(x)).$$

As result, the proposed estimator is

$$\hat{f}^{b}_{n,h}(x) = \hat{f}_{n,h}(x) - \frac{h^2}{2}\hat{f}''_{n,h}(x)\int u^2 K(u)du$$

$$= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) - \frac{1}{2nh}\sum_{i=1}^{n} K''\left(\frac{x - X_i}{h}\right)\int u^2 K(u)du, \quad (5)$$

where $\hat{f}_{n,h}(x)$ is given by (2). From the way of construction, this new estimator should be able to reduce the bias and thus the MSE. To see whether this is the case or not, they next calculated the bias and the variance of $\hat{f}^{b}_{n,h}$. These following regularity conditions on f, K and h are in need:

(1) $\int uK(u) = 0$,

(2) f is fourth differentiable in a neighbourhood of x,

(3) $h \to 0$ and $nh \to \infty$ as $n \to \infty$.

Theorem 1 (Xie and Wu [3])**.** *Under* (1), (2) *and* (3),

$$Bias(\hat{f}^{b}_{n,h}(x)) = -\frac{h^3}{6} f'''(x)\int u^3 K(u)du + o(h^3) \quad (6)$$

and

$$Var(\hat{f}^{b}_{n,h}(x)) = \frac{1}{2nh} f(x)\left(\int u^2 K(u)du\right)^2 \int K''(u)du + o((nh)^{-1}). \quad (7)$$

Consequently,

$$MSE(\hat{f}_{n,h}^{b}(x)) = o(h^6 + (nh)^{-1}).$$

Under the regularity conditions on f, K and h, the optimal MSE is of the order $O(n^{-\frac{6}{7}})$ *with* $h = O(n^{-\frac{1}{7}})$.

2.3. A brief review of Bregman divergence

Divergences are distance-like functions, widely used to assess the similarity between two objects. Several authors proposed generalized divergences which encompass these classical divergences:

(1) Csiszar's divergence [22], which is a generalization of Amari's α-divergence [23]. Both these divergences encompass the Kullback-Leibler (KL) divergence and its dual.

(2) Bregman divergence [24, 25], which encompasses the Euclidean (EUC) distance, the KL divergence and the Itakura Saito (IS) divergence.

As a distance, a divergence should be nonnegative and separable. However, a divergence does not necessarily satisfy the triangle inequality and the symmetry axiom of a distance.

Bregman (see [4-7]) introduced for a convex subset of a Hilbert space $\mathbf{S}$ and $\phi : \mathbf{S} \to \mathbb{R}$ a continuously differentiable strictly convex function; the Bregman divergence $D_{\phi}^{B} : \mathbf{S} \times \mathbf{S} \to \mathbb{R}_{+}$ as follows:

$$D_{\phi}^{B}(p, q) = \phi(p) - \phi(q) - \langle p - q, \nabla\phi(q)\rangle, \forall (p, q) \in \mathbf{S}^2, \tag{8}$$

where $\nabla\phi(y)$ stands for the gradient of ϕ evaluated at y and $\langle \cdot, \cdot \rangle$ is the standard Hermitian dot product. Thus, the Bregman divergences between two probability density functions *f* and *g* is given by

$$D_{\phi}^{B}(f(x), g(x)) := \int_{\mathbf{X}} (\phi(f(x)) - \phi(g(x)) - (f(x) - g(x))\phi'(g(x)))dx, \tag{9}$$

where $\mathbf{X}$ is a support of the two density functions, f and g; and $\phi'(t)$ the derivative of $\phi(t)$ respected to t. On the other hand, Basu et al. [26], Minami and Eguchi [27] introduced the basic beta-divergence and many researchers investigated their applications including [28, 29]. The main motivation was to develop the link between beta-divergence and Bregman divergence.

It is also interesting to note that, the beta-divergence has to be defined in limiting case for $\beta \to 0$ as the Itakura-Saito distance and for $\beta \to 1$ as the KL-divergence. For $\beta \to 2$, we obtain the standard squared Euclidean $(L_2\text{-norm})$ distance. Therefore, one can check that the beta-divergence can be generated from the Bregman divergence using the following strictly convex continuous function [28, 30]:

$$\phi(t) = \begin{cases} \dfrac{c_1}{\beta(\beta-1)} t^{\beta} + c_2 t + c_3, \beta \neq 0, 1, \\ c_1 t \log(t) + c_2 t + c_3, \beta = 1, \\ -c_1 \log(t) + c_2 t + c_3, \beta = 0. \end{cases} \tag{10}$$

Here c_1, c_2 and c_3 are some constants.

Theorem 2 (Liese and Vajda [31])**.** *If the Bregman divergence* $D_{\phi}^{B}(p, q)$ *satisfies the homogeneity condition*[1] *with* $\beta = \begin{cases} 2, \\ 1, \\ 0, \end{cases}$ *then* $\phi(t) = \begin{cases} t^2, \\ t\log(t), \\ -\log(t) \end{cases}$ *the statements hold modulo affine functions.*

Assuming that the density f, we estimate $D_{\phi}^{B}(f(x), f_{\theta}(x))$ by

[1] $D_{\phi}^{B}(kp, kq) = k^{\beta} D_{\phi}^{B}(p, q)$, for all $p, q, k > 0$ with β equal to 2, 1, 0.

$$\hat{D}_{\phi}^{B}(\hat{f}_{n,h}^{b}(x),\, f_{\theta}(x))$$

$$= \int_{A_n} \phi(\hat{f}_{n,h}^{b}(x)) - \phi(f_{\theta}(x)) - (\hat{f}_{n,h}^{b}(x) - f_{\theta}(x))\phi'(f_{\theta}(x))dx, \qquad (11)$$

where

$$A_n = \{x \in \mathbb{R}^{M_0},\, \hat{f}_{n,h}^{b}(x) \geq \gamma_n\}$$

and $\gamma_n \to 0$ is a sequence of positive constants. In this following section, we will assume $\int_{A_n} dx, \int_{A_n} \phi'(\hat{f}_{n,h}^{b}(x))dx$ and $\int_{A_n} \phi'(f_{\theta}(x))dx$ to be finite and we will use the methods developed in [32] to establish convergence results for our estimator $\hat{D}_{\phi}^{B}(\hat{f}_{n,h}^{b}(x),\, f_{\theta}(x))$.

3. Theoretical Properties of the Bregman Divergence Estimator

For proving such consistency results, one usually writes the difference $\hat{D}_{\phi}^{B}(\hat{f}_{n,h}^{b}(x),\, f_{\theta}(x)) - D_{\phi}^{B}(f(x),\, f_{\theta}(x))$ as the sum of a probabilistic term $\hat{D}_{\phi}^{B}(\hat{f}_{n,h}^{b}(x),\, f_{\theta}(x)) - \mathbb{E}\hat{D}_{\phi}^{B}(\hat{f}_{n,h}^{b}(x),\, f_{\theta}(x))$, and a deterministic term $\mathbb{E}\hat{D}_{\phi}^{B}(\hat{f}_{n,h}^{b}(x),\, f_{\theta}(x)) - D_{\phi}^{B}(f(x),\, f_{\theta}(x))$, the so-called bias. Throughout the remainder of this paper, $\mathbb{E}\hat{D}_{\phi}^{B}(\hat{f}_{n,h}^{b}(x),\, f_{\theta}(x))$ is given by

$$\mathbb{E}\hat{D}_{\phi}^{B}(\hat{f}_{n,h}^{b}(x),\, f_{\theta}(x))$$

$$:= \int_{A_n} [\mathbb{E}\phi(\hat{f}_{n,h}^{b}(x)) - \phi(f_{\theta}(x)) - (\mathbb{E}\hat{f}_{n,h}^{b}(x) - f_{\theta}(x))\phi'(f_{\theta}(x))]dx,$$

where A_n is defined in (11).

Lemma 1. *Let* $K(\cdot)$ *satisfy* (*K*1)-(*K*4), *let* $f(\cdot)$ *be a continuous bounded density*, ϕ *be strictly convex function and assume that* ϕ *is linear and satisfies the Jensen inequality*. *Then*, *for each pair of sequence* $(a_n)_{n\geq 1}$,

$(b_n)_{n\geq 1}$ *such that* $0 < a_n < b_n \leq 1$ *with* $b_n \to 0$ *and* $na_n/\log(n) \to \infty$ *as* $n \to \infty$, *we have with probability* 1

$$\sup_{a_n \leq h \leq b_n} | \hat{D}_\phi^B(\hat{f}_{n,h}^b(x), f_\theta(x)) - \mathbb{E}\hat{D}_\phi^B(\hat{f}_{n,h}^b(x), f_\theta(x)) |$$

$$= 0\left(\sqrt{\frac{\log(1/a_n) \vee \log\log n}{na_n}}\right).$$

Proof. Define

$$\Delta_{n1} := \hat{D}_\phi^B(\hat{f}_{n,h}^b(x), f_\theta(x)) - \mathbb{E}\hat{D}_\phi^B(\hat{f}_{n,h}^b(x), f_\theta(x)).$$

We have

$$| \Delta_{n1} | = \left| \int_{A_n} [\phi(\hat{f}_{n,h}^b(x)) - \phi(f_\theta(x)) - (\hat{f}_{n,h}^b(x) - f_\theta(x))\phi'(f_\theta(x))] \right.$$

$$\left. + [\mathbb{E}\phi(\hat{f}_{n,h}^b(x)) - \phi(f_\theta(x)) - (\mathbb{E}\hat{f}_{n,h}^b(x) - f_\theta(x))\phi'(f_\theta(x))]dx \right|$$

$$\leq \left| \int_{A_n} (\phi(\hat{f}_{n,h}^b(x)) - \mathbb{E}\phi(\hat{f}_{n,h}^b(x)))dx \right|$$

$$+ \left| \int_{A_n} (\hat{f}_{n,h}^b(x) - \mathbb{E}\hat{f}_{n,h}^b(x))\phi'(f_\theta(x))dx \right|,$$

ϕ verifies the Jensen inequality, i.e., $\mathbb{E}(\phi(\hat{f}_{n,h}^b(x))) \geq \phi(\mathbb{E}\hat{f}_{n,h}^b(x))$ and ϕ is linear, i.e., $\phi(\hat{f}_{n,h}^b(x)) + \phi(\mathbb{E}\hat{f}_{n,h}^b(x)) = \phi(\hat{f}_{n,h}^b(x) + \mathbb{E}\hat{f}_{n,h}^b(x))$. Therefore

$$| \Delta_{n1} | \leq \phi(\sup_{a_n \leq h \leq b_n} | \hat{f}_{n,h}^b(x) - \mathbb{E}\hat{f}_{n,h}^b(x) |)$$

$$\times \int_{A_n} dx + \sup_{a_n \leq h \leq b_n} | \hat{f}_{n,h}^b(x) - \mathbb{E}\hat{f}_{n,h}^b(x) | \int_{A_n} \phi'(f_\theta(x))dx.$$

For $0 < a_n < b_n \leq 1$, we have

$$\sup_{a_n \le h \le b_n} | f_{n,h}(x) - \mathbb{E}\hat{f}^b_{n,h}(x) | \le \| \hat{f}^b_{n,h}(x) - \mathbb{E}\hat{f}^b_{n,h}(x) \|_\infty,$$

where $\| \cdot \|_\infty$ denotes, the supremum norm, i.e., $\| \psi \|_\infty := \sup_{x \in \mathbb{R}} | \psi(x) |$.

Therefore,

$$| \Delta_{n1} | \le \phi(\| \hat{f}^b_{n,h}(x) - \mathbb{E}\hat{f}^b_{n,h}(x) \|_\infty) \times \int_{A_n} dx + \| \hat{f}^b_{n,h}(x) - \mathbb{E}\hat{f}^b_{n,h}(x) \|_\infty \int_{A_n} \phi'(f_\theta(x)) dx.$$

Finally

$$\sup_{a_n \le h \le b_n} | \Delta_{n1} | \le \phi(\sup_{a_n \le h \le b_n} \| \hat{f}^b_{n,h}(x) - \mathbb{E}\hat{f}^b_{n,h}(x) \|_\infty) \times \int_{A_n} dx + \sup_{a_n \le h \le b_n} \| \hat{f}^b_{n,h}(x) - \mathbb{E}\hat{f}^b_{n,h}(x) \|_\infty \times \int_{A_n} \phi'(f_\theta(x)) dx.$$

Whenever $K(\cdot)$ is measurable and satisfies (K3)-(K4) and by the remark 2 in [33], when $f(\cdot)$ is bounded, for each pair of sequence $(a_n)_{n \ge 1}$ and $(b_n)_{n \ge 1}$ such that $0 < a_n < b_n \le 1$ with $b_n \to 0$ and $na_n / \log(n) \to \infty$ as $n \to \infty$, we have with probability 1

$$\sup_{a_n \le h \le b_n} \| \hat{f}^b_{n,h}(x) - \mathbb{E}\hat{f}^b_{n,h}(x) \|_\infty = 0\left(\sqrt{\frac{\log(1/a_n) \vee \log\log n}{na_n}} \right). \qquad (12)$$

Since $\int_{A_n} dx < \infty$ and $\int_{A_n} \phi'(f_\theta(x)) dx < \infty$, in view of (12) and the preceding equation, we obtain with probability 1,

$$\sup_{a_n \le h \le b_n} | \Delta_{n1} | = 0\left(\sqrt{\frac{\log(1/a_n) \vee \log\log n}{na_n}} \right) + 0\left(\sqrt{\frac{\log(1/a_n) \vee \log\log n}{na_n}} \right).$$

Thus

$$\sup_{a_n \le h \le b_n} | \Delta_{n1} | = 0\left(\sqrt{\frac{\log(1/a_n) \vee \log\log n}{na_n}}\right). \tag{13}$$

It concludes the proof of the lemma.

Lemma 2. *Let* $K(\cdot)$ *satisfy* $(K1)$-$(K4)$, *let* $f(\cdot)$ *be a continuous bounded density*, ϕ *be strictly convex function. Assuming that* ϕ *linear and satisfies the Jensen inequality. Then, for each pair of sequence* $(a_n)_{n\ge1}, (b_n)_{n\ge1}$ *such that* $0 < a_n < b_n \le 1$ *with* $b_n \to 0$ *as* $n \to \infty$,

$$\sup_{a_n \le h \le b_n} | \mathbb{E}\hat{D}_\phi^B(\hat{f}_{n,h}^b(x), f_\theta(x)) - D_\phi^B(f(x), f_\theta(x)) | = 0(b_n).$$

Proof. Let $\Delta_{n2} = \mathbb{E}\hat{D}_\phi^B(\hat{f}_{n,h}^b(x), f_\theta(x)) - D_\phi^B(f(x), f_\theta(x))$, therefore

$$| \Delta_{n2} | = \left| \int_{A_n} [\mathbb{E}\phi(\hat{f}_{n,h}^b(x)) - \phi(f_\theta(x)) - (\mathbb{E}\hat{f}_{n,h}^b(x) - f_\theta(x))\phi'(f_\theta(x))]dx \right.$$

$$\left. + \int_{A_n} [\phi(f(x)) - \phi(f_\theta(x)) - (f(x) - f_\theta(x))\phi'(f_\theta(x))]dx \right|.$$

Repeat the arguments above in the terms $|\Delta_{n1}|$ with the formal change of $\hat{f}_{n,h}^b$ by f. We show that, for any $n \ge 1$,

$$| \Delta_{n2} | \le \phi(\sup_{a_n \le h \le b_n} | \hat{f}_{n,h}^b(x) - f(x) |)$$

$$\times \int_{A_n} dx + \sup_{a_n \le h \le b_n} | \hat{f}_{n,h}^b(x) - f(x) | \int_{A_n} \phi'(f_\theta(x))dx,$$

which implies

$$| \Delta_{n2} | \le \phi(\sup_{a_n \le h \le b_n} | \hat{f}_{n,h}^b(x) - f(x) |)$$

$$\times \int_{A_n} dx + \sup_{a_n \le h \le b_n} | \hat{f}_{n,h}^b(x) - f(x) | \int_{A_n} \phi'(f_\theta(x))dx. \tag{14}$$

In [33], when the density $f(\cdot)$ is uniformly continuous, we have for each pair of sequence $(a_n)_{n\geq 1}, (b_n)_{n\geq 1}$ such that $0 < a_n < b_n \leq 1$, with $b_n \to 0$ as $n \to \infty$,

$$\sup_{a_n \leq h \leq b_n} \| \hat{f}^b_{n,h}(x) - f(x) \|_\infty = 0(b_n). \tag{15}$$

Thus,

$$\sup_{a_n \leq h \leq b_n} |\Delta_{n2}| \leq \phi(\sup_{a_n \leq h \leq b_n} \| \hat{f}^b_{n,h}(x) - f(x) \|_\infty)$$

$$\times \int_{A_n} dx + \sup_{a_n \leq h \leq b_n} \| \hat{f}^b_{n,h}(x) - f(x) \|_\infty \int_{A_n} \phi'(f_\theta(x))dx,$$

where $\int_{A_n} dx$ and $\int_{A_n} \phi'(f_\theta(x))dx$ are finite. Then, in view of (15)

$$\sup_{a_n \leq h \leq b_n} |\Delta_{n2}| = 0(b_n) + 0(b_n).$$

Finally,

$$\sup_{a_n \leq h \leq b_n} |\Delta_{n2}| = 0(b_n) \tag{16}$$

is deduced the proof of the lemma.

Theorem 3. *Let* $K(\cdot)$ *satisfy* (*K*3)-(*K*4), $f(\cdot)$ *be a uniform, bounded and continuous density and* ϕ *be strictly convex function. Assume that* ϕ *linear and satisfies the Jensen inequality. Then, for each pair of sequence* $(a_n)_{n\geq 1}$, $(b_n)_{n\geq 1}$ *such that* $0 < a_n < b_n \leq 1$ *with* $b_n \to 0$ *and* $na_n/\log(n) \to \infty$ *as* $n \to \infty$, *we have with probability* 1

$$\sup_{a_n \leq h \leq b_n} | \hat{D}^B_\phi(\hat{f}^b_{n,h}(x), f_\theta(x)) - D^B_\phi(f(x), f_\theta(x)) |$$

$$= 0\left(\sqrt{\frac{\log(1/a_n) \vee \log\log n}{na_n}} \vee b_n \right).$$

Proof. We have

$$\begin{aligned}
&| \hat{D}_\phi^B(\hat{f}_{n,h}^b(x),\, f_\theta(x)) - D_\phi^B(f(x),\, f_\theta(x)) | \\
&\leq | \hat{D}_\phi^B(\hat{f}_{n,h}^b(x),\, f_\theta(x)) - \mathbb{E}\hat{D}_\phi^B(\hat{f}_{n,h}^b(x),\, f_\theta(x)) | \\
&\quad + | \mathbb{E}\hat{D}_\phi^B(\hat{f}_{n,h}^b(x),\, f_\theta(x)) - D_\phi^B(f(x),\, f_\theta(x)) |.
\end{aligned}$$

Combining Lemmas 1 and 2, we obtain

$$\begin{aligned}
&\sup_{a_n \leq h \leq b_n} | \hat{D}_\phi^B(\hat{f}_{n,h}^b(x),\, f_\theta(x)) - D_\phi^B(f(x),\, f_\theta(x)) | \\
&= 0\left(\sqrt{\frac{\log(1/a_n) \vee \log\log n}{na_n}}\right) + 0(b_n). \qquad (17)
\end{aligned}$$

This entails that, as $n \to \infty$,

$$\sup_{a_n \leq h \leq b_n} | \hat{D}_\phi^B(\hat{f}_{n,h}^b(x),\, f_\theta(x)) - D_\phi^B(f(x),\, f_\theta(x)) | \to 0.$$

It concludes the proof of the theorem.

4. Applications for Testing Hypothesis

4.1. Test for goodness-of-fit

Recall the hypothesis testing (1) written as follows:

$$H_0 : F_\theta = F \ \ against \ H_1 : F_\theta \neq F.$$

Note that for simplicity, we have omitted 0 on θ. We have to reject the null hypothesis iff $D_\phi^B(\hat{F}_{n,h}^b, F_{\hat{\theta}}) > d$, where d has to be chosen for getting a level α test. In some situations, it will be possible to get the exact distribution of the statistic $D_\phi^B(\hat{F}_{n,h}^b, F_{\hat{\theta}})$ and then the value d. But in general this is not possible and we have to use the asymptotic distribution of the statistic $D_\phi^B(\hat{F}_{n,h}^b, F_{\hat{\theta}})$. In this following theorem, we present this asymptotic distribution.

Theorem 4. *Let* $D_\phi^B(F, F_\theta)$ *be the Bregman divergence and let* $\hat{D}_\phi^B(\hat{F}_{n,h}^b, F_{\hat{\theta}})$ *be its estimator. Under the null hypothesis* $H_0 : F_\theta = F$, *we have*

$$2n\hat{D}_\phi^B(\hat{F}_{n,h}^b, F_{\hat{\theta}}) \xrightarrow{\mathcal{L}} \sum_{i=1}^{r} \beta_i Z_i^2 + \sum_{j=1}^{s} \alpha_j Z_j^2,$$

when $n \to \infty$, *where* $Z_i, i = 1, ..., r$ *and* $Z_j, j = 1, ..., s$ *are iid normal variables with mean zero and variance* 1; *we assume that* $r = s$. β_i, $i = 1, ..., r$ *are the non null eigenvalues of the matrix* $H \sum_{F_\theta}$, $\alpha_j, j = 1, ..., s$ *are the non null eigenvalues of the matrix* $B \sum_F$, $r = rank(\sum_{F_\theta} H \sum_{F_\theta})$, *and* $s = rank(\sum_F B \sum_F)$, *being* $\sum_{F_\theta} = diag(F_\theta) - F_\theta F_\theta^t$ *and* $\sum_F = diag(F) - FF^t$ *and*

$$H = \left(\frac{\partial^2}{\partial f_i \partial f_j} D_\phi^B(F, F_\theta) \right)_{i, j=1, ..., M_0},$$

$$B = \left(\frac{\partial^2}{\partial f_{i\theta} \partial f_{j\theta}} D_\phi^B(F, F_\theta) \right)_{i, j=1, ..., M_0}.$$

Proof. The second order Taylor expansion of $\hat{D}_\phi^B(\hat{F}_{n,h}^b, F_{\hat{\theta}})$ about F and F_θ gives

$$\hat{D}_\phi^B(\hat{F}_{n,h}^b, F_{\hat{\theta}}) = \frac{1}{2}(\hat{F}_{n,h}^b - F)^T B(\hat{F}_{n,h}^b - F) + \frac{1}{2}(F_{\hat{\theta}} - F_\theta)^T H(F_{\hat{\theta}} - F_\theta) + o(\| F_{n,h} - F \|^2 + \| F_{\hat{\theta}} - F_\theta \|^2).$$

One can write

$$2n\hat{D}_\phi^B(\hat{F}_{n,h}^b, F_{\hat{\theta}}) = \sqrt{n}(\hat{F}_{n,h}^b - F)^T B\sqrt{n}(\hat{F}_{n,h}^b - F) + \sqrt{n}(F_{\hat{\theta}} - F_\theta)^T H\sqrt{n}(F_{\hat{\theta}} - F_\theta) + 2no(\| \hat{F}_{n,h}^b - F \|^2 + \| F_{\hat{\theta}} - F_\theta \|^2).$$

And $\sqrt{n}(F_{\hat{\theta}} - F_\theta) \xrightarrow{\mathcal{L}} N(0, \Sigma_{F_\theta})$, when $n \to \infty$; then $\| F_{\hat{\theta}} - F_\theta \|^2 = 0_p(n^{-1})$. Therefore $2no(\| F_{n,h}^b - F \|^2 + \| F_{\hat{\theta}} - F_\theta \|^2) = o_p(1)$. The random variables $2n\hat{D}_\phi^B(\hat{F}_{n,h}^b, F_{\hat{\theta}})$ and

$$\sqrt{n}(\hat{F}_{n,h}^b - F)^T B\sqrt{n}(\hat{F}_{n,h}^b - F) + \sqrt{n}(F_{\hat{\theta}} - F_\theta)^T H\sqrt{n}(F_{\hat{\theta}} - F_\theta)$$

have the same asymptotic distribution. Now by Corollary 2.1 in Dik and de Gunst [34] the result follows.

We consider now the case when the model is not specified, i.e., $H_1 : F_\theta \neq F$. Let us introduce the two important regularity assumptions.

- (A_1) Under the regularity conditions on the dominated model, the MLE is unique and asymptotically normal under F_θ, $\forall\theta$

(1) $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} N(0, I(\theta_0)^{-1})$, where $I(\theta_0)$ is Fisher information and $n \to \infty$.

(2) $F_{\hat{\theta}} \xrightarrow{as} F_{\theta_0}$ when $n \to \infty$.

- (A_2) There exists $\theta \in \Theta$; $\wedge^* = \begin{pmatrix} \wedge_{11} & \wedge_{12} \\ \wedge_{21} & \wedge_{22} \end{pmatrix}$, with $\wedge_{12} = \wedge_{21}$ and such that

$$\sqrt{n}\begin{pmatrix} \hat{F}_{n,h}^b - F \\ F_{\hat{\theta}} - F_\theta \end{pmatrix} \xrightarrow{\mathcal{L}} N(0, \wedge^*).$$

Theorem 5. *Under* $H_1 : F_\theta \neq F$ *and we assume that the conditions* $(A_1), (A_2)$ *hold, we have*:

$$\sqrt{n}[\hat{D}_\phi^B(\hat{F}_{n,h}^b, F_{\hat{\theta}}) - D_\phi^B(F, F_\theta)] \xrightarrow{\mathcal{L}} N(0, \wedge_\phi^2),$$

where

$$\wedge_\phi^2 = K^T \wedge_{11} K + K^T \wedge_{12} N + N^T \wedge_{12} K + N^T \wedge_{22} N \qquad (18)$$

$K^T = (k_1, ..., k_{M_0})$ *with*

$$k_i = \left(\frac{\partial}{\partial f_i} D_\phi^B(F, F_\theta)\right), \quad i = 1, ..., M_0.$$

$N^T = (n_1, ..., n_{M_0})$ *with*

$$n_i = \left(\frac{\partial}{\partial f_{i\theta}} D_\phi^B(F, F_\theta)\right), \quad i = 1, ..., M_0.$$

Proof. A first order Taylor expansion gives

$$\hat{D}_\phi^B(\hat{F}_{n,h}^b, F_{\hat{\theta}}) = D_\phi^B(F, F_\theta) + K^T(\hat{F}_{n,h}^b - F) + N^T(F_{\hat{\theta}} - F_\theta)$$

$$+ o(\| F_{n,h}^b - F \| + \| F_{\hat{\theta}} - F_\theta \|).$$

One can write

$$\sqrt{n}[\hat{D}_\phi^B(\hat{F}_{n,h}^b, F_{\hat{\theta}}) - D_\phi^B(F, F_\theta)] = \sqrt{n}[K^T(\hat{F}_{n,h}^b - F) + N^T(F_{\hat{\theta}} - F_\theta)]$$

$$+ \sqrt{n}o(\| \hat{F}_{n,h}^b - F \| + \| F_{\hat{\theta}} - F_\theta \|).$$

Since $\sqrt{n}(F_{\hat{\theta}} - F_\theta) \xrightarrow{\mathcal{L}} N(0, \Sigma_{F_\theta})$, when $n \to \infty$, with Σ_{F_θ} defined in Theorem 4, $\| F_{\hat{\theta}} - F_\theta \| = 0_p(n^{-1/2})$ and $\sqrt{n}o\| F_{\hat{\theta}} - F_\theta \| = o_p(1)$. Therefore $\sqrt{n}o(\| F_{n,h}^b - F \| + \| F_{\hat{\theta}} - F_\theta \|) = o_p(1)$.

Hence

$$\sqrt{n}[\hat{D}_\phi^B(\hat{F}_{n,h}^b, F_{\hat{\theta}}) - D_\phi^B(F, F_\theta)]$$

$$= \sqrt{n}[K^T(\hat{F}_{n,h}^b - F) + N^T(F_{\hat{\theta}} - F_\theta)] + o_p(1).$$

The random variables $\sqrt{n}[\hat{D}_\phi^B(\hat{F}_{n,h}^b, F_{\hat{\theta}}) - D_\phi^B(F, F_\theta)]$ and

$$\sqrt{n}[K^T(\hat{F}_{n,h}^b - F) + N^T(F_{\hat{\theta}} - F_\theta)]$$

have the same asymptotic distribution.

In view of A_1 and A_2 we have

$$\sqrt{n}[K^T(\hat{F}^b_{n,h} - F) + N^T(F_{\hat{\theta}} - F_\theta)] \xrightarrow{\mathcal{L}} N(0, \wedge^2_\phi),$$

where $\wedge^2_\phi$ is given by (18). This completes the proof.

Remark 1. On the basis of Theorem 5, the power function at $F \neq F_\theta$ when testing $H_0 : F = F_\theta$ is given by the formula

$$\beta_{n,\phi}(F) = 1 - \Phi_n\left(\frac{t_\alpha - 2nD^B_\phi(F, F_\theta)}{2\sqrt{n}\wedge_\phi}\right); \tag{19}$$

for a sequence of distribution function $\Phi_n(x)$ tending uniformly to the standard normal distribution function $\Phi(x)$; t_α is the critical value of $T_\phi = 2n\hat{D}^B_\phi(\hat{F}^b_{n,h}, F_{\hat{\theta}})$ and $\wedge_\phi$ is given in Theorem 5.

Thus, thanks to the goodness-of-fit test, it is possible to choose the best model among a collection of candidate models to be the one which is close to the true distribution according to the Bregman divergence.

4.2. Test for model selection with Bregman divergence

Consider the situation in which we have two candidate parametric models F_θ and $F_\gamma = \{F(\cdot, \gamma); \gamma \in \Gamma \subseteq \mathbb{R}^{M_0}\}$ another candidate model. We would like to chose the best of two candidate models based on their discrimination statistic between the observations and models F_θ and F_γ defined, respectively, as follows $\hat{D}^B_\phi(\hat{F}^b_{n,h}, F_{\hat{\theta}})$ and $\hat{D}^B_\phi(\hat{F}^b_{n,h}, F_{\hat{\gamma}})$. Our major work is to propose some tests for model selection, i.e., for the null hypothesis

$H_0 : D^B_\phi(F, F_\theta) = D^B_\phi(F, F_\gamma)$ means that the two models are equivalent,

$H_{f_\theta} : D^B_\phi(F, F_\theta) < D^B_\phi(F, F_\gamma)$ means that F_θ is better than F_γ,

$H_{f_\gamma} : D^B_\phi(F, F_\theta) > D^B_\phi(F, F_\gamma)$ means that F_θ is worse than F_γ.

To define the model selection statistic, let us give this next lemma.

Lemma 3. *Under the assumptions of Theorem* 5, *we have*

(i) *for the model* F_θ

$$\hat{D}_\phi^B(\hat{F}_{n,h}^b, F_{\hat{\theta}}) = D_\phi^B(F, F_\theta) + T_\theta^T(\hat{F}_{n,h}^b - F) + V_\theta^T(F_{\hat{\theta}} - F_\theta) + o_p(1), \quad (20)$$

(ii) *for model* F_γ

$$\hat{D}_\phi^B(\hat{F}_{n,h}^b, F_{\hat{\gamma}}) = D_\phi^B(F, F_\gamma) + T_\gamma^T(\hat{F}_{n,h}^b - F) + V_\gamma^T(F_{\hat{\gamma}} - F_\gamma) + o_p(1) \quad (21)$$

with $T_\theta^T = (t_1, ..., t_{M_0})$, *where*

$$t_i = \left(\frac{\partial}{\partial f_i} D_\phi^B(F, F_\theta)\right), \quad i = 1, ..., M_0$$

and $V_\theta^T = (v_1, ..., v_{M_0})$ *with*

$$v_i = \left(\frac{\partial}{\partial f_{i\theta}} D_\phi^B(F, F_\theta)\right), \quad i = 1, ..., M_0.$$

Proof. The result follows from a first order Taylor expansion.

We define

$$\kappa^2 = (T_\theta - T_\gamma; V_\theta - V_\gamma)^T \wedge^* (T_\theta - T_\gamma; V_\theta - V_\gamma) \quad (22)$$

which is the variance of

$$\sqrt{n}(T_\theta - T_\gamma; V_\theta - V_\gamma)^T \begin{pmatrix} F_{n,h} - F \\ F_{\hat{\theta}} - F_\theta \end{pmatrix}.$$

Since $T_\theta, T_\gamma, V_\theta, V_\gamma$ and $\wedge^*$, consistently estimated by their sample analogues $T_{\hat{\theta}}, T_{\hat{\gamma}}, V_{\hat{\theta}}, V_{\hat{\gamma}}$ and $\hat{\wedge}^*$. Hence, κ^2 is consistently estimated by

$$\hat{\kappa}^2 = (T_{\hat{\theta}} - T_{\hat{\gamma}}; V_{\hat{\theta}} - V_{\hat{\gamma}})^T \hat{\wedge}^* (T_{\hat{\theta}} - T_{\hat{\gamma}}; V_{\hat{\theta}} - V_{\hat{\gamma}}).$$

Let U be the model selection statistic and be given by

$$U = \frac{\sqrt{n}}{\hat{\kappa}}[\hat{D}_\phi^B(\hat{F}_{n,h}^b, F_{\hat{\theta}}) - \hat{D}_\phi^B(\hat{F}_{n,h}^b, F_{\hat{\gamma}})]. \tag{23}$$

Theorem 6 (Asymptotic distribution of the U-statistic)**.** *Under the assumptions of Theorem* 5, *suppose that* $\kappa \neq 0$, *then under the null hypothesis* $H_0, U \to N(0,1)$.

Proof. From Lemma 3, it follows that

$$\begin{aligned}
&\hat{D}_\phi^B(\hat{F}_{n,h}^b, F_{\hat{\theta}}) - \hat{D}_\phi^B(\hat{F}_{n,h}^b, F_{\hat{\gamma}}) \\
&= D_\phi^B(F, F_\theta) - D_\phi^B(F, F_\gamma) + T_\theta^T(\hat{F}_{n,h}^b - F) - T_\gamma^T(\hat{F}_{n,h}^b - F) \\
&\quad + V_\theta^T(F_{\hat{\theta}} - F_\theta) - V_\gamma^T(F_{\hat{\gamma}} - F_\gamma) + o_p(1).
\end{aligned}$$

Under H_0, $D_\phi^B(F, F_\theta) = D_\phi^B(F, F_\gamma)$, $F_\theta = F_\gamma$ and $F_{\hat{\theta}} = F_{\hat{\gamma}}$ we have

$$\begin{aligned}
&\hat{D}_\phi^B(\hat{F}_{n,h}^b, F_{\hat{\theta}}) - \hat{D}_\phi^B(\hat{F}_{n,h}^b, F_{\hat{\gamma}}) \\
&= T_\theta^T(\hat{F}_{n,h}^b - F) - T_\gamma^T(\hat{F}_{n,h}^b - F) + V_\theta^T(F_{\hat{\theta}} - F_\theta) - V_\gamma^T(F_{\hat{\gamma}} - F_\gamma) + o_p(1) \\
&= (T_\theta - T_\gamma; V_\theta - V_\gamma)^T \begin{pmatrix} \hat{F}_{n,h}^b - F \\ F_{\hat{\theta}} - F_\theta \end{pmatrix} + o_p(1).
\end{aligned}$$

Finally, applying the central limit theorem and assumptions (A_1)-(A_2), we can now immediately obtain $U \to N(0,1)$. It concludes the proof of Theorem 6.

Theorem 6 is quite general and gives us a wide variety of asymptotic standard normal tests for model selection based on Bregman divergence type statistic.

5. Numerical Studies

5.1. Life-data

We analyze a real life-data set in which a selection between Gamma and log-normal distributions is of a prime interest.

Data set. Suppose the following observations [35] are used to test whether the data come from a Gamma or a log-normal. The data given arose in tests on endurance of deep groove ball bearings. The data are number of million revolutions before failure for each of the lifetime tests and they are: 17.88, 28.92, 33.00, 41.52, 42.12, 45.60, 48.80, 51.84, 51.96, 54.12, 55.56, 67.80, 68.44, 68.64, 68.88, 84.12, 93.12, 98.64, 105.12, 105.84, 127.92, 128.04, 173.40. Here we consider Gamma model as the component of vector of densities F_θ and log-normal as the component of F_γ defined, respectively in Subsection 4.2. Therefore, to analyze a skewed positive data set an experimenter might wish to select one of them.

A random variable X is said to have a *Gamma distribution*, denoted by $GA(\alpha, \eta)$, when it has the probability density function (PDF) of

$$f_{GA}(x; \alpha, \eta) = \begin{cases} \dfrac{\eta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\eta x}, & x \geq 0, \alpha > 0, \eta > 0, \\ 0, & x < 0, \end{cases}$$

where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$. We know that $E(X) = \frac{1}{n}\sum_{i=1}^n X_i = \frac{\hat{\alpha}}{\hat{\eta}}$, then MLE of η in terms of α is given by

$$\hat{\eta}_n = \frac{n\hat{\alpha}_n}{\sum_{i=1}^n X_i}.$$

The MLE $\hat{\alpha}_n$ is defined by the relation

$$\sum_{j=1}^n \ln(\hat{\eta}_n X_j) - \frac{n\dot{\Gamma}(\hat{\alpha}_n)}{\Gamma(\hat{\alpha}_n)} = 0 \text{ and } \sqrt{n}(\hat{\alpha}_n - \alpha) \xrightarrow{\mathcal{L}} N(0, I(\alpha)^{-1}), \qquad (24)$$

when $n \to \infty$. The sequel dot means derivation w.r.t. α.

The unfortunately we cannot consider the MLE $\hat{\alpha}_n$ as a good estimator process because the calculation of the MLE as a solution of equation (24) is computationally too complicated.

We introduce the multi-step MLE process [36], which in this case provides us an estimator $\alpha_n^\star$ such that $\sqrt{n}(\alpha_n^\star - \alpha) \xrightarrow{\mathcal{L}} N(0, I(\alpha)^{-1})$ when $n \to \infty$. Suppose that we have n i.i.d. r.v.'s $X^n = (X_1, ..., X_n)$ with smooth density function $f(x, \alpha)$ and $l(x, \alpha) = \ln f(x, \alpha)$. Here $\alpha \in \Theta$. Let us denote $\overline{\alpha}_N$ the preliminary estimator constructed by the first $N = [n^\delta]$ observations $X^N = (X_1, ..., X_N)$ with $\delta = \left(\frac{1}{2}, 1\right)$. Then the one-step MLE process $\alpha_n^\star = (\alpha_{k,n}^\star, N + 1 \le k \le n)$ is defined by the equality

$$\alpha_{k,n}^\star = \overline{\alpha}_N + I(\overline{\alpha}_N)^{-1} \frac{1}{k} \sum_{j=N+1}^{k} i(X_j, \overline{\alpha}_N),$$

for $k = [sn]$, $s \in (0, 1]$; we have the convergence

$$\sqrt{k}(\alpha_{k,n}^\star - \alpha_0) \xrightarrow{\mathcal{L}} N(0, I(\alpha_0)^{-1}).$$

Here s is fixed and $n \to \infty$. Therefore, $\alpha_n^\star$ is *a good estimator process*, *i.e.*, $\alpha_{k,n}^\star$ depends on $X^k = (X_1, ..., X_k)$, easy to calculate and is asymptotically efficient because it is asymptotically equivalent to the MLE.

Therefore for our case, the preliminary estimator $\overline{\alpha} = \frac{\hat{\eta}}{n} \sum_{i=1}^{n} X_i \to \alpha_0$, $\sqrt{n}(\overline{\alpha} - \alpha_0) \xrightarrow{\mathcal{L}} N(0, I(\alpha_0^{-1}))$. Then the one-step MLE process is given by

$$\hat{\alpha} = \overline{\alpha} + \frac{1}{nI(\overline{\alpha})} \sum_{i=1}^{n} \left[\ln(\hat{\eta} X_i) - \frac{\dot{\Gamma}(\overline{\alpha})}{\Gamma(\overline{\alpha})} \right].$$

Then

$$\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{\mathcal{L}} N(0, I(\alpha)^{-1}), \; n \to \infty,$$

where $I(\overline{\alpha}) = \dfrac{(\ddot{\Gamma}(\overline{\alpha})\Gamma(\overline{\alpha}) - \dot{\Gamma}^2(\overline{\alpha}))}{\Gamma^2(\overline{\alpha})}$ with

$$\dot{\Gamma}(\overline{\alpha}) = \int_0^\infty (\ln x) x^{\overline{\alpha}-1} e^{-x} dx,\ \ddot{\Gamma}(\overline{\alpha}) = \int_0^\infty (\ln x)^2 x^{\overline{\alpha}-1} e^{-x} dx$$

and α_0 the true value of α.

A random variable X is distributed as log-normal, denoted as $LN(\mu, \sigma^2)$, if $\ln(X)$ is normal, i.e., $\ln(X) \sim N(\mu, \sigma^2)$. The probability density of X is given by

$$f_{LN}(x; \mu, \sigma^2) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{1}{2\sigma^2}(\ln(x)-\mu)^2}, & x \geq 0, \mu > 0, \sigma > 0, \\ 0, & x < 0. \end{cases}$$

The MLE of μ and σ are given below, respectively:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \ln(X_i) \ \text{ and } \ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (\ln(X_i) - \hat{\mu})^2.$$

For $\beta = 3$ and $c_1 = 1$ the relations (9) and (10) allow us to compute the $\hat{D}_\phi^B(\hat{f}_{n,h}^b(x), f_{\widehat{GA}}(x))$ and $\hat{D}_\phi^B(\hat{f}_{n,h}^b(x), f_{\widehat{LN}}(x))$ as follows:

$$\hat{D}_\phi^B(\hat{f}_{n,h}^b(x), f_{\widehat{GA}}(x))$$

$$= \frac{1}{6}\int_0^\infty ((\hat{f}_{n,h}^b(x))^3 - 3f_{n,h}^b(x) f_{\widehat{GA}}^2(x) + 2f_{\widehat{GA}}^3(x))dx \qquad (25)$$

and

$$\hat{D}_\phi^B(\hat{f}_{n,h}^b(x), f_{\widehat{LN}}(x))$$

$$= \frac{1}{6}\int_0^\infty ((\hat{f}_{n,h}^b(x))^3 - 3f_{n,h}^b(x) f_{\widehat{LN}}^2(x) + 2f_{\widehat{LN}}^3(x))dx, \qquad (26)$$

where $\hat{f}^b_{n,h}(\cdot)$ is given by (5). We consider Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$ because it has infinitely many (nonzero) derivatives as our candidate models. Note that for the Gaussian kernel,

$$\hat{f}^b_{n,h}(x) = \frac{1}{2\sqrt{2\pi}nh} \sum_{i=1}^{n} \left[3 - \left(\frac{x - X_i}{h} \right)^2 \right] e^{-\frac{1}{2}\left(\frac{x-X_i}{h}\right)^2}.$$

To get h optimal, the cross-validation method introduced in [37] giving the simple and attractive smoothing parameter is used. Hence, $h \equiv h_{CV} = \arg\min_{h>0} CV(h)$, where $CV(h)$ is cross-validation given by $CV(h) = \int \hat{f}^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{,-i}(X_i)$ and

$$\hat{f}_{,-i}(x) = \frac{1}{(n-1)h} \sum\nolimits_{j \neq i} K\left(\frac{x - X_j}{h} \right).$$

$f_{\widehat{GA}}(x)$ and $f_{\widehat{LN}}(x)$ are parametric estimators of Gamma and log-normal models and are given by $f_{\widehat{GA}}(x) = \frac{\hat{\eta}^{\hat{\alpha}}}{\Gamma(\hat{\alpha})} x^{\hat{\alpha}-1} e^{-\hat{\eta}x}$ and $f_{\widehat{LN}}(x) = \frac{1}{\sqrt{2\pi}\hat{\sigma}x} e^{-\frac{1}{2\hat{\sigma}^2}(\ln(x)-\hat{\mu})^2}$, respectively.

For the data at hand, Ali-Akbar and Reza [2] proved that Gamma fits better in discrimination between Gamma and log-normal distributions using the Ratio of Minimized Kullback-Leibler Divergence. Therefore, we obtain for the Gamma model $\hat{\alpha} = 4.028040$ and $\hat{\eta} = 0.055767$. And for log-normal distribution $\hat{\mu} = 4.150614$ and $\hat{\sigma} = 0.521485$. From (25) and (26), one has $\hat{D}^B_{\phi_1} \equiv \hat{D}^B_{\phi}(\hat{f}^b_{n,h}(x), f_{\widehat{GA}}(x)) = 0.000021$ and

$$\hat{D}^B_{\phi_2} \equiv \hat{D}^B_{\phi}(\hat{f}^b_{n,h}(x), f_{\widehat{LN}}(x)) = 0.000024.$$

Bregman divergence is the non-symmetric measure of the difference (dissimilarity) between two probability distributions.

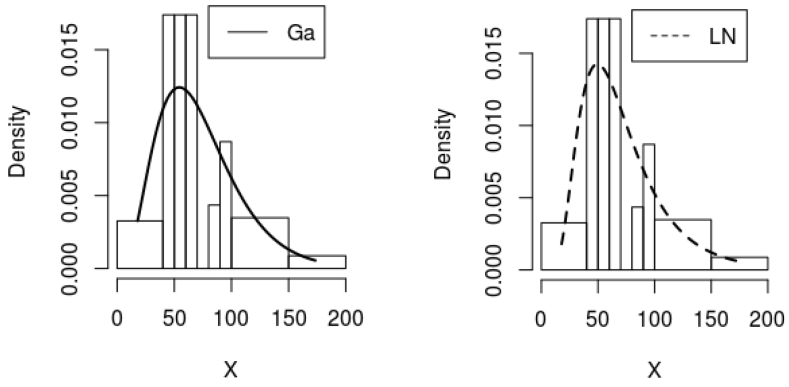

Figure 1. The histogram with log-normal and Gamma density functions for the given data set.

Being interested to select the model minimizing the BD as the best model, i.e., $U = -0.000002$. At 5% significance level, we compare U with -1.96 and 1.96. U falls between -1.96 and 1.96, we conclude that both estimated models fit the data equally well. Figures 1 and 2 show that these models may provide similar data fit for moderate sample sizes. Note that many observed data are concentrated between 40 and 80 considering the axis of the X of our figures.

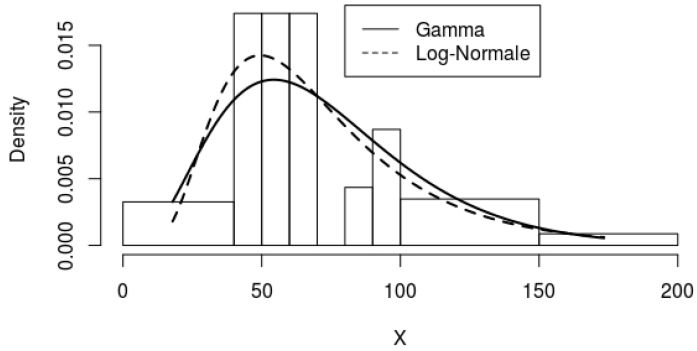

Figure 2. The histogram and two fitted density functions for the given data set.

5.2. Simulation study

To illustrate well our model selection procedure, we have defined our candidate models, Bregman divergence type statistic to measure the closeness between our candidate models and reduced bias kernel density estimator. Consider the data generated from a mixture of Gamma and log-normal distributions. These two distributions are calibrated by the multi-step MLE process from the data set defined in Subsection 5.1. Hence, the Data Generating Process (DGP) has density

$$l(\pi) = \pi Gamma(4.02804, 0.05576722) + (1-\pi) log\text{-}normal(4.150614, 0.5214847),$$

where $\pi \in (0,1)$ is specific to each set of experiments. In each set, several random samples are drawn from this mixture. The sample size varies from 20 to 90, and for each sample size the number of replications is 1000. We choose different values of π which are 0.00, 0.25, 0.5, 0.75, 1.00. Although our proposed model selection procedure does not require that the data generating process belong to either of the candidate models. We consider the two limiting cases $\pi = 0.00$ and $= 1.00$ for they correspond to the correctly specified cases. For $\pi = 0.25$ and $\pi = 0.75$ both candidate models are misspecified but not at equal distance from the DGP. These cases correspond to a DGP which is Gamma or log-normal distributions but slightly contaminated by the other distribution. The value $\pi = 0.5$ is the value for which the Gamma and log-normal distributions are approximately at equal distance to the mixture $l(\pi)$ according to statistics $\hat{D}^B_{\phi_1} \equiv \hat{D}^B_{\phi}(\hat{f}^b_{n,h}(x), f_{\widehat{GA}}(x))$ and $\hat{D}^B_{\phi_2} \equiv \hat{D}^B_{\phi}(\hat{f}^b_{n,h}(x), f_{\widehat{LN}}(x))$. Our model selection statistic is given by U. The results of our five sets of experiments are presented in Tables 1-5. For $n = 60$, we plot the histogram of datasets and overlay the curves for Gamma and log-normal distribution in order to analyze the closeness of these two models.

Table 1. DGP = *log-normal*(4.150614, 0.5214847)

n		20	40	60	80	90
$\hat{\alpha}$		4.5391	4.1526	4.0529	3.9728	3.9540
		(1.6452)	(0.9603)	(0.7491)	(0.6603)	(0.5743)
$\hat{\eta}$		0.0644	0.0579	0.0562	0.0551	0.0548
		(0.2783)	(0.0162)	(0.0125)	(0.0109)	(0.0095)
$\hat{\mu}$		4.1494	4.1522	4.1515	4.1480	4.1483
		(0.1155)	(0.0859)	(0.0686)	(0.0582)	(0.0540)
$\hat{\sigma}$		0.5019	0.5121	0.5145	0.5180	0.5184
		(0.0840)	(0.0566)	(0.0469)	(0.0418)	(0.0256)
$\hat{D}^B_{\phi_1}$		0.000031	0.000028	0.000026	0.000026	0.000026
		(0.000015)	(0.000009)	(0.000007)	(0.000006)	(0.000006)
$\hat{D}^B_{\phi_2}$		0.000026	0.000023	0.000021	0.000021	0.000021
		(0.000014)	(0.000009)	(0.000007)	(0.000006)	(0.000004)
U		1.470805	2.371773	3.035950	3.132926	3.283792
		(0.000003)	(0.000002)	(0.000002)	(0.000002)	(0.000001)
Model selection based on U	Correct	24.9%	67.0%	87.1%	91.7%	93.9%
	Indecisive	74.6%	32.8%	12.8%	8.2%	6.1%
	incorrect	0.5%	0.2%	0.1%	0.1%	0.0%

Table 2. DGP = *Gamma*(4.02804, 0.05576722)

n	20	40	60	80	90
$\hat{\alpha}$	4.7264	4.3407	4.1839	4.2493	4.1606
	(1.7632)	(0.9604)	(0.7605)	(0.7903)	(0.6042)
$\hat{\eta}$	0.6627	0.0606	0.0580	0.0589	0.0576
	(0.0262)	(0.0148)	(0.0187)	(0.0113)	(0.010)
$\hat{\mu}$	4.1499	4.1500	4.1508	4.1527	4.1525
	(0.120811)	(0.083108)	(0.052709)	(0.068731)	(0.0575)
$\hat{\sigma}$	0.509398	0.518293	0.526293	0.521632	0.5250
	(0.095431)	(0.063427)	(0.034347)	(0.053744)	(0.0432)
$\hat{D}^B_{\phi_1}$	0.000027	0.000024	0.000023	0.000023	0.000022
	(0.000012)	(0.00008)	(0.00009)	(0.000006)	(0.00000)

$\hat{D}^B_{\phi_2}$		0.000030	0.000027	0.000026	0.000026	0.000025
		(0.000012)	(0.00009)	(0.000010)	(0.000006)	(0.000005)
U		-1.177846	-1.604843	-1.968482	-2.260217	-2.277816
		(0.000003)	(0.000002)	(0.000002)	(0.000002)	(0.000001)
Model selection based on U	Correct	14.2%	31.6%	47.0%	64.3%	66.2%
	Indecisive	85.6%	68.0%	52.9%	35.4%	33.6%
	incorrect	0.2%	0.4%	0.1%	0.3%	0.2%

Table 3. DGP = $0.25Gamma(4.02804,\ 0.05576722) + 0.75log\text{-}normal(4.150614, 0.5214847)$

n		20	40	60	80	90
$\hat{\alpha}$		7.2498	6.7094	6.4145	6.3138	6.3687
		(2.5978)	(1.7032)	(1.3363)	(0.6603)	(1.0345)
$\hat{\eta}$		0.101999	0.093505	0.088910	0.087376	0.08830
		(0.041302)	(0.026984)	(0.020856)	(0.010946)	(0.01631)
$\hat{\mu}$		4.200020	4.200743	4.200762	4.200372	4.1993
		(0.091232)	(0.065416)	(0.051791)	(0.058290)	(0.0422)
$\hat{\sigma}$		0.391877	0.398275	0.403963	0.404366	0.4030
		(0.066546)	(0.046814)	(0.0 3.9236)	(0.041885)	(0.0314)
$\hat{D}^B_{\phi_1}$		0.000041	0.000040	0.000039	0.000039	0.000036
		(0.000015)	(0.000012)	(0.000011)	(0.000009)	(0.000008)
$\hat{D}^B_{\phi_2}$		0.000037	0.000036	0.000034	0.000034	0.000034
		(0.000016)	(0.000012)	(0.000011)	(0.000009)	(0.00008)
U		0.815310	1.427280	2.106998	2.934055	3.455230
		(0.000005)	(0.000003)	(0.000003)	(0.000002)	(0.000002)
Model selection based on U	Gamma	1.9%	1.5%	0.6%	0.3%	0.1%
	Indecisive	93.9%	78.1%	43.5%	10.5%	4.0%
	log-normal	4.2%	20.4%	55.9%	89.2%	95.9%

Table 4. DGP = 0.5*Gamma*(4.02804, 0.05576722) + 0.5*log-normal* (4.150614, 0.5214847)

n		20	40	60	80	90
$\hat{\alpha}$		8.836202	8.154637	8.011385	7.880364	7.8554
		(3.143420)	(1.916379)	(1.491180)	(1.337800)	(1.1732)
$\hat{\eta}$		0.122711	0.113053	0.110831	0.108976	0.1086
		(0.046264)	(0.028233)	(0.021856)	(0.019985)	(0.01761)
$\hat{\mu}$		4.219796	4.215739	4.216706	4.216613	4.2166
		(0.080647)	(0.059911)	(0.049240)	(0.042338)	(0.03988)
$\hat{\sigma}$		0.358219	0.364973	0.365404	0.367736	0.3671
		(0.059884)	(0.0 42608)	(0.034347)	(0.031361)	(0.02839)
$\hat{D}^B_{\phi_1}$		0.000046	0.000042	0.000041	0.000041	0.000040
		(0.000017)	(0.000012)	(0.000010)	(0.000009)	(0.000008)
$\hat{D}^B_{\phi_2}$		0.000049	0.000047	0.000046	0.000045	0.000045
		(0.000017)	(0.000012)	(0.000010)	(0.000009)	(0.000008)
U		-0.566535	-0.944703	-1.421004	-1.675458	-1.972893
		(0.000006)	(0.000004)	(0.000003)	(0.000002)	(0.000002)
Model selection based on U	Gamma	2.2%	6.7%	22.2%	34.7%	51.7%
	Indecisive	96.4%	91.3%	76.9%	64.4%	47.7%
	log-normal	1.4%	2.0%	0.9%	0.9%	0.6%

Table 5. DGP = 0.75*Gamma*(4.02804, 0.05576722) + 0.25*log-normal* (4.150614, 0.5214847)

n	20	40	60	80	90
$\hat{\alpha}$	7.680876	6.872101	6.768990	6.709298	6.592283
	(2.705339)	(1.545658)	(1.257030)	(1.071147)	(0.647426)
$\hat{\eta}$	0.107086	0.095342	0.093907	0.093040	0.091217
	(0.039593)	(0.022479)	(0.018754)	(0.015924)	(0.009525)
$\hat{\mu}$	4.204221	4.202126	4.201923	4.201670	4.202510
	(0.090257)	(0.065006)	(0.052709)	(0.046475)	(0.029065)
$\hat{\sigma}$	0.386789	0.400571	0.401475	0.402087	0.402912
	(0.065110)	(0.045933)	(0.034347)	(0.033488)	(0.20713)

$\hat{D}^B_{\phi_1}$		0.000040	0.000037	0.000036	0.000035	0.000035
		(0.000025)	(0.000011)	(0.00009)	(0.000008)	(0.000005)
$\hat{D}^B_{\phi_2}$		0.000043	0.000040	0.000040	0.000039	0.000038
		(0.000024)	(0.000011)	(0.000010)	(0.000008)	(0.00000)
U		-0.654543	-1.296389	-2.172371	-2.545976	-3.875059
		(0.000005)	(0.000002)	(0.000002)	(0.000002)	(0.000001)
Model selection based on U	Gamma	2.2%	18.6%	57.8%	72.1%	97.2%
	Indecisive	96.5%	80.4%	42.1%	27.8%	0.0
	log-normal	1.3%	1.0%	0.1%	0.1%	2.8%

The first half of each table gives the average values of the multi-step MLE process estimators $\hat{\alpha}$, $\hat{\eta}$, $\hat{\mu}$ and $\hat{\sigma}$, the Bregman divergence test statistics $\hat{D}^B_{\phi_1}$ and $\hat{D}^B_{\phi_2}$ and the model selection statistic U. The values in parentheses are standard errors. The second half of each table gives the probability of correct selection (PCS) which is in percentage the number of times our proposed model selection procedure based on U, favors the Gamma model, the log-normal model and indecisive. The tests are conducted at 5% nominal significance level. In the first two sets of experiments $(\pi = 0.00$ and $\pi = 1.00)$, where one model is correctly specified, we use the labels *correct*, *incorrect* and *indecisive* when a choice is made. The first halves of Tables 1-5 confirm our asymptotic results. They all show that the multi-step MLE process estimators $\hat{\alpha}$, $\hat{\eta}$, $\hat{\mu}$ and $\hat{\sigma}$ converge rapidly to their pseudo-true values in the misspecified cases and to their true values in the correctly specified cases as the sample size increases. The statistics $\hat{D}^B_{\phi_1}$ and $\hat{D}^B_{\phi_2}$ converge approximately to zero at the rate of n, as expected when the models are correctly specified and when the models are misspecified. With respect to our U, it diverges to $-\infty$ at the approximate rate of $\sqrt{n}$. In Tables 3, 4 and 5, we observed a large percentage of incorrect decisions. This is because both models are now incorrectly specified. In contrast, turning to the second halves of Tables 1 and 2, we first note that the percentage of correct choices

using model selection statistic steadily increases and ultimately converge to 100%. As a consequence, the probability of correct choice (PCS) based on Monte Carlo simulation is found to be significantly higher in choosing the correct model in this selection procedure based on Bregman divergence. The preceding comments for the second halves of Tables 1 and 2 also apply to the second halves of Tables 3 and 4. Table 5 also confirms our asymptotics results: as sample size increases, the percentage of rejection of both models steadily decreases but still keeping the highest percentage. In all figures, we plot the histogram of datasets and overlay the curves for Gamma and log-normal distributions. They all (figures) show that these two distributions (Gamma and log-normal distributions) are close and closely approximate the data. This is because these distributions are often interchangeable and commonly used to model certain lifetimes in reliability and survival analysis (Wiens [38]).

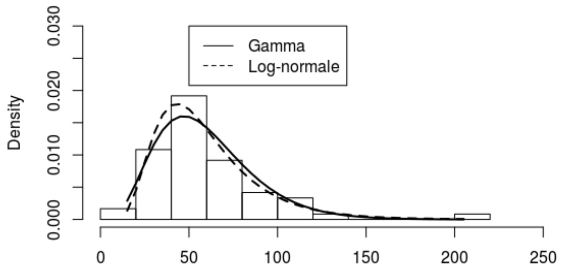

Figure 3. Histogram of DGP = log-normal(4.150614, 0.5214847), with $n = 60$ and $\pi = 0$.

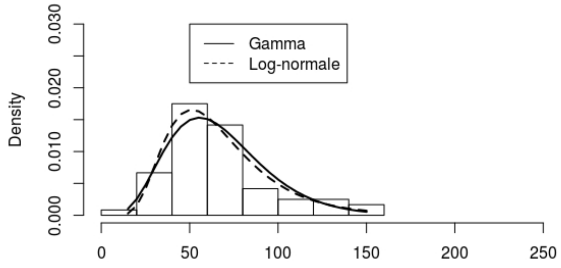

Figure 4. Histogram of DGP = Gamma(4.02804, 0.05576722), with $n = 60$ and $\pi = 1$.

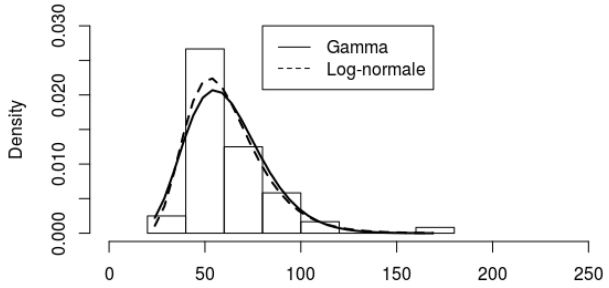

Figure 5. Histogram of DGP = 0.25Gamma(4.02804, 0.05576722) + 0.75log-normal(4.150614, 0.5214847), with $n = 60$ and $\pi = 0.25$.

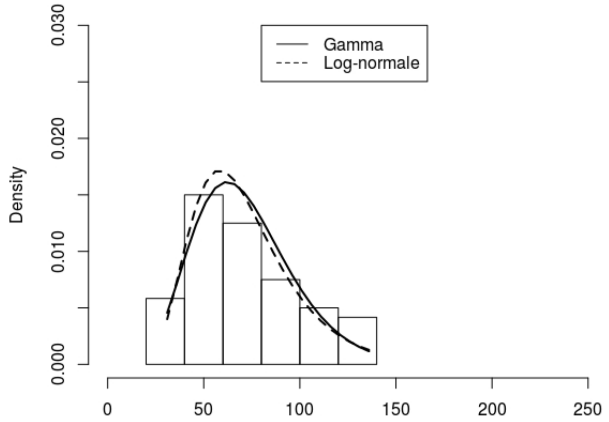

Figure 6. Histogram of DGP = 0.5Gamma(4.02804, 0.05576722) + 0.5log-normal(4.150614, 0.5214847), with $n = 60$ and $\pi = 0.5$.

When the DGP is correctly specified (Figure 1), the log-normal distribution has reasonable chance to be distinguished from Gamma distribution. Similarly, in Figure 2, as can be seen, the Gamma distribution closely approximates the data sets. In Figures 3 and 5, these two distributions are close but the log-normal (Figure 3) and the Gamma distributions (Figure 5) do appear to be much closer to the data sets. When $\pi = 0.5$, the distribution for both (Figure 4) log-normal distribution and Gamma distribution is nearly similar.

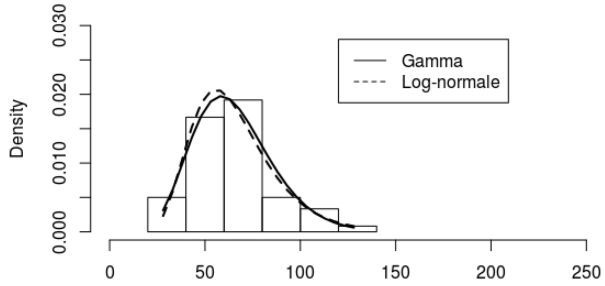

Figure 7. Histogram of DGP = 0.75Gamma(4.02804, 0.05576722) + 0.25log-normal(4.150614, 0.5214847), with $n = 60$ and $\pi = 0.75$.

6. Conclusion

In this paper, we have studied the problem of selecting estimated models using Bregman divergence type statistics. In particular, we have proposed some asymptotically standard normal and hypothesis tests based on Bregman divergence type statistics. We suggest using robust estimators such as the corresponding multi-step MLE process estimators for guaranteeing the optimality like consistent and efficient estimators. We investigate tests for model selection which are designed to determine whether the estimated candidate models are as close to the true distribution against alternative hypothesis that one estimated model is closer, where the closeness is measured according to the discrepancy implicit in the Bregman divergence type statistic used. Theoretical properties, such as consistency and rate of convergence of the Bregman divergence estimator are studied. We also show that with proper choice of the bias reduced kernel density estimator, one can ensure the improvement on convergence rate to the true distribution. Both simulated and real example show that the model selection procedure based on the Bregman divergence criterion competitively especially in small samples.

References

[1] B. Ali-Akbar, discriminating between Weibull and log-normal distribution based on Kullback-Leibler divergence, Ekonometri ve statistik Say 16 (2012), 44-54.

[2] B. Ali-Akbar and V. Reza, Discrimination between gamma and log-normal distributions by ratio of minimized Kullback-Leibler divergence, Pakistan Journal of Statistics and Operation Research IX(4) (2013), 441-451.

[3] X. Xie and J. Wu, Some improvement on convergence rates of kernel density estimator, Published Online June 2014 in SciRes. http://www.scirp.org/journal/am Applied Mathematics 5 (2014), 1684-1696. Doi.org/10.4236/am.2014.511161.

[4] P. Chen, Yunmei and M. Rao, Metrics defined by Bregman divergence, International Press 6(4) (2008), 927-948.

[5] W. Stummer and I. Vajda, On Bregman Distances and Divergences of Probability Measures, Wolfgang Stummer and Igor Vajda, Fellow, IEEE arvXiv:0911.2784v2 [cs:IT] 5 Oct. 2011.

[6] Inderjit S. Dhillon and Suvrit Sra, Generalized nonnegative matrix approximations with Bregman divergences, Dept. of Computer Sciences, Univ. of Texas at Austin, TX 78712 inderjit,suvrit@cs.utexas.edu.

[7] A. Cichocki, S. Cruces and S. Amari, Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization, Entropy 13 (2011), 134-170. www.mdpi.com/journal/entropy, Doi: 10.3390/e13010134.

[8] F. Itakura and S. Saito, Analysis Synthesis Telephony Based Upon Maximum Likelihood Method, Repts. of the 6th International Cong. Acoust., Y. Kohasi, ed., Tokyo, C-5-5, C17-20, 1968.

[9] S. Kullback and R. Leibler, On information and sufficiency, Ann. Math. Statist. 22(1) (1951), 79-86.

[10] P. C. Mahalanobis, On the generalized distance in statistics, Proceedings of National Institute of Science of India 12 (1936), 49-55.

[11] C. L. Mallows, Some comments on Cp, Technometrics 15(4) (1973), 661-675. http://links.jstor.org/sici?sici=0040-1706%28197311%2915%3A4%3C661%3ASCO%3E2.0.CO%3B2-6.

[12] H. Akaike, Information theory and an extension of the maximum likelihood principle, Proceedings of the Second International Symposium on Information Theory Akademiai Kaido, Budapest, 1973, pp. 267-281.

[13] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (1978), 461-464.

[14] S. Konishi and G. Kitagawa, Generalised information criteria in model selection, Biometrika 83 (1996), 875-890.

[15] A. Toma, Model selection criteria using divergences, Entropy 16(5) (2014), 2686-2698. Doi: 10.3390/e16052686.

[16] N. Z. Mohd Saat, A. A. Jemain and S. H. Al-Mashoor, A comparison of Weibull and gamma distributions in application of sleep apnea, Asian Journal of Mathematics and Statistics 1(3) (2008), 132-138.

[17] Q. H. Vuong, Likelihood ratio tests for model selection and non-nested hypotheses, Econometrica 57(2) (1989), 257-306. Doi: 10.2307/1912557.

[18] A. Basu, I. R. Harris and S. Basu, Tests of hypotheses in discrete models based on the penalized Hellinger distance, Statist. Probab. Lett. 27(4) (1996), 367-373. Doi: 10.1016/0167-7152(95)00101-8.

[19] M. Rosenblatt, On estimation of a probability density function and the mode, Ann. Math. Statist. 33 (1956), 1065-1076. http://dx.doi.org/10.1214/aoms/1177704472.

[20] E. Parzen, Remarks on some nonparametric estimates of a density function, Ann. Math. Statist. 27 (1962), 832-837. http://dx.doi.org/10.1214/aoms/1177728190.

[21] L. Devroye and L. Gyorfi, Nonparametric Density Estimation, The L1 View, Wiley Series in Probability and Mathematical Statistics: Tracts on Probability and Statistics, John Wiley and Sons Inc., New York, 1985.

[22] A. Cichocki, R. Zdunek and S. Amari, Csiszars divergences for non-negative matrix factorization: Family of new algorithms, Conference on Independent Component Analysis and Blind Source Separation (ICA), Charleston, SC, USA, 2006, pp. 32-39.

[23] A. Cichocki, R. Zdunek, S. Choi, Robert J. Plemmons and S. Amari, Non-negative tensor factorization using alpha and beta divergences, IEEE International Conference on Acoustics, Speech, and Signal Processing, Volume 3, Honolulu, Hawaii, USA, 2007, pp. 1393-1396.

[24] L. M. Bregman, The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming, USSR Comput. Math. Math. Phys. 7 (1967), 210-217.

[25] Inderjit S. Dhillon and Suvrit Sra, Generalized nonnegative matrix approximations with Bregman divergences, Y. Weiss, B. Scholkopf and J. Platt, eds., Neural Information Processing Systems Conference (NIPS), MIT Press, Cambridge, MA, December 2006, pp. 283-290.

[26] A. Basu, I. R. Harris, N. Hjort and M. Jones, Robust and efficient estimation by minimising a density power divergence, Biometrika 85(3) (1998), 549-559.

[27] M. Minami and S. Eguchi, Robust blind source separation by Beta-divergence, Neural Comput. 14 (2002), 1859-1886.

[28] A. Cichocki, R. Zdunek, S. Choi, R. Plemmons and S. Amari, Nonnegative tensor factorization using Alpha and Beta divergences, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Tulose, France, Volume III, May 2007, pp. 1393-1396.

[29] N. Murata, T. Takenouchi, T. Kanamori and S. Eguchi, Information geometry of U-boost and Bregman divergence, Neural Comput. 16 (2004), 1437-1481.

[30] A. Cichocki and Shun-Ichi Amari, Families of Alpha- Beta- and Gamma-Divergences: Flexible and Robust Measures of Similarities, Entropy 12 (2010), 1532-1568. www.mdpi.com/journal/entropy Doi: 10.3390/e12061532.

[31] F. Liese and I. Vajda, Convex Statistical Distances, Teubner-Texte zur Mathematik Teubner Texts in Mathematics 95 (1987), 1-85.

[32] U. Einmahl and D. M. Mason, An empirical process approach to the uniform consistency of kernel-type function estimators, J. Theoret. Probab. 13(1) (2000), 1-37.

[33] U. Einmahl and D. M. Mason, Uniform in bandwidth consistency of kernel-type function estimators, Ann. Statist. 33(3) (2005), 1380-1403.

[34] J. J. Dik and M. C. M. de Gunst, The distribution of general quadratic forms in normal variables, Statistica Neerlandica 39 (1985), 14-26.

[35] J. F. Lawless, Statistical Models and Methods for Lifetime Data, Wiley, New York, 1982. H. Linhart and W. Zucchini, Model Selection, Wiley, New York, 1986.

[36] Yury A. Kutoyants, On the multi-step MLE-process for ergodic diffusion, Stochastic Processes and their Applications 127(7) (2017), 2243-2261.

[37] Alexandre B. Tsybakov, Introduction to nonparametric estimation, Springer Series in Statistics, ISBN: 978-0-387-79051-0, Doi: 10.1007/978-0-387-79052-7. Library of Congress Control Number: 2008939894.

[38] B. L. Wiens, When log-normal and gamma models give different results: a case study, Amer. Statist. 53(2) (1999), 89-93.

Papa Ngom: papa.ngom@ucad.edu.sn