Discriminating between two models based on Bregman divergence in small samples

Papa NGOM\textsuperscript{a,b,}\, Jean de Dieu NKURUNZIZA\textsuperscript{b}, Carlos Simplice OGOUYANDJOU\textsuperscript{b}

\textsuperscript{a}LMA, Université Cheikh Anta Diop, Dakar, Senegal
\textsuperscript{b}Institut de Mathématiques et de Sciences Physiques, Porto Novo, Benin

Abstract

Recently in \cite{1, 2}, Ali-Akbar 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 in order to improve the process of choice of the model. Our approach differs from him. 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 kernel density estimator and the second estimated candidate model. In these two cases Bregman Divergence (BD) and the bias reduced kernel estimator \cite{3} focuses on improving the convergence rates of kernel density estimators are used. Our testing procedure for model selection is thus based on the comparison of the value of model selection test statistic to critical values from a standard normal table. We establish the asymptotic properties of Bregman divergence 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.

Keywords: A Bias Reduced Kernel Estimator; Bregman Divergence; Hypothesis Test.

1. Introduction

Bregman (1967) introduced for convex functions\cite{4, 5, 6, 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 pseudodistance which is reflexive but neither symmetric nor satisfying the triangle inequality). In the last decade, Bregman divergences have become an important tool in many research areas. For instance, several specific Bregman divergences, such as Itakura-Saito Distance \cite{8, 9, 10}, Kullback-Leibler Divergence (KLD) \cite{11}, and Mahalanobis Distance (P.C. Mahalanobis; 1936) 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 \cite{21}, Stone \cite{22}, Akaike \cite{23}, Schwarz \cite{24} and Konishi and Kitagawa \cite{25}, Aida Toma \cite{26} respectively. Ngom and Ntep \cite{27, 28} provided (in their outcomes of the tests) information on the
strength of the statistical evidence for the choice of a model based on its goodness-of-fit. Vuong’s tests [29] for model selection rely on the Kullback-Leibler Information Criterion (KLIC). Mohd Saat et al (2008) compared RML with Vuong’s closeness test to discriminant between Gamma and Weibull, in which they found both methods relatively similar. Similar work to compare RML with Kolmogorov-Smirnov and Chi-Squared (with asymptotic properties) has been studied and some inconsistency among them are reported (Basu et al, 2009). Despite of significant amount of work on discrimination by different methods, there is no much work to use the KLD and compare it with alternatives test statistic [1, 2]. Akbar introduced the Kullback-Leibler Divergence (KLD) test statistic in discriminating between two models. It was observed that the proposed test statistic named Ratio of minimized Kullback-Leibler divergence (RMKLD) is consistent with alternative testing statistic, say RML. The purpose of this paper is to generalize the works of Ali-Akbar Bromideh [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. 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. Following Vuong [30] the procedures considered here are testing the null hypothesis that the competing models are equally close to the data generating process (DGP) versus the alternative hypothesis that one model is closer to the DGP 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 definitions and notations in Section 2. For proving the results rigorously we briefly describe a reduced bias kernel estimator and the Bregman divergence in Section 3 and 4. In Section 5, we establish the consistancy of the Bregman divergence estimator. In Section 6 we obtain asymptotic distributions of the Bregman divergence estimator. Model selection and Bregman divergence based test Statistic are presented in Section 7. Examples and Data Analysis: Implementation of the Bregman divergence test statistic and the probability of correct selection (PCS) is obtained in Section 8 and finally the conclusion appears in Section 9.

2. Definitions and Notations

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

\[
F = \left\{ F_\theta = (f_1(\cdot, \theta), ..., f_{M_0}(\cdot, \theta))^T : \theta \in \Theta \right\},
\]

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

We assume that the probability distributions \(F_\theta\) is absolutely continuous with respect to a \(\sigma\)-finite measure \(\mu\) on \((X, \beta_X)\). For simplicity \(\mu\) is either the Lebesgue measure or a counting measure. The parametric family of models may or may not contain the true model. If \(F\) contains the true model, then there exists a \(\theta_0 \in \Theta\) such that \(F_{\theta_0} = F\) and the model \(F_\theta\) is said to be correctly specified. We are interested in testing

\[
H_0 : F = F_{\theta_0} \text{ versus } H_1 : F \neq F_{\theta_0}.
\]

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.
3. A Bias Reduced Kernel Estimator

Kernel density estimator was first introduced by Rosenblatt [12] and Parzen [32]. 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

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

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

(K.1) $K(.)$ is of bounded variation on $\mathbb{R}$
(K.2) $K(.)$ 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 $K(.)$, let $f$ be twice continuously differentiable in a neighborhood of $x$, then

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

and

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

Then from (3) and (4) we have

$$MSE(\hat{f}_{n,h}(x)) = Bias^2(\hat{f}_{n,h}(x)) + Var(\hat{f}_{n,h}(x))$$

$$= \frac{1}{4} (f''(x))^2 h^4 \left[ \int u^2 K(u)du \right]^2 + \frac{1}{nh} f(x) \int K^2(u)du + o(h^4 + (nh)^{-1})$$

Devroy [33] showed that the optimized bandwidth is $h \sim O\left(n^{-\frac{1}{4}}\right)$ and then the optimal MSE is of the order $n^{-\frac{3}{4}}$. X.Xie, J.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.

$$\hat{f}_{n,h}^0(x) = \hat{f}_{n,h}(x) - \hat{Bias}(\hat{f}_{n,h}(x))$$

As result, the proposed estimator is

$$\hat{f}_{n,h}^0(x) = \hat{f}_{n,h}(x) - \frac{h^2}{2} \hat{f}_{n,h}'(x) \int u^2 K(u)du$$  \hfill (5)

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

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 $\hat{f}_{n,h}^0$. 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. (X. Xie, J. Wu [3]). Under 1), 2) and 3),

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

and

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

Consequently,

\[ \text{MSE}(\hat{f}_{n,h}(x)) = o(h^6 + (nh)^{-1}) \]

Under the regularity conditions on \( f, K \) and \( h \), the optimal MSE is of the order \( O\left(n^{-\frac{7}{6}}\right) \) with \( h = O\left(n^{-\frac{1}{2}}\right) \).

4. 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 [34], which is a generalization of Amari’s \( \alpha \)-divergence [35]. Both these divergences encompass the Kullback-Leibler(KL) divergence and its dual.
2. Bregman divergence [36, 37], which encompasses the euclidean(EUC) distance, the KL divergence and the Itakura Saito(IS) divergence.

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

Bregman (see [4, 5, 6, 7]) introduced for a convex subset of a Hilbert space \( S \) and \( \phi : S \to \mathbb{R} \) a continuously differentiable strictly convex function; the Bregman divergence \( D^\phi_{\phi} : S \times S \to \mathbb{R} \) as follows

\[ D^\phi_{\phi}(p, q) = \phi(p) - \phi(q) - \langle p - q, \nabla \phi(q) \rangle, \ \forall (p, q) \in S^2 \]  

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

\[ D^\phi_{\phi}(f(x), g(x)) := \int_X \left( \phi(f(x)) - \phi(g(x)) - (f(x) - g(x))\phi'(g(x)) \right) dx \]  

where \( X \) is a support of the two density functions, \( f \) and \( g \); and \( \phi'(.) \) the derivative of \( \phi(.) : \mathbb{R} \to \mathbb{R} \) respected to \( x \). On the other hand Basu al. [38], Minami, Eguchi [39] introduced the basic Beta-divergence and many researchers investigated their applications including [40, 41]. 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 (\( L_2 \)-norm) distance. Therefore one can check that the Beta-divergence can be generated from the Bregman divergence using the following strictly convex continuous function [40]

\[ \phi(t) = \begin{cases} 
\frac{c_1}{\beta(\beta-1)}t^\beta + c_2t + c_3, & \beta \neq 0, 1 \\
c_1t \log(t) + c_2t + c_3, & \beta = 1 \\
-c_1 \log(t) + c_2t + c_3, & \beta = 0
\end{cases} \]  

Here \( c_1, c_2 \) and \( c_3 \) are some constants.
Lemma 1. where 

\[ b \hat{A} \]

5. Consistency of the Bregman divergence estimator

P (Liese and Vajda [42]) we estimate \( \beta \)

we estimate \( f \) is theoretically known and satisfies: \( \int_{\mathbb{R}^{d_h}} f_0(x) dx \) is finite, we estimate \( D^B_\phi(f(x), f_0(x)) \) by

\[
\hat{D}_\phi^B(f_{n,h}(x), f_0(x)) = \int_{A_h} \phi(\hat{f}_{n,h}(x)) - \phi(f_0(x)) - (\hat{f}_{n,h}(x) - f_0(x))\phi'(f_0(x))dx.
\]

where \( A_h = \{ x \in \mathbb{R}^{d_h}, \hat{f}_{n,h}(x) \geq \gamma_n \} \) and \( \gamma_n \rightarrow 0 \) is a sequence of positive constants. In this following section, we will use the methods developed in [43] to establish convergence results for our estimator \( \hat{D}_\phi^B(f_{n,h}(x), f_0(x)) \).

5. Consistency of the Bregman divergence estimator

For proving such consistency results, one usually writes the difference \( \hat{D}_\phi^B(f_{n,h}(x), f_0(x)) - D^B_\phi(f(x), f_0(x)) \) as the sum of a probabilistic term \( \hat{D}_\phi^B(f_{n,h}(x), f_0(x)) - \mathbb{E}\hat{D}_\phi^B(f_{n,h}(x), f_0(x)), \) and a deterministic term

\[ \mathbb{E}\hat{D}_\phi^B(f_{n,h}(x), f_0(x)) - D^B_\phi(f_{n,h}(x), f_0(x)), \]

the so-called bias. Throughout the remainder of this paper, \( \mathbb{E}\hat{D}_\phi^B(f_{n,h}(x), f_0(x)) \) is given by

\[
\mathbb{E}\hat{D}_\phi^B(f_{n,h}(x), f_0(x)) := \int_{A_h} \left[ \mathbb{E}\phi(f_{n,h}(x)) - \phi(f_0(x)) - (\mathbb{E}f_{n,h}(x) - f_0(x))\phi'(f_0(x)) \right] dx,
\]

where \( A_h \) is defined in (11).

Lemma 1. Let \( K_\beta \) satisfy (K1)-(K4), let \( f(.) \) be a continuous bounded density, \( \phi \) be linear and strictly convex function 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 \( b_n \rightarrow 0 \) and \( n a_n / \log(n) \rightarrow \infty \) as \( n \rightarrow \infty \), we have with probability \( 1 \)

\[
\sup_{a_n \leq h \leq b_n} | \hat{D}_\phi^B(f_{n,h}(x), f_0(x)) - \mathbb{E}\hat{D}_\phi^B(f_{n,h}(x), f_0(x)) | = 0 \left( \sqrt{\log(1/a_n) \vee \log \log n} \right).
\]

PROOF. Define

\[
\Delta_n := \hat{D}_\phi^B(f_{n,h}(x), f_0(x)) - \mathbb{E}\hat{D}_\phi^B(f_{n,h}(x), f_0(x))
\]

We have

\[
| \Delta_n | = \left| \int_{A_h} \left[ \phi(\hat{f}_{n,h}(x)) - \phi(f_0(x)) - (\hat{f}_{n,h}(x) - f_0(x))\phi'(f_0(x)) \right] + \\
- \left[ \mathbb{E}\phi(\hat{f}_{n,h}(x)) - \phi(f_0(x)) - (\mathbb{E}\hat{f}_{n,h}(x) - f_0(x))\phi'(f_0(x)) \right] dx \right| \]

\[
\leq \left| \int_{A_h} \left[ \phi(\hat{f}_{n,h}(x)) - \mathbb{E}\phi(\hat{f}_{n,h}(x)) \right] dx + \left| \int_{A_h} (\hat{f}_{n,h}(x) - \mathbb{E}\hat{f}_{n,h}(x)) \phi'(f_0(x))dx \right| \right|
\]

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

\[
| \Delta_n | \leq \phi(\sup_{a_n \leq h \leq b_n} | f_{n,h}(x) - \mathbb{E}f_{n,h}(x) |) \int_{A_h} dx + \sup_{a_n \leq h \leq b_n} | f_{n,h}(x) - \mathbb{E}f_{n,h}(x) | \int_{A_h} \phi(f_0(x)) dx.
\]

1. \( D^B_\phi(kp, kq) = k^p D^B_\phi(p, q), \) for all \( p, q, k > 0 \) with \( \beta \) equal to 2, 1, 0.
For $0 < a_n < b_n \leq 1$, we have $\sup_{a_n \leq x \leq b_n} | f_{n,h}(x) - \mathbb{E} \hat{f}_{n,h}^b(x) | \leq \| f_{n,h}^b(x) - \mathbb{E} \hat{f}_{n,h}^b(x) \|_\infty$ where $\| \cdot \|_\infty$ denotes the supremum norm, i.e., $\| \psi \|_\infty := \sup_{x \in \mathbb{R}} \| \psi(x) \|.$ Therefore,

$$| \Delta_n | \leq \phi \left( \| f_{n,h}^b(x) - \mathbb{E} \hat{f}_{n,h}^b(x) \|_\infty \right) \int_{A_n} d\pi + \| f_{n,h}^b(x) - \mathbb{E} \hat{f}_{n,h}^b(x) \|_\infty \int_{A_n} \phi'(f_0(x)) dx.$$ 

Finally,

$$\sup_{a_n \leq x \leq b_n} | \Delta_n | \leq \phi \left( \sup_{a_n \leq x \leq b_n} \| f_{n,h}^b(x) - \mathbb{E} \hat{f}_{n,h}^b(x) \|_\infty \right) \int_{A_n} d\pi + \sup_{a_n \leq x \leq b_n} \| f_{n,h}^b(x) - \mathbb{E} \hat{f}_{n,h}^b(x) \|_\infty \int_{A_n} \phi'(f_0(x)) dx.$$ 

Whenever $K(.)$ is measurable and satisfies (K3)-(K4) and by the remark 2 in [44], when $f(.)$ is bounded, for each pair of sequence $(a_n)_{n \geq 1}$ and $(b_n)_{n \geq 1}$ 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

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

(13)

Since $\int_{\mathbb{R}} d\pi < \infty$ and $\int_{\mathbb{R}} \phi'(f_0(x)) dx < \infty$, in view of (12) and (13), we obtain with probability 1,

$$\sup_{a_n \leq x \leq b_n} | \Delta_n | = 0 \left( \sqrt{\log(1/a_n) \vee \log \log n} / na_n \right) + 0 \left( \sqrt{\log(1/a_n) \vee \log \log n} / na_n \right).$$

Thus

$$\sup_{a_n \leq x \leq b_n} | \Delta_n | = 0 \left( \sqrt{\log(1/a_n) \vee \log \log n} / na_n \right).$$ 

(14)

It concludes the proof of the lemma.

**Lemma 2.** Let $K(.)$ satisfy (K1)-(K4), let $f(.)$ be a continuous bounded density, $\phi$ linear and strictly convex function 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 $b_n \to 0$ as $n \to \infty$,

$$\sup_{a_n \leq x \leq b_n} | \mathbb{E} \hat{D}_{\phi}^b(f_{n,h}(x), f_0(x)) - D_{\phi}^b(f(x), f_0(x)) | = 0 (b_n).$$

**Proof.** Let $\Delta_n = \mathbb{E} \hat{D}_{\phi}^b(f_{n,h}(x), f_0(x)) - D_{\phi}^b(f(x), f_0(x))$, therefore

$$| \Delta_n | = \int_{A_n} \left[ \mathbb{E} \phi(f_{n,h}(x)) - \phi(f_0(x)) - (\mathbb{E} \hat{f}_{n,h}^b(x) - f_0(x)) \phi'(f_0(x)) \right] dx + \int_{A_n} \phi(f(x)) - \phi(f_0(x)) - (f(x) - f_0(x)) \phi'(f_0(x)) dx$$

Repeat the arguments above in the terms $| \Delta_n |$ with the formal change of $f_{n,h}$ by $f$. We show that, for any $n \geq 1$,

$$| \Delta_n | \leq \phi \left( \sup_{a_n \leq x \leq b_n} \| f_{n,h}^b(x) - f(x) \|_\infty \right) \int_{A_n} dx + \sup_{a_n \leq x \leq b_n} \| f_{n,h}^b(x) - f(x) \|_\infty \int_{A_n} \phi'(f_0(x)) dx,$$
which implies
\[
| \Delta a_2 | \leq \phi \left( \sup_{a_n \leq h \leq b_n} | f_{n,h}^b(x) - f(x) | \right) \int_{R^{n_0}} dx + \sup_{a_n \leq h \leq b_n} | f_{n,h}^b(x) - f(x) | \int_{R^{n_0}} \phi (f_0(x)) dx.
\]  
(15)

In [12], when the density \( f(.) \) 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 \( n \to \infty \),
\[
\sup_{a_n \leq h \leq b_n} \| f_{n,h}^b(x) - f(x) \|_\infty = 0(b_n).
\]  
(16)

Thus,
\[
\sup_{a_n \leq h \leq b_n} \ | \Delta a_2 | \leq \phi \left( \sup_{a_n \leq h \leq b_n} \| f_{n,h}^b(x) - f(x) \|_\infty \right) \int_{R^{n_0}} dx + \sup_{a_n \leq h \leq b_n} \| f_{n,h}^b(x) - f(x) \|_\infty \int_{R^{n_0}} \phi (f_0(x)) dx,
\]
where \( \int_{R^{n_0}} dx \) and \( \int_{R^{n_0}} \phi (f_0(x)) dx \) are finite. Then, in view of (16),
\[
\sup_{a_n \leq h \leq b_n} \ | \Delta a_2 | = 0(b_n) + 0(b_n),
\]
Finally,
\[
\sup_{a_n \leq h \leq b_n} \ | \Delta a_2 | = 0(b_n)
\]  
(17)
is deduced the proof of the lemma.

**Theorem 3.** Let \( K(.) \) satisfy (K3)-(K4), \( f(.) \) be a uniform, bounded and continuous density and \( \phi \) linear and strictly convex function and satisfy 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 \( n a_n / \log(n) \to \infty \) as \( n \to \infty \), we have with probability 1
\[
\sup_{a_n \leq h \leq b_n} \ | \delta a_n | \leq 0 \left( \sqrt{\log(1/a_n) \vee \log n} / n a_n \right). \]

**Proof.** We have
\[
| \delta a_n | \leq | \delta a_n (f_{n,h}^b(x), f_0(x)) | - | \delta a_n (f_{n,h}^b(x), f_0(x)) - E \delta a_n (f_{n,h}^b(x), f_0(x)) | + | E \delta a_n (f_{n,h}^b(x), f_0(x)) - \delta a_n (f(x), f_0(x)) |.
\]

Combinating the Lemma (1) and (2), we obtain
\[
\sup_{a_n \leq h \leq b_n} \ | \delta a_n (f_{n,h}^b(x), f_0(x)) | = 0 \left( \sqrt{\log(1/a_n) \vee \log n} / n a_n \right) + 0(b_n).
\]  
(18)

This entails that, as \( n \to \infty \),
\[
\sup_{a_n \leq h \leq b_n} \ | \delta a_n (f_{n,h}^b(x), f_0(x)) | \to 0.
\]

It concludes the proof of the Theorem.
6. Asymptotic behavior of Bregman divergence estimator

Recall the hypothesis testing \( H_0 : F_\theta = F \) against \( H_1 : F_\theta \neq F \).

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

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

\[
2n\hat{D}^b_{\theta}(\hat{F}_{n,h}^b, F_\theta) \overset{d}{\to} \sum_{i=1}^{r} \beta_i Z_i^2 + \sum_{j=1}^{s} \alpha_j Z_j^2,
\]

when \( n \to \infty \). Where \( Z_i, i = 1, \ldots, r \) and \( Z_j, j = 1, \ldots, s \) are iid normal variables with mean zero and variance 1; we assume that \( r = s \). \( \beta_i, i = 1, \ldots, r \) are the non null eigenvalues of the matrix \( HH^T \), \( \alpha_j, j = 1, \ldots, s \) are the non null eigenvalues of the matrix \( BS_r \), \( r = \text{rank}(\Sigma_F H \Sigma_{F,\theta}) \), and \( s = \text{rank}(\Sigma_F B \Sigma_{F,\theta}) \), being \( \Sigma_F = \text{diag}(F) - FF^T \) and \( \Sigma_M = \text{diag}(F) - FF^T \) and

\[
H = \left( \frac{\partial^2}{\partial f_i \partial f_j} D^b_{\theta}(F, F_\theta) \right)_{i,j=1,\ldots,M_0}, \quad B = \left( \frac{\partial^2}{\partial f_i \partial f_j} \hat{D}^b_{\theta}(F, F_\theta) \right)_{i,j=1,\ldots,M_0}.
\]

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

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

One can write

\[
2n\hat{D}^b_{\theta}(\hat{F}_{n,h}^b, F_\theta) = \sqrt{n} (\hat{F}_{n,h}^b - F)^T B \sqrt{n} (\hat{F}_{n,h}^b - F) + \sqrt{n} (F_\theta - F)^T H \sqrt{n} (F_\theta - F) + 2n o(|| F_{n,h} - F ||^2 + || F_\theta - F ||^2).
\]

And \( \sqrt{n} (F_\theta - F_\theta) \overset{d}{\to} N(0, \Sigma_{F,\theta}) \), when \( n \to \infty \); then \( || F_\theta - F_\theta ||^2 = o_p \left( n^{-1} \right) \). Therefore \( 2n o(|| F_{n,h} - F ||^2 + || F_\theta - F_\theta ||^2) = o_p(1) \). The random variables \( 2n\hat{D}^b_{\theta}(\hat{F}_{n,h}^b, F_\theta) \) and \( \sqrt{n} (\hat{F}_{n,h}^b - F)^T B \sqrt{n} (\hat{F}_{n,h}^b - F) + \sqrt{n} (F_\theta - F)^T H \sqrt{n} (F_\theta - F) \) have the same asymptotic distribution. Now by corollary 2.1 in Dik and Gunst \([45]\) 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 \( F_\theta, \forall \theta \)

\[
1) \sqrt{n} (\hat{\theta} - \theta_0) \overset{d}{\to} N(0, I(\theta_0)^{-1}) \quad \text{where} \quad I(\theta_0) \text{ is Fisher information}
\]

\[
2) F_\theta \overset{d}{\to} F_{\theta_0} \quad \text{when} \quad n \to \infty.
\]

- \((A_2)\) There exists \( \theta \in \Theta, \lambda^* = \left( \begin{array}{c} \lambda_{11} \\ \lambda_{21} \\ \lambda_{12} \\ \lambda_{22} \end{array} \right) \), with \( \lambda_{12} = \lambda_{21} \) and such that

\[
\sqrt{n} \left( \begin{array}{c} \hat{F}_{n,h}^b - F \\ F_\theta - F_\theta \end{array} \right) \overset{d}{\to} N(0, \lambda^*).
\]
Theorem 5. Under $H_1: F_{\theta} \neq F$ and we assume that the conditions (A1), (A2) hold, we have:

$$\sqrt{n} \left[ D_{\phi}^{b}(\hat{F}_{n,h}, F_{\theta}) - D_{\phi}^{b}(F, F_{\theta}) \right] \overset{L}{\rightarrow} N(0, \lambda_{\phi}^2)$$

where

$$\lambda_{\phi}^2 = K^T \wedge_{11} K + K^T \wedge_{12} N + N^T \wedge_{12} K + N^T \wedge_{22} N \tag{19}$$

$K^T = (k_1, \ldots, k_{M_0})$ with

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

$N^T = (n_1, \ldots, n_{M_0})$ with

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

Proof. A first order Taylor expansion gives

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

One can write

$$\sqrt{n} \left[ D_{\phi}^{b}(\hat{F}_{n,h}, F_{\theta}) - D_{\phi}^{b}(F, F_{\theta}) \right] = \sqrt{n} \left[ K^T(\hat{F}_{n,h} - F) + N^T(F_{\theta} - F_{\theta}) \right] + \sqrt{n}o(\parallel F_{n,h} - F \parallel + \parallel \hat{F}_{n,h} - F \parallel).$$

Since $\sqrt{n}(F_{\theta} - F_{\theta}) \overset{L}{\rightarrow} N(0, \Sigma_{F_{\theta}})$, then $\parallel F_{\theta} - F_{\theta} \parallel = o_p(1)$ and $\sqrt{n}o(\parallel F_{\theta} - F_{\theta} \parallel) = o_p(1)$. Therefore $\sqrt{n}o(\parallel F_{n,h} - F \parallel + \parallel \hat{F}_{n,h} - F \parallel) = o_p(1)$.

Hence

$$\sqrt{n} \left[ D_{\phi}^{b}(\hat{F}_{n,h}, F_{\theta}) - D_{\phi}^{b}(F, F_{\theta}) \right] = \sqrt{n} \left[ K^T(\hat{F}_{n,h} - F) + N^T(F_{\theta} - F_{\theta}) + o_p(1) \right] \overset{L}{\rightarrow} N(0, \lambda_{\phi}^2)$$

The random variables $\sqrt{n} \left[ D_{\phi}^{b}(\hat{F}_{n,h}, F_{\theta}) - D_{\phi}^{b}(F, F_{\theta}) \right]$ and $\sqrt{n} \left[ K^T(\hat{F}_{n,h} - F) + N^T(F_{\theta} - F_{\theta}) \right]$ have the same asymptotic distribution. In view of A1 and A2 we have

$$\sqrt{n} \left[ K^T(\hat{F}_{n,h} - F) + N^T(F_{\theta} - F_{\theta}) \right] \overset{L}{\rightarrow} N(0, \lambda_{\phi}^2)$$

where $\lambda_{\phi}^2$ is given by (19). This completes the proof.

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

$$\beta_{n,\theta}(F) = 1 - \Phi_{\alpha} \left( \frac{t_{\alpha} - 2nD_{\phi}^{b}(F, F_{\theta})}{2\sqrt{n}\lambda_{\phi}} \right); \tag{20}$$

for a sequence of distribution function $\Phi_{\theta}(x)$ tending uniformily to the standard normal distribution function $\Phi(x)$; $t_{\alpha}$ is the critical value of $T_{\theta} = 2nD_{\phi}^{b}(\hat{F}_{n,h}, F_{\theta})$ and $\lambda_{\phi}$ is given in thereom 5.

Thus thanks to the goodness-of-fit test, it is possible to chose 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.

9
7. Model Selection and Bregman Divergence Based Test Statistic

Consider the situation in which we have two candidate parametric models \( F_\theta \) and \( F_\gamma = \{ F(\cdot, \gamma); \gamma \in \Gamma \subseteq \mathbb{R}^m \} \) another candidate model. We would like to choose the best of two candidate models based on their discrimination statistic between the observations and models \( F_\theta \) and \( F_\gamma \), defined respectively as follows \( \hat{D}_\phi(F_{n,h}, F_\theta) \) and \( \hat{D}_\phi(F_{n,h}, F_\gamma) \).

Our major work is to propose some tests for model selection, i.e. for the null hypothesis

\[ H_0: \hat{D}_\phi(F_{n,h}, F_\theta) = \hat{D}_\phi(F_{n,h}, F_\gamma) \]

Under the assumptions of theorem 5, suppose that \( \kappa \).

\[ \text{Theorem 6. (Asymptotic distribution of the U-statistic).} \]

Under the assumptions of theorem 5 suppose that \( \kappa \neq 0 \), then under the null hypothesis \( H_0 \), \( U \to N(0,1) \).

\[ U = \frac{\sqrt{n}}{\hat{\kappa}} \left[ \hat{D}_\phi(F_{n,h}, F_\theta) - \hat{D}_\phi(F_{n,h}, F_\gamma) \right] \]
The MLE \( \hat{P} \) provides us an estimator \( n \) when the density function (PDF) of \( F_{93.12, 98.64, 105.12, 105.84, 127.92, 128.04, 173.40} \). Here we consider Gamma model as the component of vector of groove ball bearings. The data are number of million revolutions before failure for each of the lifetime tests and they test whether the data come from a Gamma or a Log-Normal. The data given arose in tests on endurance of deep data set 8.

Example with real 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 (as given by Lieblein and Zelen (1956) for the lifetime) 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 \( F_\theta \) and log-normal as the component of \( F_\gamma \) defined respectively in section 7. 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

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

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

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

The MLE \( \hat{\alpha}_n \) can not be easy to compute because its construction requires the solution of the maximum likelihood equation (MLEq)

\[
\sum_{j=1}^n \ln(\hat{\eta}_n X_j) - n \frac{\Gamma(\hat{\alpha}_n)}{\Gamma(\alpha_n)} = 0 \quad \text{and} \quad \sqrt{n}(\hat{\alpha}_n - \alpha) \xrightarrow{\mathcal{L}} N\left(0, I(\alpha)^{-1}\right),
\]

when \( n \to \infty \). The sequel dot means derivation w.r.t. \( \alpha \). We introduce the multi-step MLE-process [47], which in this case provides us an estimator \( \hat{\alpha}_n^* \) such that \( \sqrt{n}(\hat{\alpha}_n^* - \alpha) \xrightarrow{\mathcal{L}} N\left(0, I(\alpha)^{-1}\right) \) 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 \( \tilde{\alpha}_N \).
the preliminary estimator constructed by the first \( N = \lfloor n^\delta \rfloor \) observations \( X^k = (X_1, \ldots, X_k) \) with \( \delta = \left( \frac{1}{2}, 1 \right) \). Then the one-step MLE-process \( \alpha_n^* = (\alpha_{k,n}^*) \) \( N + 1 \leq k \leq n \) is defined by the equality

\[
\alpha_{k,n}^* = \bar{\alpha}_n + I(\bar{\alpha}_n)^{-1} \frac{1}{k} \sum_{j=N+1}^{k} l(X_j, \bar{\alpha}_n),
\]

for \( k = \lfloor sn \rfloor \), \( s \in (0, 1] \); we have the convergence

\[
\sqrt{K}(\alpha_n^* - \alpha_0) \xrightarrow{L} N(0, I(\alpha_0)^{-1}),
\]

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

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

\[
\hat{\alpha} = \bar{\alpha} + \frac{1}{n\hat{\alpha}(\bar{\alpha})} \sum_{i=1}^{n} \left[ \ln(\hat{\alpha}X_i) - \frac{1}{\hat{\alpha}(\bar{\alpha})} \right].
\]

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

where \( I(\bar{\alpha}) = \left( \frac{\Gamma(\bar{\alpha})(\bar{\alpha})^2}{\Gamma(\bar{\alpha})^2} \right) \) with \( \Gamma(\bar{\alpha}) = \int_0^\infty (ln x)^{\bar{\alpha}} x^{\bar{\alpha}-1} e^{-x} dx \), \( \bar{\alpha}(\bar{\alpha}) = \int_0^\infty (ln x)^2 x^{\bar{\alpha}-1} e^{-x} dx \) and \( \alpha_0 \) the true value of \( \alpha \).

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

\[
f_{LN}(x; \mu, \sigma^2) = \left\{ \begin{array}{ll}
\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(ln(x) - \mu)^2}{2\sigma^2}} & x \geq 0, \mu > 0, \sigma > 0 \\
0 & x < 0
\end{array} \right.
\]

The MLE of \( \mu \) and \( \sigma \) 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 \( \ref{eq:2a} \) and \( \ref{eq:10} \) allow us to compute the \( \hat{D}_\beta(\hat{\beta}_n^h(x), f_{GA}(x)) \) and \( \hat{D}_\beta(\hat{\beta}_n^h(x), f_{LN}(x)) \) as follows

\[
\hat{D}_\beta(\hat{\beta}_n^h(x), f_{GA}(x)) = \frac{1}{6} \int_0^{\infty} \left( (f_{n,h}^h(x))^3 - 3 f_{n,h}^h(x) f_{GA}^2(x) + 2 f_{GA}^3(x) \right) dx
\]

and

\[
\hat{D}_\beta(\hat{\beta}_n^h(x), f_{LN}(x)) = \frac{1}{6} \int_0^{\infty} \left( (f_{n,h}^h(x))^3 - 3 f_{n,h}^h(x) f_{LN}^2(x) + 2 f_{LN}^3(x) \right) dx
\]

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

\[
f_{n,h}^h(x) = \frac{1}{2\sqrt{2\pi\sqrt{n}h}} \sum_{i=1}^{n} \left[ 3 - \left( \frac{X_i - \bar{X}}{h} \right)^2 \right] e^{-\frac{(X_i - \bar{X})^2}{2h^2}}.
\]

To get \( h \) optimal, the cross-validation method introduced by Rudemo (1982) and Bowman (1984) giving the simple and attractive smoothing parameter is used. Hence \( h = h_{CV} = \arg \min_{h>0} CV(h) \) where \( CV(h) \) is cross-validation given by

\[
CV(h) = \int f^2 dx - \frac{h}{2} \sum_{j=1}^{N} f_{n,h}^h(X_j) - \frac{1}{h} \sum_{j=i}^n K \left( \frac{X_j - X_i}{h} \right).
\]

\( f_{GA}(x) \) and \( f_{LN}(x) \) are parametric estimators of Gamma and log-normal models and are given by \( f_{GA}(x) = \frac{\bar{\gamma}}{\Gamma(\bar{\gamma})} x^{\bar{\gamma}-1} e^{-x/\bar{\gamma}} \) and \( f_{LN}(x) = \frac{1}{\sqrt{2\pi\bar{\sigma}}} e^{-\frac{(ln(x) - \bar{\mu})^2}{2\bar{\sigma}^2}} \).
For the data at hand, Ali-Akbar Bromideh and Reza Valizadeh (2013) 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 \( \hat{D}_B^1 \equiv \hat{D}_B^1(\hat{f}_b(x), \hat{f}_{GA}(x)) = 0.000021 \) and \( \hat{D}_B^2 \equiv \hat{D}_B^2(\hat{f}_b(x), \hat{f}_{LN}(x)) = 0.000024 \). Bregman divergence is the non-symmetric measure of the difference (dissimilarity) between two probability distributions. 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. The Figure 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 1: The histogram with Log-Normale and Gamma density functions for the given data set.](image1)

![Figure 2: The histogram and two fitted density functions for the given data set.](image2)
8.2. 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 sub-section 8.1. Hence the Data Generating Process (DGP) has density

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

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 \( \pi \) 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 \( \pi = 1.00 \) 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 \( \hat{D}^1_B \equiv \hat{D}^1_B(f_{\hat{f}_{n,b}(x), n,b}, f_{\hat{f}_{GA}(x)}) \) and \( \hat{D}^2_B \equiv \hat{D}^2_B(f_{\hat{f}_{n,b}(x), n,b}, f_{\hat{f}_{LN}(x)}) \). 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 analyse 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)   |
| \( \hat{D}^B_{\hat{f}_{\pi}} \) | 0.000031   | 0.000028   | 0.000026   | 0.000026   | 0.000026   |
|     | (0.000015) | (0.000009) | (0.000007) | (0.000006) | (0.000006) |
| \( \hat{D}^B_{\phi_{\pi}} \) | 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 = $\text{Gamma}(4.02804, 0.05576722)$

| 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 |
|     | (1.120811) | (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) | (0.037444) | (0.0432) |
| $\hat{D}_{\phi_1}^B$ | 0.000027 | 0.000024 | 0.000023 | 0.000023 | 0.000022 |
|     | (0.000012) | (0.000008) | (0.000009) | (0.000006) | (0.00000) |
| $\hat{D}_{\phi_2}^B$ | 0.000030 | 0.000027 | 0.000026 | 0.000026 | 0.000025 |
|     | (0.000012) | (0.000009) | (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 = 0.25$\text{Gamma}(4.02804, 0.05576722) + 0.75\text{log-normal}(4.150614, 0.5214847)$

| n  | 20     | 40     | 60     | 80     | 90     |
|----|--------|--------|--------|--------|--------|
| $\hat{\alpha}$ | 7.2398 | 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) | (0.039236) | (0.041885) | (0.0314) |
| $\hat{D}_{\phi_1}^B$ | 0.000051 | 0.000040 | 0.000039 | 0.000039 | 0.000036 |
|     | (0.000015) | (0.000012) | (0.000011) | (0.000009) | (0.000008) |
| $\hat{D}_{\phi_2}^B$ | 0.000037 | 0.000036 | 0.000034 | 0.000034 | 0.000034 |
|     | (0.000016) | (0.000012) | (0.000011) | (0.000009) | (0.000008) |
| $U$ | -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 = 0.5Gamma(4.02804, 0.05576722) + 0.5log-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) | (0.042608) | (0.034347) | (0.031361) | (0.02839) |
| \(\hat{D}_{\phi_1}\) | 0.000046 | 0.000042 | 0.000041 | 0.000041 | 0.000040 |
|     | (0.000017) | (0.000012) | (0.000010) | (0.000009) | (0.000008) |
| \(\hat{D}_{\phi_2}\) | 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 | 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% 9.9% 9.9% 6.6% |

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

| n  | 20      | 40      | 60      | 80      | 200     |
|----|---------|---------|---------|---------|---------|
| \(\hat{\alpha}\) | 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) | (0.020713) |
| \(\hat{D}_{\phi_1}\) | 0.000040 | 0.000037 | 0.000036 | 0.000035 | 0.000035 |
|     | (0.000025) | (0.000011) | (0.000009) | (0.000008) | (0.000005) |
| \(\hat{D}_{\phi_2}\) | 0.000043 | 0.000040 | 0.000039 | 0.000038 | 0.000038 |
|     | (0.000024) | (0.000011) | (0.000010) | (0.000008) | (0.000000) |
| \(U\) | -0.654543 | -1.296389 | -2.172371 | -2.545976 | -3.875059 |
|     | (0.000005) | (0.000002) | (0.000002) | (0.000002) | (0.000001) |
| Model selection | 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 $\hat{D}_{\phi_1}^B$ and $\hat{D}_{\phi_2}^B$, 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 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 $\hat{D}_{\phi_1}^B$ and $\hat{D}_{\phi_2}^B$ 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. The 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 approximates the data. This is because these distributions are often interchangeable and commonly used to model certain lifetimes in reliability and survival analysis (Wiens, [46]).

**Figure 1. Histogram of DGP**

$= \text{Log-normal}(4.150614, 0.5214847)$,

with $n=60$ and $\pi = 0$.

**Figure 2. Histogram of DGP**

$= \text{Gamma}(4.02804, 0.05576722)$,

with $n=60$ and $\pi = 1$. 
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) does 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 are nearly similar.
9. 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 for model selection based on Bregman divergence type statistics that use the corresponding multi-step MLE process estimators. The tests 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. We have established a fundamental property such as the strong consistency of the Bregman divergence estimator. To facilitate the choice, we have used the bias reduced kernel density estimator to assure the improvement on convergence rate to the true distribution. For model selection procedure based on divergence measures, the Bregman divergence criterion performs well especially in small sample.

References

References

[1] Bromideh Ali-akbar, discriminating between Weibull and Log-Normal distribution based on Kullback-Leibler divergence, 44-54, 2012.
[2] Bromideh Ali Akbar, Valizadeh Reza, Discrimination between gamma and log-normal distributions by ratio of minimized Kullback-Leibler divergence, Pakistan Journal of Statistics and Operation Research, vol.9, issue 4, 441-451, 2013.
[3] Aiaaon Xie, Jingjing Wu Some Improvement on Convergence Rates of Kernel Density Estimator Published Online June 2014 in SciRes. [http://www.scirp.org/journal/entropy, http://dx.doi.org/10.4236/entropy.2014.511161] Applied Mathematics, 2014, 5, 1684-1696.
[4] Pengwen Chen, Yunmei and Murali Rao, Metrics defined by Bregman Divergence, Internat. Press, vol.6, No. 4, pp. 927-948, 2008.
[5] Wolfgang Stummer and Igor Vajda, On Bregman Distances and Divergences of Probability Measures, Fellow, IEEE.
[6] Inderjit S. Dhillon and Suvrit Sra. Generalized nonnegative matrix approximations with Bregman divergences. In Y. Weiss, B. Schlkopf, and J. Platt, editors, Proceedings of the 20th Annual Conference on Neural Information Processing Systems (NIPS'07), pp. 356-364, Vancouver, Canada, December 2007.
[7] Andrzej Cichocki, Sergio Cruces, and Shin-ichi Amari Generalized Alpha-Beta Divergences and Their Application to Robust Nonnegative Matrix Factorization entropy ISSN 1099-4300 www.mdpi.com/journal/entropy Received: 13 December 2010; in revised form: 4 January 2011 / Accepted: 4 January 2011 / Published: 14 January 2011.
[8] Papa Ngom, Hamza Dhaker, El Hadji Deme and Pierre Mendy, Kernel-Type Estimators of Divergence Measures and Its Strong Uniform Consistency, American Journal of Theoretical and Applied Statistics ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online), Published online February 16, 2016.
[9] F. Aakura and S. Saito, Analysis Synthesis Telephony Based Upon Maximum Likelihood Method, Repts. of the 6th International. Cong. Acoust., Y. Kohbasi, ed., Tokyo, C-5-5, C17-20, 1968.
[10] Andrzej Cichocki 1,2, and Shin-ichi Amari Fannilies of Alpha- Beta- and Gamma- Divergences: Flexible and Robust Measures of Similarities entropy ISSN 1099-4300 www.mdpi.com/journal/entropy Received: 26 April 2010 / Accepted: 1 June 2010 / Published: 14 June 2010
[11] Solomon Kullback and Richard A. Leibler. On information and sufficiency. The Annals of Mathematical Statistics, 22(1):79-86, March 1951.
[12] Rosenblatt, M. (1956) On Estimation of a Probability Density Function and the Mode. The Annals of Mathematical Statistics, 33, 1065-1076. [http://dx.doi.org/10.1214/aoms/1177704772]
[13] Dey, Arabin Kumar Kundu, Debasis, Discriminating among the log-normal, weibull, and generalized exponential distributions, IEE Transactions on Reliability, vol.58, issue 3, 416-424, 2009.
[14] Kundu, D.Manglick, A., Discriminating between the Weibull and Log- Normal distributions, Naval Research Logistics, vol.51, 893-905, 2004.
[15] Dey, Arabin Kumar Kundu, Debasis, Discriminating Between the Weibull and Log-normal Distributions for Type-II Censored Data, issue 1, 1-28.
[16] Kundu, Debasis Manglick, Anubhav, Discriminating Between The Log-normal and Gamma Distributions, Journal of the Applied Statistical Sciences, vol.14,175-187, 2005.
[17] Alzaid, A.Sultan, K. S., Discriminating between gamma and log-normal distributions with applications, Journal of King Saud University - Sciences, vol.21, issue 2.99-108, 2009.
[18] Elsherpieny, Elsayed A Muhammed, Hiba Z Radwan, Noha U, Discriminating Between Weibull and Log- Logistic Distributions in Presence of Progressive Type II Censoring, 7281-7290, 2015.
[19] Gupta, Rameshwar D., Discriminating between the Weibull and Log-Normal distributions, American Journal of Theoretical and Applied Statistics, volume 2, number 3, pages 927-948, 2013.
[20] Dey, Arabin Kumar Kundu, Debasis, Discriminating between Gamma and Generalized exponential distributions, vol.74, issue 3, 293-302, 2010.
