Divergence Measures Estimation and Its Asymptotic Normality Theory Using Wavelets Empirical Processes I

Amadou Diadié Ba
LERSTAD, Gaston Berger University,
Saint-Louis, SENEGAL
amadou-diadie.ba@edu.ugb.en

Gane Samb LO
LERSTAD, Gaston Berger University, Saint-Louis, SENEGAL *
Associate Researcher, LASTA, Pierre et Marie University, Paris, FRANCE
Assiated Professor, African University of Sciences and Technology, Abuja, NIGERIA
gane-samb.lo@ugb.edu.sn, gslo@aust.edu.ng

Diam Ba
LERSTAD, Gaston Berger University,
Saint-Louis, SENEGAL
diam.ba@edu.ugb.en

Received 8 May 2017

Accepted 21 December 2017

We deal with the normality asymptotic theory of empirical divergences measures based on wavelets in a series of three papers. In this first paper, we provide the asymptotic theory of the general of ϕ-divergences measures, which includes the most common divergence measures : Renyi and Tsallis families and the Kullback-Leibler measures. Instead of using the Parzen nonparametric estimators of the probability density functions whose discrepancy is estimated, we use the wavelets approach and the geometry of Besov spaces. One-sided and two-sided statistical tests are derived. This paper is devoted to the foundations the general asymptotic theory and the exposition of the mains theoretical tools concerning the ϕ-forms, while proofs and next detailed and applied results will be given in the two subsequent papers which deal important key divergence measures and symmetrized estimators.

Keywords: Divergence measures estimation; Asymptotic normality; Wavelet theory; wavelets empirical processes; Besov spaces.

2000 Mathematics Subject Classification: 62G05; 62G20; 62G07

*1178, Evanston Drive,NW, Calgary, Canada, T3P 0J9

1. Introduction

1.1. *General Introduction*

In this paper, we deal with divergence measures estimation using essentially wavelets density function estimation. Let $\mathscr{P}$ be a class of probability measures on $\mathbb{R}^d$, $d \geq 1$, a divergence measure on $\mathscr{P}$ is a function

$$\begin{array}{cccc} \mathscr{D}: & \mathscr{P}^2 & \longrightarrow & \overline{\mathbb{R}} \\ & (\mathbb{Q},\mathbb{L}) & \longmapsto & \mathscr{D}(\mathbb{Q},\mathbb{L}) \end{array} \tag{1.1}$$

such that $\mathscr{D}(\mathbb{Q},\mathbb{Q}) = 0$ for any $\mathbb{Q}$ such that $(\mathbb{Q},\mathbb{Q})$ in the domain of application of $\mathscr{D}$.

The function $\mathscr{D}$ is not necessarily an application. And if it is, it is not always symmetrical and it does neither have to be a metric. A great number of them are based on probability density functions (*pdf*). So let us suppose that any $\mathbb{Q} \in \mathscr{P}$ admits a *pdf* $f_{\mathbb{Q}}$ with respect to a σ-finite measure ν on $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$, which is usually the Lebesgue measure λ_d (with $\lambda_1 = \lambda$) or a counting measure on $\mathbb{R}^d$.

We may present the following divergence measures.

(1) The L_2^2-divergence measure :

$$\mathscr{D}_{L_2}(\mathbb{Q},\mathbb{L}) = \int_{\mathbb{R}^d} (f_{\mathbb{Q}}(x) - f_{\mathbb{L}}(x))^2 d\nu(x). \tag{1.2}$$

(2) The family of Renyi's divergence measures indexed by $\alpha \neq 1$, $\alpha > 0$, known under the name of Renyi-α :

$$\mathscr{D}_{R,\alpha}(\mathbb{Q},\mathbb{L}) = \frac{1}{\alpha - 1} \log \left(\int_{\mathbb{R}^d} f_{\mathbb{Q}}^{\alpha}(x) f_{\mathbb{L}}^{1-\alpha}(x) d\nu(x) \right). \tag{1.3}$$

(3) The family of Tsallis divergence measures indexed by $\alpha \neq 1$, $\alpha > 0$, also known under the name of Tsallis-α :

$$\mathscr{D}_{T,\alpha}(\mathbb{Q},\mathbb{L}) = \frac{1}{\alpha - 1} \left(\int_{\mathbb{R}^d} f_{\mathbb{Q}}^{\alpha}(x) f_{\mathbb{L}}^{1-\alpha}(x) - 1 \right) d\nu(x); \tag{1.4}$$

(4) The Kullback-Leibler divergence measure

$$\mathscr{D}_{KL}(\mathbb{Q},\mathbb{L}) = \int_{\mathbb{R}^d} f_{\mathbb{Q}}(x) \ \log(f_{\mathbb{Q}}(x)/f_{\mathbb{L}}(x)) \ d\nu(x). \tag{1.5}$$

The latter, the Kullback-Leibler measure, may be interpreted as a limit case of both the Renyi's family and the Tsallis' one by letting $\alpha \to 1$. As well, for α near 1, the Tsallis family may be seen as derived from $\mathscr{D}_{R,\alpha}(\mathbb{Q},\mathbb{L})$ based on the first order expansion of the logarithm function in the neighborhood of the unity.

From this small sample of divergence measures, we may give the following remarks.

(a) The L_2^2-divergence measure is both an application and a metric on $\mathscr{P}^2$, where $\mathscr{P}$ is the class of probability measures on $\mathbb{R}^d$ such that

$$\int_{\mathbb{R}^d} f_{\mathbb{Q}}^2(x)\, d\nu(x) < +\infty.$$

(b) For example, for both the Renyi and the Tsallis families, we may have integrability problems and lack of symmetry. From this sample tour, we have to be cautious, when speaking about divergence measures as applications and/or metrics. In the most general case, we have to consider the divergence measure between two specific probability measures as a number or a real parameter.

Originally, divergence measures came as extensions and developments of information theory that was first set for discrete probability measures. In such a situation, the boundedness of these discrete probability measures above zero and below $+\infty$ was guaranteed. That is, the following assumption holds :

Boundedness Assumption (BD). There exist two finite numbers $0 < \kappa_1 < \kappa_2 < +\infty$ such that

$$\kappa_1 \leq f_{\mathbb{Q}}, f_{\mathbb{L}} \leq \kappa_2. \tag{1.6}$$

If Assumption (1.6) holds, we do not have to worry about integrability problems, especially for Tsallis, Renyi and Kullback-Leibler measures, in the computations arising in the estimation theories. This explains why Assumption (1.6) is systematically used in a great number of works in that topic, for example, in [Singh and Poczos (2014)], [Krishnamurthy *et al.*(2014)], [Hall(1987)], to cite a few. But instead of Assumption (1.6), we use the following

Modified Boundedness Condition : There exist $0 < \kappa_1 < \kappa_2 < +\infty$ and a **compact** domain D as large as possible such that

$$\kappa_1 \leq f_{\mathbb{Q}} 1_D, f_{\mathbb{L}} 1_D \leq \kappa_2. \tag{1.7}$$

This implies that the modified divergence measure, denoted by $\mathscr{D}^{(m)}$, is applied to the modified **pdf**'s :

$$f_{\mathbb{Q}}^{(m)} = D_1^{-1} f_{\mathbb{Q}} 1_D \text{ and } f_{\mathbb{P}}^{(m)} = D_2^{-1} f_{\mathbb{L}} 1_D,$$

where D_1 and D_2 are the integrals of $f_{\mathbb{Q}}$and $f_{\mathbb{L}}$ of D, respectively. Based on this technique, that we apply in case of integrability problems, we will suppose, when appropriate, **that Assumption (1.6) holds on a compact set** D.

Although we are focusing on the aforementioned divergence measures in this paper, it is worth mentioning that there exist quite a few number of them. Let us cite for example the ones named after : Ali-Silvey or f-divergence [Topsoe(2000)], Cauchy-Schwarz, Jeffrey divergence (see [Evren(2012)]), Chernoff (See [Evren(2012)]) , Jensen-Shannon (See [Evren(2012)]). According to [Cichocki and Amari(2010)], there is more than a dozen of different divergence measures in the literature. In a longer version of this paper (see [Ba *et al.*(2017)]), some important applications of them are highlighted with there references. The reader, who is interested by a so important review topic is referred to that paper.

In the next subsection, we describe the frame in which we place the estimation problems we deal in this paper.

1.2. *Statistical Estimation*

The divergence measures may be applied to two statistical problems among others.

(A) First, it may be used as a fitting problem as described here. Let $X_1, X_2, ...$ a sample from X with an unknown probability distribution $\mathbb{P}_X$ and we want to test the hypothesis that $\mathbb{P}_X$ is equal to a known and fixed probability $\mathbb{P}_0$. Theoretically, we can answer this question by estimating a divergence measure $\mathscr{D}(\mathbb{P}_X, \mathbb{P}_0)$ by a plug-in estimator $\mathscr{D}(\mathbb{P}_X^{(n)}, \mathbb{P}_0)$ where, for each $n \geq 1$, $\mathbb{P}_X$ is replaced by an estimator $\mathbb{P}_X^{(n)}$ of the probability law, which is based on sample X_1, X_2, ..., X_n, to be precised.

From there establishing an asymptotic theory of $\Delta_n = \mathscr{D}(\mathbb{P}_X^{(n)}, \mathbb{P}_0) - \mathscr{D}(\mathbb{P}_X, \mathbb{P}_0)$ is thought to be necessary to conclude.

(B) Next, it may be used as tool of comparing for two distributions. We may have two samples and wonder whether they come from the same probability measure. Here, we also may two different cases.

(B1) In the first, we have two independent samples $X_1, X_2, ...$ and $Y_1, Y_2, ...$ respectively from a random variable X and Y. Here the estimated divergence $\mathscr{D}(\mathbb{P}_X^{(n)}, \mathbb{P}_Y^{(m)})$, where n and m are the sizes of the available samples, is the natural estimator of $\mathscr{D}(\mathbb{P}_X, \mathbb{P}_Y)$ on which depends the statistical test of the hypothesis : $\mathbb{P}_X = \mathbb{P}_Y$.

(B2) But the data may also be paired (X,Y), $(X_1,Y_1), (X_2,Y_2), ...$, that is X_i and Y_i are measurements of the same case $i = 1, 2, ...$ In such a situation, testing the equality of the margins $\mathbb{P}_X = \mathbb{P}_Y$ should be based on an estimator $\mathbb{P}^{(n)}_{(X,Y)}$ of the joint probability law of the couple (X,Y) based on the paired observations (X_i, Y_i), $i = 1, 2, \dots, n$.

We did not encounter the approach (B2) in the literature. In the (B1) approach, almost all the papers used the same sample size, at the exception of [Poczos and Jeff(2011)], for the double-size estimation problem. In our view, the study case should rely on the available data so that using the same sample size may lead to a loss of information. To apply their method, one should take the minimum of the two sizes and then loose information. We suggest to come back to a general case and then study the asymptotic theory of $\mathscr{D}(\mathbb{P}_X^{(n)}, \mathbb{P}_Y^{(m)})$ based on samples $X_1, X_2, .., X_n.$ and $Y_1, Y_2, ..., Y_m$. In this paper, we will systematically use arbitrary samples sizes.

In the context of the situation (B1), there are several papers dealing with the estimation of the divergence measures. As we are concerned in this paper by the weak laws of the estimators, our review on that problematic did return only of a few results. Instead, the literature presented us many kinds of results on almost-sure efficiency of the estimation, with rates of convergences and laws of the iterated logarithm, L^p $(p = 1,2)$ convergence, etc. To be precise, [Dhakher *et al.*(2016)] used recent techniques based on functional empirical process to provide a series of interesting rates of convergence of the estimators in the case of one-sided approach for the class de Renyi, Tsallis, Kullback-Leibler to cite a few. Unfortunately, the authors did not address the problem of integrability, taking for granted that the divergence measures are finite. Although the results should be correct under the boundedness assumption *BD* we described earlier, a new formulation in that frame would be welcome.

The paper of [Krishnamurthy *et al.*(2015)] is exactly what we want to do, except that it is concentrated on the L^2-divergence measure and used the Parzen approach. Instead, we will handle the most general case of ϕ-divergence measure and will use the wavelets probability density estimators.

In the context of the situation (B1), we may cite first the works of [Krishnamurthy *et al.*(2014)] and [Singh and Poczos (2014)]. They both used divergence measures based on probability density functions and concentrated on Renyi-α, Tsallis-α and Kullback-Leibler. **In the description of the results below, the estimated *pfd*'s - f and g - are usually in a periodic Hőlder class of a known smoothness** s**.**

Specifically, [Krishnamurthy *et al.*(2014)] defined Renyi and Tsallis estimators by correcting the plug-in estimator and established that, as long as $\mathscr{D}_{R,\alpha}(f,g) \geq c$ and $\mathscr{D}_{T,\alpha}(f,g) \geq c$, for some constant $c > 0$, then

$$\mathbb{E}\left|\mathscr{D}_{R,\alpha}(f_n,g_n) - \mathscr{D}_{R,\alpha}(f,g)\right| \leq c\left(n^{-1/2} + n^{-\frac{3s}{2s+d}}\right)$$

and

$$\mathbb{E}\left|\mathscr{D}_{T,\alpha}(f_n,g_n) - \mathscr{D}_{T,\alpha}(f,g)\right| \leq c\left(n^{-1/2} + n^{-\frac{3s}{2s+d}}\right),$$

[Poczos and Jeff(2011)] used a $k-$nearest-neighbor approach to prove that if $|\alpha - 1| < k$, $(\alpha \neq 1)$ then

$$\lim_{n,m\to\infty} \mathbb{E}\left[\mathscr{D}_{T,\alpha}(f_n,g_m) - \mathscr{D}_{T,\alpha}(f,g)\right]^2 = 0$$

and

$$\lim_{n,m\to\infty} \mathbb{E}\left(\mathscr{D}_{R,\alpha}(f_n,g_m)\right) = \mathscr{D}_{R,\alpha}(f,g).$$

There has been a recent interest in deriving convergence rates for divergence estimators ([Moon and Hero(2014)], [Krishnamurthy *et al.*(2014)]). The rates are typically derived in terms of smoothness s of the densities :

The estimator of [Liu *et al.*(2012)] converges at rate $n^{-\frac{s}{s+d}}$, achieving the parametric rate when $s > d$.

Similarly, [Sricharan et al.(2012)] showed that when $s > d$ a k-nearest-neighbor style estimator achieves the rate $n^{-2/d}$ (in absolute error) ignoring logarithmic factors. In a follow up work, the authors improved this result to $O(n^{-1/2})$ by using a set of weak estimators, but they required $s > d$ orders of smoothness. One can also see [Singh and Poczos (2014)], [Kallberg and Seleznjev(2012)] for other contributions.

The majority of the aforementioned articles worked with densities in Hőlder classes, whereas our work applies for densities in the Besov classes.

Here, we will focus on divergence measures between absolutely continuous probability laws with respect to the Lebesgue measure. As well, our results applied to the approaches (A) and (B1) defined above. As a sequence, we estimate divergence measures by their plug-in counterparts, meaning that we replace the probability density functions (*pdf*) in the expression of the divergence measure by a nonparametric estimators of the *pdf*'s. From now, we have on our probability space, two independent sequences :

(-) a sequence of independent and identically distributed random variables with common *pdf* $f_{\mathbb{P}_X}$:

$$X_1, X_2, ... \tag{1.8}$$

(-) a sequence of independent and identically distributed random variables with common *pdf* $g_{\mathbb{P}_Y}$:

$$Y_1, Y_2, ... \tag{1.9}$$

To make the notations more simple, we write

$$f = f_{\mathbb{P}_X} \text{ and } g = f_{\mathbb{P}_Y}.$$

We focus on using *pdf*'s estimates provided by the wavelets approach. We will deal on the Parzen approach in a forthcoming study. So, we need to explain the frame in which we are going to express our results.

We also wish to get, first, general laws for an arbitrary functional of the form

$$J(f,g) = \int_D \phi(f(x), g(x))dx, \tag{1.10}$$

where $\phi(x,y)$ is a measurable function of $(x,y) \in \mathbb{R}_+^2$ on which we will make the appropriate conditions. The results on the functional $J(f,g)$, which is also known under the name of ϕ-divergence, will lead to those on the particular cases of the Renyi, Tsallis, and Kullback-Leibler measures.

The exposure of all our results will be given in three a series of three papers. This paper is devoted to the foundations the general asymptotic theory and the exposition of the mains theoretical tools concerning the ϕ-forms. The second paper will deal with important key divergence measures and symmetrized estimators. Finally a third paper will focus on the proofs.

1.3. *Wavelets estimation of pdf's*

To begin with the wavelets theory and its statistical applications, we say that the wavelets setting involves two functions φ and ψ in $L_2(\mathbb{R})$ respectively called *father* and *mother* such that

$$\left\{\varphi(.-k),\ 2^{j/2}\psi(2^j(.)-k),(j,k)\in\mathbb{Z}^2\right\},$$

is a orthonormal basis of $L_2(\mathbb{R})$. We adopt the following notation, for $j\geq 0$, $k\in\mathbb{Z}$:

$$\varphi j,k = 2^{j/2}\varphi(2^j(.)-k) \text{ and } \psi_{j,k} = 2^{j/2}\psi(2^j(.)-k).$$

Thus, any function f in $L_2(\mathbb{R})$ is characterized by its coordinates in the orthonormal basis, in the form

$$f = \sum_{k\in\mathbb{Z}} \alpha_{0,k}\varphi_{0,k} + \sum_{k\in\mathbb{Z}}\sum_{j\geq 1}\beta_{j,k}\psi_{j,k} \tag{1.11}$$

with for $j\geq 0$, $k\in\mathbb{Z}$,

$$\alpha_{0,k} = \int_{\mathbb{R}} f(t)\varphi_{0,k}(t)\,dt \text{ and } \beta_{j,k} = \int_{\mathbb{R}} f(t)\psi_{j,k}(t)\,dt.$$

For an easy introduction to the wavelets theory and to its applications to statistics, see for instance [Hardle *et al.*(1998)], [Daubechies(1992)], [Blatter(1998)], etc. In this paper we only mention the unavoidable elements of this frame.

Based on the orthonormal basis defined below, the following Kernel function is introduced

$$\mathbb{R}^2 \ni (x,y) \mapsto K(x,y) = \sum_{k\in\mathbb{Z}} \varphi(x-k)\varphi(y-k).$$

For any $j\geq 1$ fixed, called a resolution level, we define

$$K_j(x,y) = 2^j K(2^j x, 2^j y)$$

and for measurable function h, we define the operator projection K_j of h onto the space V_j of $L_2(\mathbb{R})$ (spanned by $2^{j/2}\varphi(2^j(.)-k)$), by

$$\mathbb{R}\ni x\mapsto K_j(h)(x) = \int K_j(x,y)h(y)dy.$$

Therefore we can write, for all $x\in\mathbb{R}$,

$$\begin{aligned} K_j(h)(x) &= 2^j \int K(2^j x, 2^j y) h(y) dy \\ &= 2^j \int \sum_k \varphi(2^j x - k)\varphi(2^j y - k) h(y) dy. \end{aligned} \tag{1.12}$$

In the frame of this wavelets theory, for each $n \geq 1$, we fix the resolution level depending on n and denoted by $j = j_n$, and we use the following estimator of the *pdf* f associated to X, based on the sample of size n from X, as defined in (1.8),

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_{j_n}(x, X_i). \tag{1.13}$$

As well, in a two samples problem, we will estimate the *pdf* g associated to Y, based on a sample of size n from Y, as defined in (1.9), by

$$g_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_{j_n}(x, Y_i). \tag{1.14}$$

The aforementioned estimator is known under the name of linear wavelets estimators.

Before we give the main assumptions on the wavelets we are working, we have to define the concept of weak differentiation. Denote by $\mathscr{D}(\mathbb{R})$ the class of functions from $\mathbb{R}$ to $\mathbb{R}$ with compact support and infinitely differentiable. A function $f : \mathbb{R} \to \mathbb{R}$ is weak differentiable if and only if there exists a function $g : \mathbb{R} \to \mathbb{R}$ locally integrable (on compact sets) such that, for any $\phi \in \mathscr{D}(\mathbb{R})$, we have

$$\int f(u)\phi'(u)du = -\int g(u)\phi(u)du.$$

In such a case, g is called the weak derivative function of f and denoted $f^{[1]}$. If the first weak derivative has itself a weak derivative, ans so forth up to the $p-1$-th derivative, we get the p-th derivative function $f^{[p]}$. Now we may expose the four assumptions we require on the wavelets.

Assumption 1. . The wavelets φ and ψ are bounded and have compact support and either (i) the father wavelet φ has weak derivatives up to order T in $L_p(\mathbb{R})$ ($1 \leq p \leq \infty$) or (ii) the mother wavelet ψ associated to φ satisfies $\int x^m \psi(x)dx = 0$ for all $m = 0, \ldots, T$.

and

Assumption 2. $\varphi : \mathbb{R} \to \mathbb{R}$ is of bounded p-variation for some $1 \leq p < \infty$ and vanishes on $(B_1, B_2]^c$ for some $-\infty < B_1 < B_2 < \infty$.

Wavelets generators with compact supports are available in the literature. We may cite those named after Daubechies, Coiflets and Symmlets (See [Hardle *et al.*(1998)]). The cited generators fulfill our two main assumption.

Under **Assumption** 2, the summation over k, in (1.12), is finite since only a number of the terms in the summation are non zeros (see [Giné and Nickl(2009)]).

Assumption 3. There exists a non-negative symmetrical and continuous function $\Phi(t)$ of $t \in \mathbb{R}$ with a compact support $\mathscr{K}$ such that :

$$\forall (x,y) \in \mathbb{R}^2, |K(x,y)| \leq \Phi(x-y).$$

The fourth assumption concerns the resolution level we choose. We set for once an increasing sequence $(j_n)_{n\geq 1}$ such that

Assumption 4. $\lim_{n\to+\infty} n^{-1/4}2^{j_n} = 1$.

By the way, we have as $n \to \infty$, and

$$\sqrt{\frac{j_n 2^{j_n}}{n}} + 2^{-tj_n} \approx \sqrt{\frac{1}{4\log 2}\frac{\log n}{n^{3/4}}} + n^{-t/4} \to 0, \;\; \forall t > 0 \tag{1.15}$$
$$\frac{j_n}{\log\log n} \to \infty \quad \text{and} \quad \sup_{n\geq n_0}(j_{2n} - j_n) = \frac{1}{4}.$$

These conditions allow the use the [Giné and Nickl(2009)]'s results.

We also denote

$$\begin{aligned} a_n &= \|f_n - f\|_\infty, \; b_n = \|g_n - g\|_\infty, \; n \geq 1 \\ c_n &= a_n \vee b_n, \; c_{n,m} = a_n \vee b_m, n \geq 1, \; m \geq 1, \\ c^*_{n,m} &= c_{n,m} \vee c_{m,n}, \; n \geq 1, \; m \geq 1. \end{aligned} \tag{1.16}$$

where $\|h\|_\infty$stands for $\sup_{x\in D(h)} |h(x)|$, and $D(h)$ is the domain of application of h.

In the sequel we suppose the densities f and g belong to the Besov space $\mathscr{B}^t_{\infty,\infty}(\mathbb{R})$. We will say a word of simple conditions under which our **pdf**'s do belong to such spaces.

Suppose that the densities f and g belong to $\mathscr{B}^t_{\infty,\infty}(\mathbb{R})$, that φ satisfies **Assumption** 2, and φ, ψ satisfy **Assumption** 1. Then Theorem 3 [Giné and Nickl(2009)] implies that the rates of convergence a_n, b_n and c_n are of the form

$$O\left(\sqrt{\frac{1}{4\log 2}\frac{\log n}{n^{3/4}}} + n^{-t/4}\right)$$

almost-surely and converge all to zero at this rate (with $0 < t < T$).

In order to establish the asymptotic normality of the divergences estimators, we need this key tool concerning the wavelets empirical process denoted by $\mathbb{G}^w_{n,X}(h)$, where $h \in \mathscr{B}^t_{\infty,\infty}(\mathbb{R})$ and defined as follows by

$$\mathbb{G}_{n,X}^{w}(h)=\sqrt{n}\left(\mathbb{P}_{n,X}^{w}-\mathbb{E}_X\right)(h),$$

where $\mathbb{P}_{n,X}^{w}(h)=\mathbb{P}_{n,X}\left(K_{j_n}(h)\right)=\frac{1}{n}\sum_{i=1}^{n}K_{j_n}(h)(X_i)$ and $\mathbb{E}_X(h)=\int h(x)f(x)dx$ denotes the expectation of the measurable function h with respect to the probability distribution function $\mathbb{P}_X$. The superscript w refers to *wavelets*. We have

$$\mathbb{G}_{n,X}^{w}(h)=\sqrt{n}\int(f_n(x)-f(x))h(x)dx \tag{1.17}$$

since, by Fubini's Theorem,

$$\begin{aligned}\sqrt{n}\left(\mathbb{P}_{n,X}^{w}-\mathbb{E}_X\right)(h)&=\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}K_{j_n}(h)(X_i)-\int f(x)h(x)dx\right)\\&=\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\int K_{j_n}(x,X_i)h(x)dx-\int f(x)h(x)dx\right)\\&=\sqrt{n}\int\left(\frac{1}{n}\sum_{i=1}^{n}K_{j_n}(x,X_i)-f(x)\right)h(x)dx\\&=\sqrt{n}\int(f_n(x)-f(x))h(x)dx.\end{aligned}$$

We are ready to give our results on the functional J introduced in Formula (1.10).

2. RESULTS

2.1. *Main Results*

Here, we present a general asymptotic theory of a class of divergence measures estimators including the Renyi and Tsallis families and the Kullback-Leibler ones.

Actually, we gather them in the ϕ-divergence measure form. We will obtain a general frame from which we will derive a number of corollaries. The assumption (1.6) will be used in the particular cases to ensure the finiteness of the divergence measure as mentioned in the beginning of the article. However, in the general results, the assumption (1.6) is part of the general conditions.

We begin to state a result as a general tool for establishing asymptotic normality and related to the wavelets empirical process, which we will use for establishing the asymptotic normality of divergence measures.

Theorem 2.1. *Given the $(X_n)_{n\geq 1}$, defined in (1.8) such that $f\in\mathscr{B}_{\infty,\infty}^{t}(\mathbb{R})$ and let f_n defined as* (1.13) *and $\mathbb{G}_{n,X}^{w}$ defined as in* (1.17). *Then, under* ***Assumption*** *(1-3) and for any bounded h, defined* *on D, belonging to $\mathscr{B}^t_{\infty,\infty}(\mathbb{R})$, we have*

$$\sigma_{h,n}^{-1}\mathbb{G}^w_{n,X}(h) \rightsquigarrow \mathscr{N}(0,1) \ \ as\ n \to \infty,$$

where we have

$$\sigma^2_{h,n} = \mathbb{E}_X\left(K_{j_n}(h)(X)\right)^2 - \left(\mathbb{E}_X(K_{j_n}(h)(X)\right)^2 \to \mathbb{V}ar(h(X)) \ \ as \ \ n \to \infty.$$

Based on that result which will be proved later, we are going to state all results of the functional J defined in Formula 1.10, regarding its almost-sure and Gaussian asymptotic behavior. Let us begin by some notations. Let us assume that ϕ have continuous second order partial derivatives defined as follows :

$$\phi_1^{(1)}(s,t) = \frac{\partial \phi}{\partial s}(s,t), \ \phi_2^{(1)}(s,t) = \frac{\partial \phi}{\partial t}(s,t)$$

and

$$\phi_1^{(2)}(s,t) = \frac{\partial^2 \phi}{\partial s^2}(s,t), \ \phi_2^{(2)}(s,t) = \frac{\partial^2 \phi}{\partial t^2}(s,t), \ \phi_{1,2}^{(2)}(s,t) = \phi_{2,1}^{(2)}(s,t) = \frac{\partial^2 \phi}{\partial s \partial t}(s,t).$$

Define the functions h_i, $i = 1, \ldots 4$:

$$h_1(x) = \phi_1^{(1)}(f(x), g(x)), \ h_2(x) = \phi_2^{(1)}(f(x), g(x)),$$

$$h_3(x) = \phi_1^{(1)}(g(x), f(x)) \text{ and } h_4(x) = \phi_2^{(1)}(g(x), f(x))$$

Set

$$A_1 = \int_D |h_1(x)|\,dx \ \ \text{and} \ A_2 = \int_D |h_2(x)|\,dx$$

and

$$A_3 = \int_D |h_3(x)|\,dx \ \ \text{and} \ A_4 = \int_D |h_4(x)|\,dx.$$

We require the following general conditions.

C-A. All the constants A_i are finite.

C-h. All the functions h_i used in the theorem below are bounded and lie in a Besov space $\mathscr{B}^t_{\infty\infty}$ for some t such that $t > 1/2$.

C1-ϕ. The following integral

$$\int \left\{ |\phi_1^{(1)}(f(x),g(x))| + |\phi_2^{(1)}(f(x),g(x))| \right\} dx < +\infty.$$

us finite.

C2-ϕ. For any measurable sequences of functions $\delta_n^{(1)}(x)$, $\delta_n^{(2)}(x)$, $\rho_n^{(1)}(x)$, and $\rho_n^{(2)}(x)$ of $x \in D$, uniformly converging to zero, that is

$$\max_{i=1,2,\ j=1,2} \sup \left\{ \left|\delta_n^{(i)}(x)\right| + \left|\rho_n^{(j)}(x)\right| \right\} < +\infty,$$

we have as $n \to \infty$

$$\int_D \phi_1^{(2)} \left(f(x) + \delta_n^{(1)}(x), g(x) \right) dx \to \int_D \phi_1^{(2)}(f(x),g(x))dx, \tag{2.1}$$

$$\int_D \phi_2^{(2)} \left(f(x), g(x) + \delta_n^{(2)}(x) \right) dx \to \int_D \phi_2^{(2)}(f(x),g(x))dx, \tag{2.2}$$

and

$$\int_D \phi_{1,2}^{(2)} \left(f(x) + \rho_n^{(1)}(x), g(x) + \rho_n^{(2)}(x) \right) dx \to \int_D \phi_{1,2}^{(2)}(f(x),g(x))dx. \tag{2.3}$$

Remark 2.1.
(a) To check C-h, we may use criteria based on properties of Besov spaces derived on high order differentiability and on the fact we work on compact sets, as it will be seen in the second part of this paper, or in the Appendix section on [Ba *et al.*(2017)]. These techniques show that our results apply to all the usual distributions.

(b) The conditions in C2-ϕ may be justified by the Dominated Convergence Theorem or the monotone Convergence Theorem or from other limit theorems. We may either express conditions on the general function ϕ under which these results hold true. But here, we choose to state the final results and next, to check them for particular cases, in which we may use convergence theorems.

Based on (1.13) and (1.14), we will use the following estimators

$$J(f_n,g) = \int_D \phi(f_n(x),g(x))dx, \quad J(f,g_n) = \int_D \phi(f(x),g_n(x))dx,$$
$$\text{and} \quad J(f_n,g_n) = \int_D \phi(f_n(x),g_n(x))dx.$$

Here are our main results.

I - Statements of the main results.

The first concerns the almost sure efficiency of the estimators.

Theorem 2.2. *Under the assumptions 1-3, C-A, C-h, C1-ϕ, C2-ϕ and (BD), we have*

$$\limsup_{n\rightarrow+\infty}\frac{|J(f_n,g)-J(f,g)|}{a_n}\leq A_1, a.s \tag{2.4}$$

$$\limsup_{n\rightarrow+\infty}\frac{|J(f,g_n)-J(f,g)|}{b_n}\leq A_2, a.s \tag{2.5}$$

$$\limsup_{(n,m)\rightarrow(+\infty,+\infty)}\left|\frac{J(f_n,g_m)-J(f,g)}{c_{n,m}}\right|\leq A_1+A_2\ \ a.s \tag{2.6}$$

where a_n, b_n and c_n are as in (1.16).

The second concerns the asymptotic normality of the estimators.

Theorem 2.3. *Under the assumptions 1-3, C-A, C-h, C1-ϕ, C2-ϕ and (BD), we have*

$$\sqrt{n}(J(f_n,g)-J(f,g))\rightsquigarrow\mathscr{N}\left(0,\mathbb{V}ar(h_1(X))\right),\ as\ n\rightarrow+\infty \tag{2.7}$$

$$\sqrt{n}(J(f,g_n)-J(f,g))\rightsquigarrow\mathscr{N}\left(0,\mathbb{V}ar(h_2(Y))\right),\ as\ n\rightarrow+\infty \tag{2.8}$$

and as $n\rightarrow+\infty$ and $m\rightarrow+\infty$,

$$\left(\frac{nm}{m\mathbb{V}ar(h_1(X)+n\mathbb{V}ar(h_2(Y))}\right)^{1/2}\left(J(f_n,g_m)-J(f,g)\right)\rightsquigarrow\mathscr{N}\left(0,1\right). \tag{2.9}$$

3. Comments and Announcements

In a second paper, we will give versions of our main results on specific and classical divergence measures. The references below, in general, will not be repeated in the two other papers.

Acknowledgment The fourth (1 & 2 & 3) author acknowledges support from the World Bank Excellence Center (CEA-MITIC) that is continuously funding his research activities from starting 2014.

References

[Ba *et al.*(2017)] Ba, A. D., LO, Lo, G.S and Ba, Diam B. (2017) Divergence Measures Estimation and Its Asymptotic Normality Theory Using Wavelets Empirical Processes. ArXiv:1704.04536

[Dhakher *et al.*(2016)] Dhaker H., Ngom P., Deme E. and Mendy Pierre (2016). Kernel-Type Estimators of Divergence Measures and Its Strong Uniform Consistency. American Journal of Theoretical and Applied Statistics. Vol. 5 (1), pp. 13-22. doi: 10.11648/j.ajtas.20160501.13

[Daubechies(1992)] Daubechies, I.(1992). *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia.

[Topsoe(2000)] Topsoe, F. (2000), Some inequalities for information divergence and related measures of discrimination, *IEEE Transactions on Informations Theory*, vol.46, pp.1602-1609.

[Evren(2012)] Evren, A. (2012). Some Applications of Kullback-Leibler and Jeffreys' Divergences in Multinomial Populations. *Journal of Selcuk University natural and Applied Science*,Vol.1(4), pp 48-58.

[Cichocki and Amari(2010)] Cichocki, A. and Amari, S.(2010). Families of Alpha-Beta-and Gamma-Divergences: Flexible and Robust Measures of Similarities. *Entropy*, Vol.12(6), pp 1532-1568.

[Hall(1987)] Hall,P. (1987). On Kullback-Leibler loss and density estimation. *The Annals of Statistics*, Vol.15(4), pp.1491-1519.

[Kullback and Leibler(1951)] Kullback, S. and Leibler, R.(1951). On information and sufficiency. *The Annals of Mathematical Statistics* Vol.22,(1), pp 79-86.

[Singh and Poczos (2014)] Singh S. and Poczos, B. (2014). Generalized Exponential Concentration Inequality for Rényi Divergence Estimation. *Journal of Machine Learning Research*.Vol.6. Carnegie Mellon University.

[Krishnamurthy *et al.*(2014)] Akshay K., Kirthevasan K., Poczos B., and Wasserman, L.(2014). Nonparametric Estimation of Rényi Divergence and Friends. *Journal of Machine Learning Research* Workshop and conference Proceedings, 32. Vol.3, pp. 2.

[Krishnamurthy *et al.*(2015)] KrishnamurthyA., Kandasamy K., Poczós B. and and Wasserman L.(2015) To appear in *Proceedings of the 18th International Con- ference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38*. Copyright 2015 by the authors.

[Moon and Hero(2014)] Moon, K.R. and Hero, III. A.O. , (2014). Ensemble estimation of multivariate f-divergence. in *IEEE Internatonal Symposium on Information Theory*, pp. 356-360.

[Poczos and Jeff(2011)] Poczós, B. and Jeff, S.(2011). On the estimation of $\alpha-$Divergences. In *International Conference on Artificial Intelligence and Statistics,* pp 609-617.

[Liu *et al.*(2012)] Liu, H., Lafferty, J., and Wasserman, L.(2012). Exponential concentration inequality for mutual information estimation . In *Neural Information Processing Systems (NIPS).*

[Giné and Nickl(2009)] Giné, E. and Nickl, R.(2009). Uniform limit theorems for wavelet density estimators. *The Annals of Probability*, Vol.37(4), pp.1605-1646.

[Hardle *et al.*(1998)] Hardle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A.(1998). *Wavelets, Approximation, and Statistical Applications*. Lecture Notes in Statistics.

[Blatter(1998)] Blatter, C. (1998) *Wavelets, a Primer*. A. K. Peters, Natick. MA.

[Valiron(1966)] Valiron, G. (1966). *Théorie des fonctions*. Masson, Paris Milan Melbourne.

[Sricharan et al.(2012)] Sricharan, K., Wei, D., and Hero, A. O. Ensemble estimators for multivariate entropy estimation. *arXiv:1203.5829, 2012*.

[Kallberg and Seleznjev(2012)] Kallberg D. and Seleznjev O. 2012. Estimation of entropy-type integral functionals. *arXiv:1209.2544*.

[Love, (1972)] Love, M.(1972). *Probabily Theory I* 4^{th} Edition. Springer.