International Journal of Mathematical, Engineering and Management Sciences
Vol. 6, No. 3, 688-707, 2021
https://doi.org/10.33889/IJMEMS.2021.6.3.043

On Entropy-type Measures and Divergences with Applications in Engineering, Management and Applied Sciences

C. Koukoumis
Lab of Statistics and Data Analysis,
Department of Statistics and Actuarial-Financial Mathematics,
University of the Aegean, Samos, Greece.
E-mail: sasm19008@sas.aegean.gr

A. Karagrigoriou
Lab of Statistics and Data Analysis,
Department of Statistics and Actuarial-Financial Mathematics,
University of the Aegean, Samos, Greece.
Corresponding author: alex.karagrigoriou@aegean.gr

(Received on February 21, 2021; Accepted on March 12, 2021)

Abstract
In this work we review Entropy-type measures and Divergences, discuss their properties and unfold their diverse applicability. In addition, we compare distances between populations and distributions via weighted Entropy-type measures relying mainly on Relative Entropy and Jeffrey's Distance with weights. Finally, we introduce the Absolute Weighted Relative Entropy and the Absolute Weighted Jeffrey's Distance. Two applications are presented for illustration, one from Geosciences and one from Financial Mathematics.

Keywords- Entropy, Divergence measures, Weighted Entropy-type measures, Absolute Entropy-type measures, Financial mathematics.

1. Introduction

Information theory is a branch of pure and applied sciences that deals with the quantification of information with its roots in modern communication theory where a communication system was formulated as a stochastic process. Tuller (1950) initially and Pierce (1956) later observed the strong similarities between the underlying mechanisms of communication theory and information theory.

The evolution of the field as well as the mathematical rigor that governs it are attributed to Fisher (1956), Shannon (1956) and Wiener (1956). The most fundamental measure in information theory is entropy which was first recognized, formulated and defined in statistical mechanics (Fisher, 1936; Shannon and Weaver, 1949) and consequently triggered the enormous development of the field. In this work we review Entropy-type measures and Divergences, discuss their properties and unfold their diverse applicability. In should be noted that the concept of entropy was used firstly in Physics, in the field of thermodynamics (Clausius, 1865) while its statistical definition was developed by Boltzmann (around 1870) but its applications go beyond Physics.

In the present work we attempt to approach the entropy from a probabilistic or stochastic viewpoint and combine it with the concept of distance which has numerous applications in Applied Sciences, Financial Mathematics, Engineering or Management Sciences. The concept of divergence is fundamental in data analysis since it quantifies the distance between two populations, two models

International Journal of Mathematical, Engineering and Management Sciences
Vol. 6, No. 3, 688-707, 2021
https://doi.org/10.33889/IJMEMS.2021.6.3.043

or two functions. By combining the two concepts and relying on Entropy-type divergences or measures we could provide both researchers and practitioners with useful probabilistic tools for modelling purposes in various scientific areas including goodness of fit in Reliability Theory or Survival Analysis, portfolio selection in Financial Mathematics, decision making in Management Sciences, Geosciences etc.

The sense of entropy has essential role in information, since the middle of the 20th century, when engineers and scientists used the term "information" to quantify something. Claude Shannon (Shannon and Weaver, 1949) with his work "The Mathematical Theory of Communication" was the pioneer of the branch of information theory.

The first scientist who try to quantify the information of a message source with only two numbers was Ralph V. Hartley (1928). In 1948 Shannon provided a generalized form of Hartley's information measure which represents the information (or uncertainty) on average carried by a variable. In that article, Shannon suggests and examines the notions of entropy and mutual information. The entropy is a measure for quantifying the uncertainty of a random variable.

The mutual information measures the mutual dependence between two variables by quantifying the "amount of information" (in bits, nats or bans depending on the base of the logarithm used) which is collected regarding one of the variable by the observation of the other variable. Many scientists after the definition of Shannon entropy, tried to define other types of entropy. One particular generalization is Havrda and Charvát (1967) structural α-entropy; Different values of the parameters result in distinct entropy measures. Shannon entropy is a special case of Havrda–Charvát when this single parameter tends to 1. Tsallis entropy (1988) introduced by Constantino Tsallis is another generalization of Shannon entropy and is similar with Havrda–Charvát structural α-entropy (with a different multiplying factor). Tsallis proposed to replace the usual Shannon with his non-extensive entropy and maximize it. Tsallis entropy has plenty of applications in astrophysics, fractal random walks, time series analysis and classification. Tsallis Relative entropy (1998) introduced also by Tsallis is a generalization of Kullback Cross-entropy which is one of the simplest measures for distance (see below).

The Rényi (1961) entropy that generalizes the Shannon entropy involves a single parameter called order that modifies the Rényi entropy. Also, Rényi entropy has a straightforward relation with Tsallis entropy. However, the axiomatic characterizations are not so simple as Tsallis entropy. Rényi entropy has a variety of applications in many applied fields such as Information theory, Time series, Classification and Cryptography. The relation between Information theory and Statistics was proposed by Kullback and Leibler (1951) who extended the notion of entropy by Shannon and created the Kullback-Liebler measure of divergence also known as "Relative Entropy". Their book "Information Theory and Statistics" was the beginning of a new mathematical field called Statistical Information Theory.

Before Kullback and Leibler, scientists such as Mahalanobis (1936) and later Bhattacharyya (1943) proposed various types of divergences but the work of Kullback and Leibler made the divergences mainstream to the scientific community. Divergence measures have various applications in many scientific fields such as Applied Mathematics, Probability theory, Statistics and Financial Mathematics. With the notion of divergence measures we established the "distance" between samples or two distributions but, divergence measures are not metrics with the mathematical sense of metric because they are not symmetric and most of them do not fulfilled the triangular inequality.

International Journal of Mathematical, Engineering and Management Sciences
Vol. 6, No. 3, 688-707, 2021
https://doi.org/10.33889/IJMEMS.2021.6.3.043

At this point Jeffreys (1946) with his work "An invariant form for the prior probability in estimation problems" proposed the Jeffrey's Distance which is the symmetric version of Relative entropy.

In statistical conjectures on Entropy-type measures, Divergence measures play significant role. In the field of Model Selection, Akaike (1973) was the first scientist who proposed the well-known Akaike Information Criterion (AIC) by constructing an unbiased estimator of the expected Relative entropy. It should be also noted that the use of Relative entropy is a very useful tool in clustering. Yang et al. (2019) used the hierarchical clustering analysis method based on Relative entropy and their application was held on geochemical exploration data. They observed that the Relative entropy can describe the dissimilarity of pairwise geochemical datasets.

Mager et al. (2004) used the Relative entropy as clustering technique to measure the Power spectral analysis of beat-to-beat heart rate variability (HRV). The goodness of fit tests are important tools for examining whether a dataset is compatible with a theoretical probability distribution or whether two datasets share or not the same distribution. The relative entropy goodness of fit test was proposed by Song (2002).

In this work we extend the classical Entropy-type measures to the weighted ones. The Weighted Entropy-type measures play a very significant role in many scientific fields as we mentioned above. If we wish to focus on a specific characteristic of two populations more than others then, we have to give different weights on different parts of the support of the distribution. This desire drives the scientists to re-build the original Shannon entropy to Weighted one. Guiaşu (1971) was the first who proposed the weighted entropy and established the properties for this new type of entropy. The remaining sections of the paper are organized as follows. In Section 2 we review the basic definitions of Divergence measures, different entropies and Entropy-type measures. In Section 3 we introduce the Weighted Entropy-type and Absolute Weighted Entropy-type measures. By using these types of measure, one can focus on "specific" parts of a distribution or population. In Sections 4, 5 and 6 we present some of the areas of application of Entropy-type measures. For the applications we have chosen to present an example on Geosciences and a second one on Financial Mathematics.

2. Entropy-type Measures and Divergence Measures

The sense of distance plays a very important role in probability theory and mathematical statistics, engineering, applied sciences etc. The concept of measuring distances provides various significant results for populations that we wish to study. Usually, we try to measure a characteristic of a population and a reference point to export some useful results. Now, we will present the definition of some of the most popular distance/divergence measures and provide some important identities. Among other we present the definitions for a metric and a divergence, the Shannon entropy and some generalizations and extensions of entropy such as Rényi and Tsallis entropies.

Definition 1. (Metric) A Metric on a set X is a function $d(\cdot,\cdot) : X \times X \to \mathbb{R}_+ \cup \{0\}$ that satisfies the following conditions:

(i) $d(x,y) = 0 \ iff \ x = y \ \forall \ x,y \in X$.
(ii) $d(x,y) = d(y,x) \ \forall \ x,y \in X$.
(iii) $d(x,y) \leq d(x,z) + d(z,y) \ \forall \ x,y,z \in X$.

International Journal of Mathematical, Engineering and Management Sciences
Vol. 6, No. 3, 688-707, 2021
https://doi.org/10.33889/IJMEMS.2021.6.3.043

Definition 2. (Divergence) Suppose S is a space of all probability distributions with same support. Then a Divergence on S is a function $D(\cdot,\cdot) : S \times S \rightarrow \mathbb{R}_+ \cup \{0\}$ satisfying:

$$D(P,Q) = 0 \text{ iff } P = Q \ \forall P, Q \in S.$$

It is important to observe the distinction between divergences and metrics. Indeed, the Divergence measures are not necessarily metrics because they do not have to be symmetric or fulfil the triangular inequality. A classic example of a non-metric is the Relative Entropy also known as Kullback-Leibler divergence (see Definitions 9 and 10 below) which is not symmetric but its role especially in statistical inference, is fundamental (e.g. maximum likelihood estimation).

The notion of Entropy in Information Theory plays an essential role in measuring information or the lack of it (uncertainty). The entropy is related to the uncertainty through the probability of occurrence of the event of interest. Shannon (1948) connected the two concepts for measuring or quantifying the information from a discrete stochastic signal. The following definition is self-explanatory.

Definition 3. (Shannon Entropy) Let a stochastic source described by a discrete random variable X with distribution P_X, support S_X and probability mass function p_X. The Entropy of X is defined by:

$$H(X) = E\left[\log \frac{1}{P_X(X)}\right] = \sum_{x \in S_x} p_X(X) \log \frac{1}{P_X(X)}.$$

Three are the standard identities of Shannon Entropy, namely:

(i) It is always positive.
(ii) It is zero if and only if X describes a certain event.
(iii) It increases by adding an independent component and decreases by conditioning.

Definition 4. (Cross Entropy) Consider two distributions P, Q with probability mass functions $p = (p_1, .., p_n)^T$ and $q = (q_1, \dots, q_n)^T$ respectively. Then the discrete version of Cross Entropy is the following:

$$H(P,Q) = -\sum_{i=1}^{n} p_i \log q_i.$$

Definition 5. (Tsallis Entropy) For any positive real number a the Tsallis Entropy (1988) of order α of a probability measure p on a finite set X is defined as

$$H_a(p) = \begin{cases} \frac{1}{a-1}\left(1 - \sum_{i \in X} p_i^a\right), & \text{if } a \neq 1 \\ -\sum_{i \in X} p_i \log pi\,, & \text{if } a = 1 \end{cases}.$$

Note that such type of entropies, had been studied by Havrda and Charvát (1967) long before

International Journal of Mathematical, Engineering and Management Sciences
Vol. 6, No. 3, 688-707, 2021
https://doi.org/10.33889/IJMEMS.2021.6.3.043

Tsallis. The characterization of the Tsallis entropy is the same as that of Shannon entropy except that for the Tsallis Entropy, the degree of homogeneity under convex linearity condition is α instead of 1.

Definition 6. (The Tsallis Relative Entropy) Tsallis (1998) introduced a generalization of Cross Entropy called the Tsallis Relative Entropy which is given by:

$$T_q(P,P_0)=\int_{\Sigma} P(x)\frac{\left[\frac{P(x)}{P_0(x)}\right]^{q-1}-1}{q-1}dx$$

where $\Sigma \in (-\infty,\infty)$ is the support of the random variable, $P(x)$ is a probability distribution and $P_0(x)$ is a reference distribution.

Definition 7. (The Rényi Entropy) The Rényi Entropy is defined by

$$H_a(p)=\frac{1}{1-a}\log\int_0^{\infty}p^a(x)\,dx$$

where a is a positive constant and $p(x)$ is the probability density function.

Definition 8. (The Havrda–Charvát Entropy) The Havrda–Charvát Entropy (1967) is defined by

$$S(\mu_1,..,\mu_N;a)=\frac{2^{a-1}}{2^{a-1}-1}\left(1-\sum_{i=1}^{N}\mu_i^a\right)\text{ for } a>0,\qquad a\neq 1$$

$$S(\mu_1,\dots,\mu_N;1)=-\sum_{i=1}^{N}\mu_i\log\mu_i$$

where $(\mu_1,\dots,\mu_N)$ is a probability vector.

The main focus of this work is on Relative entropy (Kullback and Leibler, 1951) which is an extremely useful tool in many scientific fields including engineering and applied sciences. Its definition is given below for both the discrete and the continuous case.

Definition 9. (Relative Entropy discrete case) Consider two distributions P,Q with probability mass functions $p=(p_1,..,p_n)^T$ and $q=(q_1,\dots,q_n)^T$ respectively. Then the discrete version of Relative Entropy is the following:

$$D(P,Q)=\sum_{i=1}^{n}p_i\log\left(\frac{p_i}{q_i}\right)=\sum_{i=1}^{n}p_i\log(p_i)-\sum_{i=1}^{n}p_i\log(q_i)$$

International Journal of Mathematical, Engineering and Management Sciences
Vol. 6, No. 3, 688-707, 2021
https://doi.org/10.33889/IJMEMS.2021.6.3.043

Definition 10. (Relative Entropy continuous case) Consider two distributions P, Q with probability mass functions $p = (p_1, .., p_n)^T$ and $q = (q_1, \dots, q_n)^T$ respectively. Then the continuous version of Relative Entropy is the following:

$$D(P,Q) = \int_{-\infty}^{+\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx.$$

As it is easily seen, the Relative entropy does not fulfill the property of symmetricity. For this case, Jeffrey's Distance (Jeffreys, 1946) which is symmetric and closely associated to the Relative entropy, is preferred.

Definition 11. (Jeffrey's Distance) Consider two distributions P, Q with probability mass functions $p = (p_1, .., p_n)^T$ and $q = (q_1, \dots, q_n)^T$ respectively. Then the discrete version of Jeffrey's Distance is the following:

$$D_J(P,Q) = D(P,Q) + D(Q,P) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right) + \sum_{i=1}^{n} q_i \log\left(\frac{q_i}{p_i}\right).$$

Proposition 1. The Relative Entropy $D(P,Q)$ between two distributions P, Q is non-negative on average.

Proof. This proof shows the connection between Relative entropy, Cross-entropy & Shannon entropy and also provides a proof that the Relative entropy is non-negative. We take $I = -\sum_{i=1}^{n} p_i \log p_i$ and introduce appropriate weights $w = (w_1, \dots, w_n)^T$:

$$I_w = -\sum_{i=1}^{n} w_i p_i \log p_i.$$

Let $q_1, \dots, q_k$ the objective probabilities associated with the $p_1, \dots, p_k$ (i.e. p_i is the estimate of the theoretical q_i).

We take, $w_i = \frac{q_i}{p_i}$. Then, $I_w = -\sum_{i=1}^{n} q_i \log p_i$ which coincides with the Cross-entropy. Observe that by Jensen's inequality (Jensen, 1906) we have

$$-D(Q,P) = -\sum_{i=1}^{n} q_i \log\left(\frac{q_i}{p_i}\right) = \sum_{i=1}^{n} q_i \log\left(\frac{p_i}{q_i}\right) \leq \log\left(\sum_{i=1}^{n} q_i \frac{p_i}{q_i}\right) = \log \sum_{i=1}^{n} p_i = \log 1 = 0$$

$$-D(Q,P) \leq 0 \Rightarrow -\sum_{i=1}^{n} q_i \log\left(\frac{q_i}{p_i}\right) \leq 0 \Rightarrow -\sum_{i=1}^{n} q_i \log q_i \leq -\sum_{i=1}^{n} q_i \log p_i.$$

Thus $-\sum_{i=1}^{n} q_i \log\left(\frac{q_i}{p_i}\right) \leq 0$, which implies that

International Journal of Mathematical, Engineering and Management Sciences
Vol. 6, No. 3, 688-707, 2021
https://doi.org/10.33889/IJMEMS.2021.6.3.043

$$-\sum_{i=1}^{n} q_i \log q_i \leq -\sum_{i=1}^{n} q_i \log p_i$$

and finally

Relative entropy = $\sum_{i=1}^{n} q_i \log\left(\frac{q_i}{p_i}\right) = E_q\left[\log\left(\frac{q_i}{p_i}\right)\right]$= *Cross entropy – Shannon entropy* ≥ *0.*

i.e.

subjective-objective measure of uncertainty > measure of objective uncertainty.

This is due to the fact that the uncertainty of the objective probabilities q_i is increased as a result of the uncertainty associated with the estimators of the q_i's by the p_i's. If $p_i = q_i$ then we have equality.

Similarly, for $E_p\left[\log\left(\frac{p_i}{q_i}\right)\right]$ where p are assumed to be the theoretical probabilities and $q_i$is the estimate of p_i. Note that this is a totally different setting since the uncertainty of the true is not the same as before and the same goes for the Cross entropy $-\sum_{i=1}^{n} p_i \log q_i$.

For the non-negativity of Jeffrey's Distance, the proof is an immediate consequence of the previous proposition.

3. Weighted Entropy-type Measures

Sometimes the entropy is not useful enough. For example, if we want to focus on a specific characteristic of two populations more than others then, we have to give different weights on different parts. This desire drives the scientist to re-build the original Shannon Entropy to a weighted one. Guiaşu (1971) was the first who proposed Weighted Entropy. In addition, in this section we provide the definition of the Weighted Relative Entropy.

Definition 12. (Weighted Shannon Entropy) Let a stochastic source described by a discrete random variable X of n possible events, with distribution P_X, probability mass function $p = (p_1, \dots, p_n)^T$ and $w = (w_1, \dots, w_n)^T$ be a vector of weights associated with these states, where $w_i \geq 0$, $i = 1, \dots n$. The weighted Shannon Entropy measure is defined by:

$$H^w(X) = \sum_{i=1}^{n} w_i p_i \log\left(\frac{1}{p_i}\right).$$

Some of the standard properties of the weighted Shannon entropy (for details see Guiaşu, 1971) are:

(i) $H^w(X) \geq 0$.
(ii) If $w_1 = I = w_n = w$, then $H^w(X) = wH^w(X)$.
(iii) If $p_i = 1$ for some $i = 1, \dots, n$ then $H^w(X) = 0$ irrespectively of the values of the weights w.
(iv) If $p_i = 0, w_i \neq 0\ \forall i \in I$ and $p_j \neq 0$, $w_i = 0\ \forall j \in J$ where $I \cup J = \{1,2, \dots, n\}$, $I \cap J = \emptyset$, then $H^w(X) = 0$.

International Journal of Mathematical, Engineering and Management Sciences
Vol. 6, No. 3, 688-707, 2021
https://doi.org/10.33889/IJMEMS.2021.6.3.043

(v) $H^w(w_1, \dots, w_{n+1}; p_1, \dots, p_n, 0) = H^w(w_1, \dots, w_n; p_1, \dots, p_n) = H^w(X)$, for any w_{n+1}.
(vi) For every non-negative, real number λ we have $H^w(\lambda w; p) = \lambda H^w(w, p) = \lambda H^w(X)$.

Suppose that E, F are two incompatible events of the experiment. We require that the weight of the union of these events is equal to the mean value of the weights of the respective events, i.e.

$$w(E \cup F) = \frac{p(E)w(E) + p(F)w(F)}{p(E) + p(F)} \tag{1}$$

where, $w(F)$ is the weight of the event F and $p(F)$ the probability of the same event. In addition, if E, F are complementary events, then

$$w(E \cup F) = p(E)w(E) + (1 - p(E))w(F).$$

If the rule (1) for the weights holds and $w_n = \frac{p'w' + p''w''}{p' + p''}$, $p_n = p' + p''$, then:

$$H^w(w_1, \dots, w_n, w', w''; p_1, \dots, p_{n-1}, p', p'') = H^w(w_1, \dots, w_n; p_1, \dots, p_n) + p_n H^w\left(w', w''; \frac{p'}{p_n}, \frac{p''}{p_n}\right).$$

Definition 13. (Weighted Relative Entropy) Consider two probability mass functions $p = (p_1, \dots, p_n)^T, q = (q_1, \dots, q_n)^T$ and $w = (w_1, \dots, w_n)^T$ be a vector of weights. Then the discrete version of Weighted Relative Entropy would be the following:

$$D^w(p, q) = \sum_{i=1}^{n} w_i p_i \log\left(\frac{p_i}{q_i}\right).$$

This form of Relative Entropy is not a proper distance measure, because it could take negative values. This disadvantage of the Weighted Relative Entropy method forced us to propose the Absolute Weighted Relative Entropy given in the definition below:

Definition 14. (Absolute Weighted Relative Entropy (A.W.R.E)) Consider two distributions P, Q with probability mass functions $p = (p_1, \dots, p_n)^T, q = (q_1, \dots, q_n)^T$ and $w = (w_1, \dots, w_n)^T$ be a vector of weights. Then the discrete version of Absolute Weighted Relative Entropy would be the following:

$$D^{wabs}(p, q) = \sum_{i=1}^{n} \left| w_i p_i \log\left(\frac{p_i}{q_i}\right) \right|.$$

Through the modification of the above definition the resulting Absolute Weighted Relative Entropy (A.W.R.E) is always non-negative. Using as a proper distance tool the A.W.R.E we could proceed and give more "attention" to a special part of a distribution. In order to incorporate the symmetry property into the previous definition we propose below the Absolute Weighted Jeffrey's Distance which is both non-negative and fulfills the symmetric property:

International Journal of Mathematical, Engineering and Management Sciences
Vol. 6, No. 3, 688-707, 2021
https://doi.org/10.33889/IJMEMS.2021.6.3.043

Definition 15. (Absolute Weighted Jeffrey's Distance (A.W.J.D)) Consider two distributions P, Q with probability mass functions $p = (p_1, .., p_n)^T$ and $q = (q_1, \dots, q_n)^T$ respectively and $w = (w_1, \dots, w_n)^T$ a vector of weights. Then the discrete version of Absolute Weighted Jeffrey's Distance is the following:

$$D_J^{wabs}(P, Q) = \sum_{i=1}^{n} \left| w_i p_i \log\left(\frac{p_i}{q_i}\right) \right| + \sum_{i=1}^{n} \left| w_i q_i \log\left(\frac{q_i}{p_i}\right) \right|.$$

The use of weights and absolute values allow the research to focus exclusively on specific parts of the distribution and the corresponding Relative entropy will be non-negative. In the three sections that follow we present three scientific areas where the Entropy-type measures find some of their numerous applications.

4. Clustering based on Entropy-type Measures

The problem of clustering is related to grouping a set of objects in the same group or classes within each of which the objects are similar (homogeneous). Frequently, we wish to quantify the dissimilarity between two populations. The clustering is a classical method for distributing populations into clusters. The greater the number of populations, the greater the number of clusters. There are many ways to measure the dissimilarity between two clusters with one of them being the Relative entropy defined previously. Note that the Relative entropy technique is quite similar with the Mahalanobis distance (1936). The main difference being that through the Relative entropy we can represent the value in terms of difference between the two clusters. With the evaluation of Relative entropy we can statistically show whether two clusters are similar or not making the Relative entropy a very useful key in clustering.

Yang et al. (2019) used the hierarchical clustering analysis method based on Relative entropy and their application was held on geochemical exploration data. They observed that the Relative entropy can describe the dissimilarity of pairwise geochemical datasets. Mager et al. (2004) used the Relative entropy as clustering technique to measure the power spectral analysis of beat-to-beat heart rate variability (HRV). The research concerned with the developing of an algorithm that utilizes continuous wavelet transform (CWT) parameters as inputs to a Kohonen self-organizing map (Kohonen, 1990), providing a method of clustering.

All the above, clearly show that the Relative entropy is a powerful and useful tool for the comparison between populations for clustering purposes.

5. Goodness of Fit based on Entropy-type Measures

The goodness of fit tests are important tools for testing whether a dataset is compatible with a theoretical probability distribution or whether two datasets share or not the same distribution. The Relative entropy goodness of fit test was proposed by Song (2002). The relation of goodness of fit test and the Relative entropy will be shown below. Assume the test hypothesis:

$$H_0: q = p \text{ vs } H_1: q \neq p.$$

The previous test hypothesis about the possible equality between two densities p, q is equivalent to the following test based on the measure $D(\cdot,\cdot)$ defined in the previous section.

International Journal of Mathematical, Engineering and Management Sciences
Vol. 6, No. 3, 688-707, 2021
https://doi.org/10.33889/IJMEMS.2021.6.3.043

$$H_0: D(Q,P) = 0 \text{ vs } H_1: D(Q,P) > 0.$$

Let A a category, O_i the frequency of results belongs to A_i and $e_i = E(O_i),\ \ i = 1, \dots, k$ then the maximum likelihood (ML) estimator of q_i is $\widehat{q_i} = \frac{O_i}{n}$. And $p_i = \frac{e_i}{n}$, The ML-estimator of $D(Q,P)$ is:

$$\widehat{D}(Q,P) = \frac{1}{n}\sum_{i=1}^{k} O_i \log\frac{O_i}{e_i}.$$

The vector $O = (O_1, \dots, O_k) \sim M_k(n, (q_1, \dots, q_k))$, where M_k is k-dimnsional multinomial distribution. For the very big sample size the vector O an asymptotic multivariate normal distribution $N_k(nq, n(D_p - qq'))D_q$ is a diagonal matrix with diagonal elements $q_i\ \ i = 1, \dots, k$ and $q = (q_1, \dots, q_k)$. Thus

$$\sqrt{n}\left(\frac{1}{n}O - q\right) \to N_k\left(0, D_q - qq'\right).$$

Simple algebra shows that:

$$Z = \sqrt{n}\left(\frac{\widehat{D}(Q,P) - D(Q,P)}{\hat{\sigma}}\right) \to N(0,1)$$

where

$$\hat{\sigma}^2 = \frac{1}{n}\left[\sum_i O_i \left(\log\frac{O_i}{e_i}\right)^2 - \left(\sum_i O_i \log\frac{O_i}{e_i}\right)^2\right].$$

Then, in testing hypothesis $H_0: D(Q,P) = 0$ in favor of $H_1: D(Q,P) > 0$ we reject H_0 if $Z_0 > z_a$ where,

$$Z_0 = \frac{\sqrt{n} \cdot \widehat{D}(Q,P)}{\hat{\sigma}}$$

and z_a is the 1- α quantile of the standard normal distribution.

The previous result is very close to G^2 which is the well-known likelihood ratio test statistic (Neyman and Pearson 1933) but in simulations (Sharifdoost et al., 2009) it appears to be more sensitive than G^2. Goodness of fit tests based on the Relative entropy are more sensitive than usual methods for rejecting distributions which are close to the distribution we have (Sharifdoost et al., 2009).

6. Model Selection based on Entropy-type Measures

Model selection is the field of statistics for selecting an ideal statistical model from a set of candidate models. As we know, model selection plays important role in any scientific field. The first scientist who studied deeply this concept was Akaike (1973) who proposed the Akaike Information Criterion (AIC) by constructing an unbiased estimator of the expected Relative entropy. Below we briefly discuss the main characteristics of AIC as well as the Divergence Information Criterion (DIC).

International Journal of Mathematical, Engineering and Management Sciences
Vol. 6, No. 3, 688-707, 2021
https://doi.org/10.33889/IJMEMS.2021.6.3.043

Let f be the true model and g a model which is used to estimate f. The Relative entropy (equivalent to definition 10) between f and g is:

$$D(f,g) = \int_X f(x) \log \frac{f(x)}{g(x|\theta)} dx$$

where θ a parameter associated with g for the estimation of which one uses the available data. $D(f,g)$ with support X, represents the information lost when g is used to estimate f. Equivalently we can write:

$$D(f,g) = \int_X f(x) \log f(x) dx - \int_X f(x) \log\big(g(x|\theta)\big) dx$$

$$= E_f[\log f(x)] - E_f[\log(g(x|\theta))].$$

The first expectation is constant, say z, irrespectively of the model g used, so

$$D(f,g) = z - E_f\big[\log\big(g(x|\theta)\big)\big] \Rightarrow$$

$$D(f,g) - z = -E_f\big[\log\big(g(x|\theta)\big)\big].$$

By computing $E_f[\log(g(x|\theta))]$ which is the continuous version of the Cross entropy (see definition 4), we easily obtain the relative distance $D(f,g) - z$ distance between f and g. Instead of this quantity which cannot be computed Akaike found that the expectation $E_f\left[E_f[\log(g(x|\theta))]\right]$ can be computed. For the above quantity, which is known as the *expected Relative entropy information*, the asymptotically unbiased estimator is found by Akaike to be $\log(\mathcal{L}(\hat{\theta}|x)) - p$ where p is the dimension of parameter θ and $\hat{\theta}$ is a consistent estimate of θ. Then the AIC is:

$$AIC = -2 \log\left(\mathcal{L}\left(\hat{\theta}\middle|x\right)\right) + 2p$$

where $\hat{\theta}$ is the maximum likelihood estimator (or equivalently the minimum Relative entropy estimator). Selecting among various candidate models g, the model with the smallest AIC value is related to the model with the least Relative entropy between the true distribution f and the estimated one.

Now, we present another useful information criterion for model selection called Divergence Information Criterion (DIC, Mattheou et al., 2009). For this type of criterion Mattheou et al. (2009) based on the same methodology as AIC criterion used the BHHJ Divergence (Basu et al., 1998) for developing a new criterion.

International Journal of Mathematical, Engineering and Management Sciences
Vol. 6, No. 3, 688-707, 2021
https://doi.org/10.33889/IJMEMS.2021.6.3.043

Consider a random sample $X_1, \ldots, X_n$ from the distribution f (the true model) and a candidate model g_θ. For constructing the DIC the following formula will be useful.

$$W_\theta = E_{g_\theta}(g_\theta^a(Z)) - (1-\alpha^{-1})E_f\left(g_\theta^a(Z)\right), \quad a > 0$$

which is the same as the BHHJ divergence without the last term, which remaining constant independent of the model g_θ. Now the formula which gives an unbiased estimator is $EW_\theta = E\left(W_\theta \middle| \theta = \hat{\theta}\right)$ where $\hat{\theta}$ is an asymptotically normal estimator of θ. We can also say that the above expression is the average distance between f and g_θ. Now, we present an unbiased estimator of the expected overall discrepancy:

$$Q_\theta = \int g_\theta^{1+a}(z)dz - \left(1+\frac{1}{a}\right)\frac{1}{n}\sum_{i=1}^{n} g_\theta^a(X_i).$$

The asymptotically unbiased estimator of n-times the expected overall discrepancy evaluated at $\hat{\theta}$ is given by

$$DIC = nQ_{\hat{\theta}} + (\alpha+1)(2\pi)^{-\frac{a}{2}}\left(\frac{1+a}{1+2a}\right)^{1+\frac{p}{2}} p.$$

The adjusted DIC model is given below (Mantalos et al., 2010) for the case of the MLE of θ:

$$DIC_{MLE} = nQ_{\hat{\theta}} + (2\pi)^{-\frac{a}{2}}(1+a)^{-\frac{p}{2}}p\,.$$

We must point out that the MLE method is faster in computations than other methods. Also, the DIC criterion has highly performance of accuracy in simulations. Note also that it could be used in applications with outlier and contaminated observations. Based on all the above observations we conclude that DIC is a powerful criterion for model selection.

7. Applications

In this section we study two examples one from Geosciences focusing on a dataset on earthquakes and a second one on Financial Mathematics focusing on the price comparison of two stocks. The purpose of the analysis is to check the performance and the capabilities of the proposed absolute weighted entropy-type measures as opposed to standard entropy-type measures. For the first example we compare the distribution of the dataset with a specific candidate distribution while the second compares the distributions of logarithmic return prices of two stocks.

The analysis is based on an algorithmic procedure the steps of which are presented below. The method requires the split of the support of the distribution into a number of subintervals for each of which the associated probabilities are evaluated. The method which is called "*Middle method*" uses the probabilities obtained for each of the subintervals of the support. The method is then applied to Entropy-type measures discussed in the previous sections and compare them. For the examples considered in this work we use n=10 subintervals.

International Journal of Mathematical, Engineering and Management Sciences
Vol. 6, No. 3, 688-707, 2021
https://doi.org/10.33889/IJMEMS.2021.6.3.043

The steps of the "*Middle method*" algorithm are the following:

1. Take all the intervals and calculate the Relative entropy (weights: $w_i = 1$ for each interval).

2. Remove the 2 middle intervals and use the remaining by calculating the Relative entropy (weights: $w_i = \frac{n}{n-2}$ for each interval).

3. Repeat by removing 2 middle intervals at each step of the algorithm and increasing accordingly by an equal amount the weights so that $\sum w_i = n$.

4. Repeat the steps 1-3 for (a) Jeffrey's Distance (b) Absolute Relative Entropy and (c) Absolute Jeffrey's Distance.

Remark: In addition to the Middle method that focuses on the analysis of both tails, similar methods can be implemented in case the focus should be placed on a single tail of the distribution. Indeed, the so-called Left method (LR, Left to Right method) removes one-by-one the intervals starting from the left while in the Right method (RL, Right to Left method) the intervals are removed starting from the right. The Relative entropy is calculated in each step of the algorithmic procedure applying each time, the appropriate weights.

7.1 The Geosciences Example

We collected data from the Hellenic Institute of Geodynamics (National Observatory of Athens) www.gein.noa.gr. The data concern 5384 earthquakes from 1973 to 2004 in Greece with magnitude below 4 in Richter scale. In this part of the work, we try to see the relationship between our dataset and the Shifted Exponential Distribution which is a displacement of Exponential Distribution to the right, by 4 units. Firstly, we present the histogram of the data together with the Shifted Exponential Distribution (Figure 1). The average of the data is equal 4.280758 and the standard deviation is 0.3451494. We suppose the distribution that fits better the data is the Shifted Exponential Distribution by 4 units with parameter estimated by $\lambda = 4.280758$.

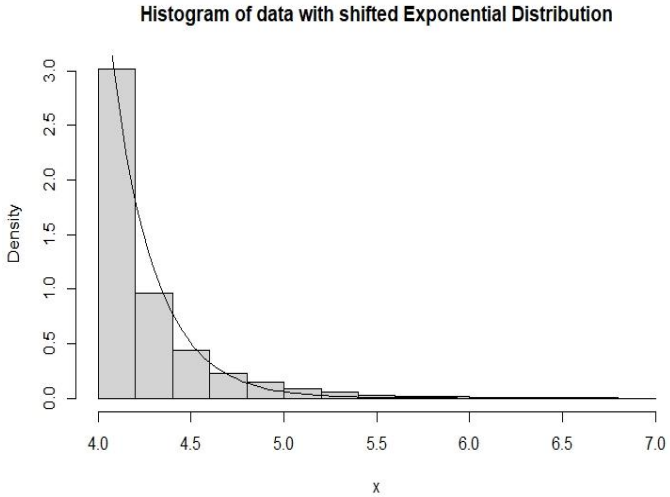

Figure 1. Histogram of dataset and shifted exponential distribution.

International Journal of Mathematical, Engineering and Management Sciences
Vol. 6, No. 3, 688-707, 2021
https://doi.org/10.33889/IJMEMS.2021.6.3.043

Now, for the implementation of the proposed method using the algorithm defined earlier, we divide the support as follows:

$$[4, 4.25), [4.25, 4.5), [4.5, 4.75), [4.75, 5), [5, 5.25),$$

$$[5.25, 5.5), [5.5, 6), [6, 6.25), [6.25, 6.5), [6.5,7]$$

The main idea for splitting the support of the data is to add specific weights on each interval. Then, we calculate the required Entropy-type measures. Table 1 provides the percentage of data in every interval for the real data and the Shifted Exponential Distribution respectively.

Table 1. Percentages by interval dataset vs shifted exponential distribution.

Percentages per Interval										
Intervals	[4, 4.25)	[4.25, 4.5)	[4.5, 4.75)	[4.75, 5)	[5, 5.25)	[5.25, 5.5)	[5.5, 6)	[6, 6.25)	[6.25, 6.5)	[6.5,7]
Data	0.6027	0.2423	0.0624	0.0514	0.0170	0.0126	0.0083	0.0013	0.0005	0.00111
Shifted Exponential	0.6570	0.2253	0.0772	0.0265	0.0090	0.0031	0.0014	0.0001	0.00004	0.00001

Figure 2 describes the "Middle method" and the comparison of 4 Weighted Entropy-type techniques. From Figure 2 we observe that the Relative entropy is not useful. Indeed, firstly, it is not symmetric and secondly it takes negative values. The latter is due to the fact that the numerator of the logarithm is smaller than the denominator and when this happens the measure is negative. The defects stated above can be resolved if one uses Jeffrey's Distance. Observe that Jeffrey's measure is both symmetric and always positive. Note though that Jeffrey's Distance is not very useful because although each term is positive, the elements of each term are not both positive. One is positive and one is negative so that the result (even with the use of a large weight) will not be as extreme as it should. This defect can be resolved if one uses the Absolute Jeffrey's Distance which combines the advantages of Jeffrey's Distance and the absolute value. It should be noted that the use of squares instead of the absolute value, was not going to have the same effect since each term in each summation is less than 1 and the squares were going to reduce the magnitude of the contribution of the most significant intervals (terms). Observe further that the use of Jeffreys together with the absolute value increases when we focus on the last two intervals where the difference is maximum.

International Journal of Mathematical, Engineering and Management Sciences
Vol. 6, No. 3, 688-707, 2021
https://doi.org/10.33889/IJMEMS.2021.6.3.043

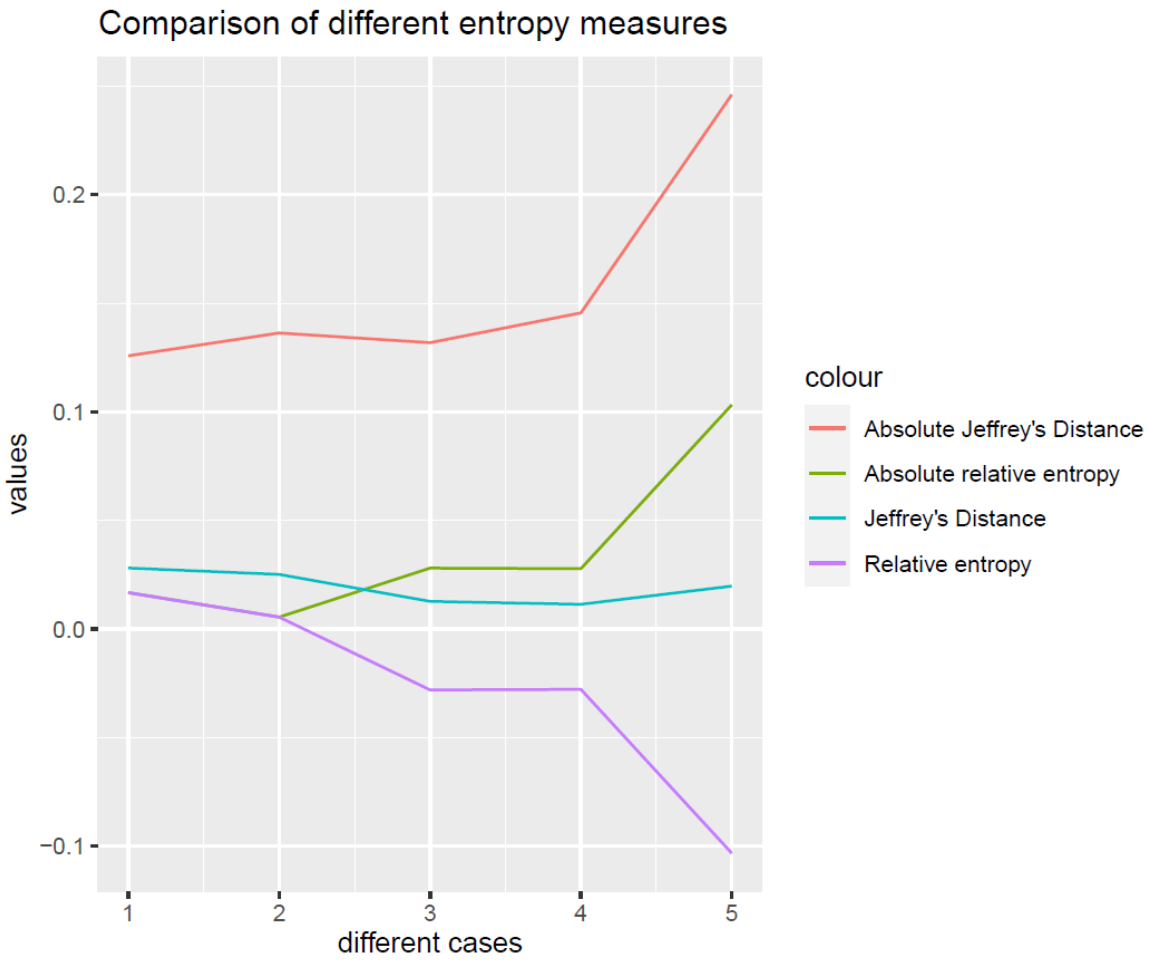

Figure 2. Dataset vs. shifted exponential distribution.

7.2 The Financial Mathematics Example

For the second application, we collect from www.finance.yahoo.com, logarithmic return prices of the index S&P500 and logarithmic return prices of Barrick Gold Corporation (GOLD) for a five-year period from 05/Jan/2016 until 05/Jan/2021. Our purpose is to find the relationship ("distance") between the two stocks via Absolute Weighted and Weighted Entropy-type measures. The dataset contains 1258 observations for each stock. Firstly, we present the histograms of the data separately (Figure 3). Note that although the choice of the above two stocks is purely illustrative for revealing the capabilities of the proposed algorithmic procedure, financial products are often representative examples due to the negative correlation they frequently exhibit as a result of the belief that gold moves higher when economic conditions worsen and stock markets go down.

For S&P500 returns the average is 0.000483 and the standard deviation is 0.012184876 while for Barrick Gold Corporation returns the average is 0.000929 and the standard deviation is 0.025566474.

International Journal of Mathematical, Engineering and Management Sciences
Vol. 6, No. 3, 688-707, 2021
https://doi.org/10.33889/IJMEMS.2021.6.3.043

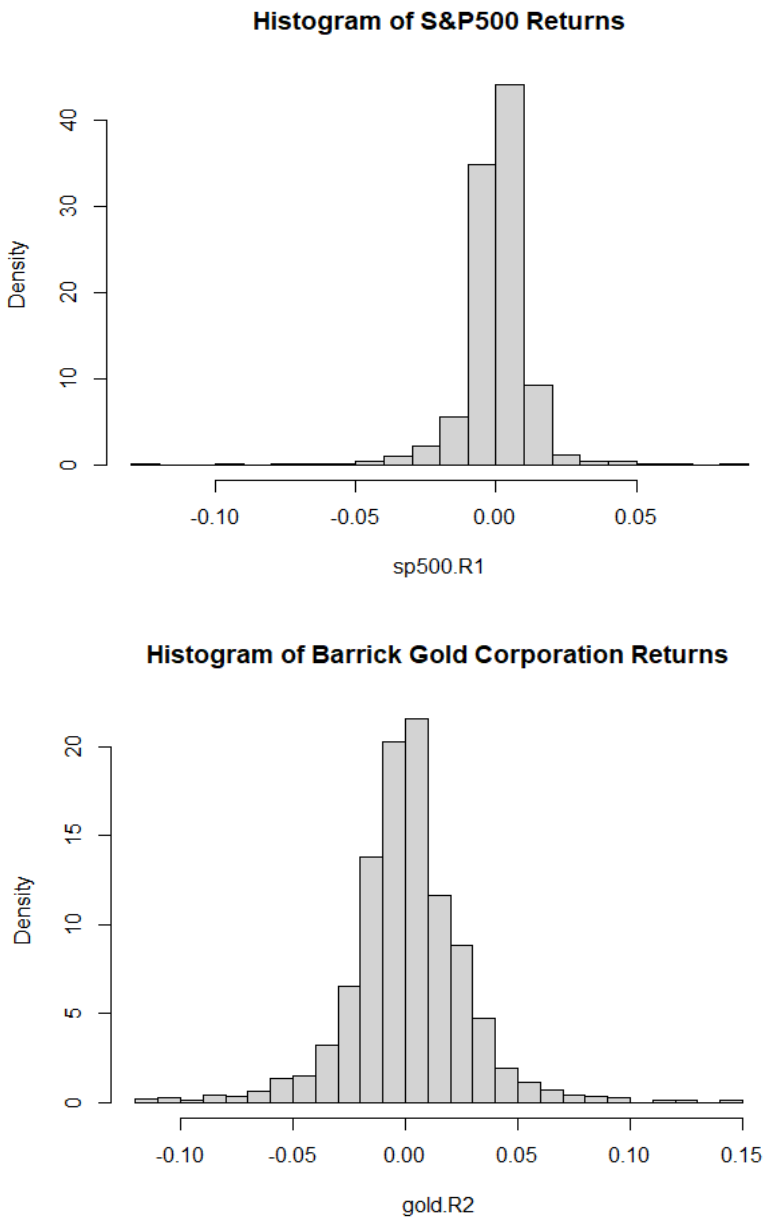

Figure 3. Histograms of S&P500 and GOLD returns.

For the implementation of the method we will divide the support of the dataset in the following way:

$$(-\infty, -0.08), [-0.08, -0.05), [-0.05, -0.02), [-0.02, -0.01),$$

$$[-0.01, 0), [0, 0.01), [0.01, 0.02), [0.02, 0.05), [0.05, 0.08), [0.08, \infty).$$

International Journal of Mathematical, Engineering and Management Sciences
Vol. 6, No. 3, 688-707, 2021
https://doi.org/10.33889/IJMEMS.2021.6.3.043

Table 2 provides the percentage of data in every interval of the dataset.

Table 2. Percentages by interval S&P500 vs GOLD.

Percentages per Interval										
Intervals	$(-\infty,-0.08]$	$[-0.08,-0.05]$	$[-0.05,-0.02]$	$[-0.02,-0.01]$	$[-0.01,0)$	$[0,0.01)$	$[0.01,0.02]$	$[0.02,0.05)$	$[0.05,0.08]$	$[0.08,\infty)$
S&P500	0.0015	0.0031	0.034	0.0548	0.3489	0.441	0.0922	0.0190	0.00238	0.005
GOLD	0.0087	0.0230	0.112	0.1375	0.2027	0.214	0.1160	0.1542	0.02225	0.007

Applying again the "Middle method" algorithm (as in previous example) we reveal the relation ("distance") between the stocks S&P500 and GOLD and compare the relevant Entropy-type measures used (Figure 4).

Observe again that the Relative entropy takes negative values while the Absolute Relative entropy is non-negative but it takes smaller values than Jeffrey's Distance. Further, the largest differences between the two stocks are reported for the Absolute Jeffrey's Distance method. In conclusion, we observe that the Absolute Jeffrey's Distance gives the higher differences of the distance between the two stocks. Observe also that if the Relative entropy is used the two stocks appear to be almost equidistant.

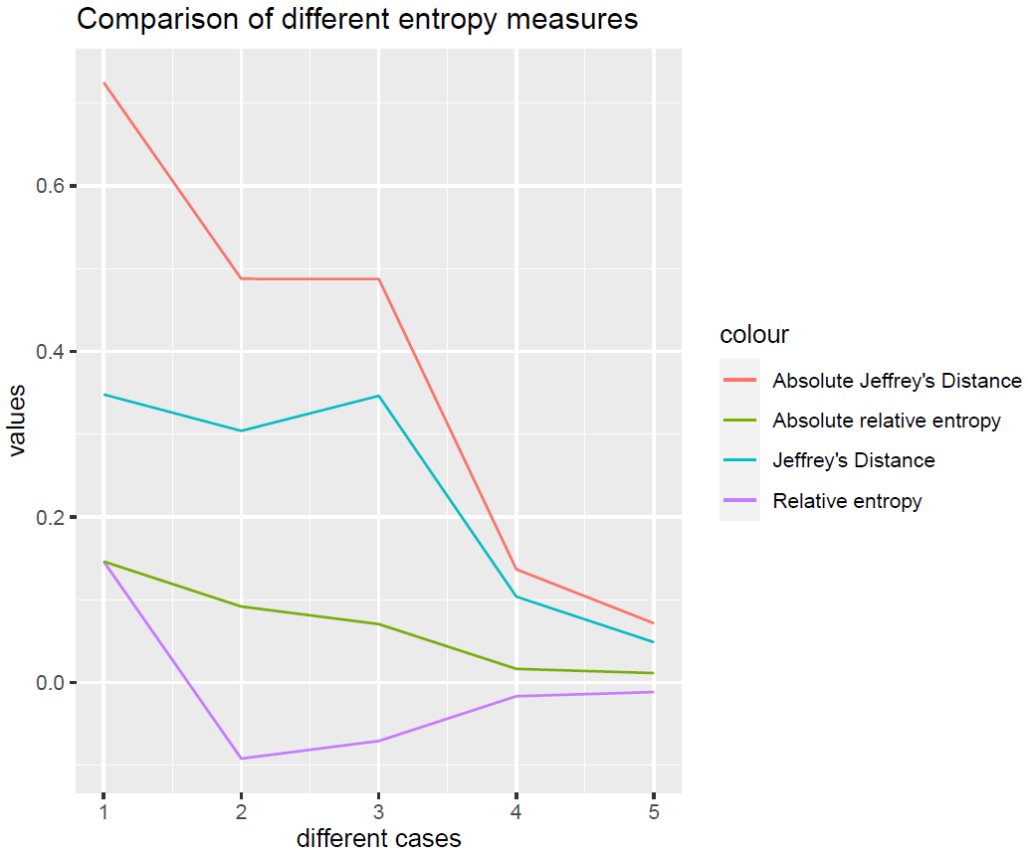

Figure 4. Middle method S&P500 vs GOLD.

8. Conclusions

The main purpose of this work is to review of Entropy-type measures and Divergences, discuss their properties and unfold their diverse applicability. After the presentation of the necessary theory on Divergences and Entropy, we proposed and compared Weighted Entropy-type measures and revealed and explored their advantages as distance measures. More specifically, we first observed that the Weighted Relative Entropy technique is less accurate because it takes negative values which violates the main idea of distance. Then, we presented the Weighted Jeffrey's Distance which is symmetric but not accurate. Finally, we introduced the Absolute Weighted Relative Entropy (A.W.R.E) and the Absolute Weighted Jeffrey's Distance (A.W.J.D) both of which and especially the second one, gave larger distance values between two datasets and at the same time fulfilled the properties of symmetricity and non-negativity.

For checking the performance of the proposed methodology, we applied all previous theoretical results in two applications. The first experiment on Geosciences focuses on the closeness of the distribution of earthquakes and a fitted distribution while the other experiment on Financial Mathematics deals with the measuring of the distance and the relation between two stocks. By introducing the Absolute Weighted Entropy-type methods we observed that the Absolute Jeffrey's Distance provides the best results (higher values) among all methods considered. Although the entropy-type measures are important and useful in many fields, the Absolute Entropy-type measures proposed in this work can be found extremely useful in special cases.

In conclusion, based on the two real applications we conclude that the Absolute Jeffrey's Distance appears to be *the most sensitive* Entropy-type measure among all studied techniques. This means that it produces larger values when we focus on specific parts which otherwise have indistinguishable dissimilarities and therefore it provides the researcher with a useful tool for many scientific fields where the interest focusses not on the entire distribution but on specific (special) parts of it.

Conflict of Interest

The authors confirm that there is no conflict of interest to declare for this publication.

Acknowledgments

The authors wish to express their appreciation to the Editor and two anonymous reviewers for their valuable comments and recommendations that greatly improve the quality of the manuscript. This work is part of the thesis of the first author under the supervision of the second author and was completed as part of the research activities of the Lab of Statistics and Data Analysis (LabSTADA) of the University of the Aegean (http://actuarweb.aegean.gr/labstada/index.html).

References

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika*, *60*(2), 255-265.

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

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, *35*, 99-109.

Clausius, R. (1865). *Ueber verschiedene für die Anwendung bequeme Formen der Hauptgleichungen der mechanischen Wärmetheorie: vorgetragen in der naturforsch. Gesellschaft den 24. April 1865*.

Fisher, R.A. (1936). Statistical methods for research workers (rev. & enl, pp. xiii, 339). *Edinburgh and London: Oliver & Boyd*.

Fisher, R.A. (1956). *Statistical methods and scientific inference.* Hafner Publishing Co.

Guiaşu, S. (1971). Weighted entropy. *Reports on Mathematical Physics*, *2*(3), 165-179.

Havrda, J., & Charvát, F. (1967). Quantification method of classification processes. Concept of structural a-entropy. *Kybernetika*, *3*(1), 30-35.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical, Physical and Engineering Sciences*, *186*(1007), 453-461.

Jensen, J.L.W.V. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, *30*, 175-193.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464-1480.

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

Mager, D.E., Merritt, M.M., Kasturi, J., Witkin, L.R., Urdiqui-Macdonald, M., Sollers 3rd, J.J., & Thayer, J.F. (2004). Kullback-Leibler clustering of continuous wavelet transform measures of heart rate variability. *Biomedical Sciences Instrumentation*, *40*, 337-342.

Mahalanobis, P.C. (1936). On the generalized distance in statistics. *National Institute of Science of India*, *2*, 49-55.

Mantalos, P., Mattheou, K., & Karagrigoriou, A. (2010). An improved divergence information criterion for the determination of the order of an AR process. *Communications in Statistics—Simulation and Computation*, *39*(5), 865-879.

Mattheou, K., Lee, S., & Karagrigoriou, A. (2009). A model selection criterion based on the BHHJ measure of divergence. *Journal of Statistical Planning and Inference*, *139*(2), 228-235.

Neyman, J., & Pearson, E.S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *231*(694-706), 289-337.

Pierce, J.F. (1956). *Raw material inventory control at the Boston Woven Hose and Rubber Company: the information system, control quantities*, (*Doctoral dissertation*, Massachusetts Institute of Technology. School of Industrial Management.)

Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 547-561.

Shannon, C. (1956). The zero error capacity of a noisy channel. *IRE Transactions on Information Theory*, *2*(3), 8-19.

Shannon, C.E. & Weaver, W. (1949). *The mathematical theory of communication*. The University of Illinois Press, Urbana, IL.

Sharifdoost, M., Nematollahi, N., & Pasha, E. (2009). Goodness of fit test and test of independence by entropy. *Journal of Mathematical Extension*, *3*(2), 43-59.

Song, K.S. (2002). Goodness-of-fit tests based on Kullback-Leibler discrimination information. *IEEE Transactions on Information Theory*, *48*(5), 1103-1117.

Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, *52*(1), 479-487.

Tsallis, C. (1998). Generalized entropy-based criterion for consistent testing. *Physical Review E*, *58*(2), 1442.

Tuller, W.G. (1950). Information theory applied to system design. *Transactions of the American Institute of Electrical Engineers*, *69*(2), 1612-1614.

Wiener, N. (1956). The theory of prediction. In Beckenbach, E.F. (ed) *Modern Mathematics for Engineers*, McGraw-Hill, New York, pp.165–190.

Yang, J., Grunsky, E., & Cheng, Q. (2019). A novel hierarchical clustering analysis method based on Kullback–Leibler divergence and application on dalaimiao geochemical exploration data. *Computers & Geosciences*, *123*, 10-19.