Entropy Measures with Applications in Financial Mathematics

Koukoumis Charalampos

Supervisor:
Prof. Karagrigoriou Alexandros
Co-Supervisors:
Assoc. Prof. Tachtsis Eleftherios
Assis. Prof. Vakeroudis Stavros

Department of Statistics and Actuarial-Financial Mathematics
University of the Aegean
Greece

A thesis presented for the degree of
Master of Science (M.Sc.)

Supervisors

Karagrigoriou Alexandros

Professor, Department of Statistics and Actuarial-Financial Mathematics, University of the Aegean, Greece.

Tachtsis Eleftherios

Associate Professor, Department of Statistics and Actuarial-Financial Mathematics, University of the Aegean, Greece.

Vakeroudis Stavros

Assistant Professor, Department of Statistics and Actuarial-Financial Mathematics, University of the Aegean, Greece.

Dedicated to Georgios, Maria, Charalampos, Maria and Aggelita.

Abstract

Many times in statistics and in general in sciences one of the basic goals is the comparison and quantification of the distance between distributions and populations. One very important tool is the Entropy-type measures. By using Entropy-type measures we easily quantify the "distance" between random processes. At the same time, many times, researchers wish to focus on "specific parts" of random processes. A solution to this problem is given by the Weighted Entropy-type measures. The use of this type of Entropy-type measures has many advantages research-wise. Moreover they can be used in plethora of scientific fields like Financial Risk Analysis, Geosciences, Meteorology etc. where researchers often give more attention to the tails of a distribution.

Περίληψη

Πολλές φορές στη στατιστική αλλά και γενικά στις θετικές επιστήμες ένας από τους βασικότερους σκοπούς είναι η σύγκριση και η ποσοτικοποίηση της απόστασης μεταξύ κατανομών και πληθυσμών. Ένα πάρα πολύ χρήσιμο εργαλείο είναι τα Εντροπικά Μέτρα Απόκλισης. Με τη χρήση αυτών των μέτρων μπορούμε να ποσοτικοποιήσουμε την 'απόσταση' μεταξύ τυχαίων διεργασιών. Αρκετές φορές όμως οι ερευνητές θέλουν να επικεντρωθούν σε πιο 'ειδικά τμήματα' των τυχαίων διεργασιών. Λύση σε αυτό το πρόβλημα έρχονται να δώσουν τα Σταθμισμένα Εντροπικά Μέτρα Απόκλισης. Χρησιμοποιώντας αυτού του τύπου μέτρα απόκλισης μπορεί η επιστημονική έρευνα να εκμεταλευθεί και να αξιοποιήσει τα πολλαπλά πλεονεκτήματά τους. Επίσης μπορούν να χρησιμοποιηθούν σε πληθώρα επιστημονικών πεδίων όπως για παράδειγμα η Χρηματοοικονομική Ανάλυση Ρίσκου, οι Γεωεπιστήμες, η Μετεωρολογία κλπ. όπου οι ερευνητές δίνουν συχνά περισσότερη βαρύτητα στις ουρές των κατανομών.

Contents

Chapter 1

Introduction

Information theory is a branch of pure and applied sciences that deals with the quantification of information. It started as being a key player in modern communication theory by formulating a communication system 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 three great researchers, namely 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 (1948), Shannon and Weaver (1949)) and consequently triggered the enormous development of the field, in the years that follow.

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, and more specifically in the field of thermodynamics (Clausius, 1865) while its statistical definition was developed by Boltzmann (1872) 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 can find 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 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.

Chapter 2

Literature Review

In the days of industrial age, around 1760 to around 1840, engineers try to construct a perpetual motion machine. After many failed attempts, they formulated a law of conservation of energy. They name it the "First law of thermodynamics". Engineers try again to construct a machine that derives energy in the form of heat from a hot body and convert it to equal amount of work. After many attempts and designs Clausius (1865) proposed the "Second law of thermodynamics". This law describes that does not exist a cyclic process that transfers heat from a cold body to a warm body (every process needs help by external work). The existence of entropy is based inevitably on the first and second law of thermodynamics. The role of entropy in thermodynamics is to quantify the irreversibility of a thermodynamic process.

Boltzmann (1866) re-interpreted the second law of thermodynamics in terms of the number of possible atomic arrangements. By this situation, Boltzmann laid the foundation of statistical approach to thermodynamics. Boltzmann's entropy is the basis to all statistical concepts of entropy.

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. Theoretical information scientists and communication engineers are more interested with transmit messages of a given form, despite the content of specified messages. Claude Shannon (Shannon & 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. For a simple coin with probability of tails equal to p, the entropy is 1 if $p = \frac{1}{2}$ (full uncertainty since we cannot expect one outcome over the other) and 0 if $p = 0$ or $p = 1$ (no uncertainty since the outcome is certain). The mutual information measures the mutual dependence between two variables by quantifying the "amount of information" (in unit of bit) 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–Charvát structural α-

entropy (1967). 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 entropy (1961) 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 Maximum Entropy Principle which proposed by Jaynes (1957) states that the probability distribution which describes better the dataset is that with the largest entropy. That is, it is that distribution obtained by maximizing the entropy measure (i.e. Shannon's) subject to given constraints, a process that resorts to the familiar method of constrained maximization using Lagrange multipliers. It has plenty of applications in Finance.

Minimum Cross-Entropy Principle (Jeffreys, 1946) is a measure of correlation of two probability distributions. Minimum Cross-entropy of two distributions constraints measure the closeness of two distribution according to Kullback–Leibler divergence measure (see below). Finally, Minimum Cross-Entropy Principle has important applications to Finance.

The relation between Information theory and Statistics was proposed by Kullback and Leibler (1951). They extended the notion of Shannon entropy and created a measure of divergence called Kullback-Leibler Divergence or "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. 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 in his work the Akaike Information Criterion (AIC) by constructing an unbiased estimator of the expected Relative entropy. 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 to 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 will 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. The advantages of the Weighted Entropy-type measures will be clear at the following example. 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. Guiasu (1971) was the first who proposed Weighted entropy and established the properties for this new type of entropy.

Chapter 3

Entropy-type Measures

In Sections 3.1-3.3 we provide the mathematical background and the relevant definitions associated with entropies and divergences. Section 3.4 is denoted to a brief discussion of applications.

3.1 Mathematical Background

Definition 1. *(Topological Space)*

A topological space $(X, \mathcal{T})$ is a set X and a collection $\mathcal{T} \subseteq \mathcal{P}(X)$ of subsets of X, called open sets, such that

1. $\emptyset, \quad X \in \mathcal{T}$
2. *if $\{U_a \in \mathcal{T} : a \in I\}$ is an arbitrary collection of open sets, then their union is open, hence:*

$$\bigcup_{a \in I} U_a \in \mathcal{T}$$

3. *if $\{U_i \in \mathcal{T} : i = 1, 2, ..., N\}$ is a finite collection of open sets, then their intersection is open, hence:*

$$\bigcap_{i=1}^{N} U_i \in \mathcal{T}$$

The complement of an open set in X is called a closed set, and $\mathcal{T}$ is called a topology on X.

Note*: $\mathcal{P}(X)$ is the power set of X which is the set of all possible subsets of X.*

Definition 2. *(σ-Algebra)*

A σ-algebra on a set X is a collection A of subsets of X such that:

1. $\emptyset, X \in \mathcal{A}$
2. *if* $A \in \mathcal{A}$ *then* $A^c \in \mathcal{A}$
3. *if* $A_i \in \mathcal{A}$ *for* $i \in \mathbb{N}$ *then*

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}, \qquad \left(\textit{or equivently} \quad \bigcap_{i=1}^{\infty} A_i \in \mathcal{A} \right)$$

From de Morgan's laws, a collection of subsets is σ-algebra if it contains $\emptyset$ and is closed under the operations of taking complements and countable unions (or, equivalently, countable intersections).

Note: If the union of 3 is finite then the collection $\mathcal{A}$ of subsets is called Algebra.

Definition 3. *(Borel σ-Algebra)*

Let $(\mathcal{X}, \mathcal{T})$ be a topological space. The Borel σ-algebra

$$\mathcal{B}(X) = \sigma(\mathcal{T})$$

is the σ-algebra generated by the collection $\mathcal{T}$ of open sets on X.

Definition 4. *(Measurable Space)*

A measurable space $(X, \mathcal{A})$ is a non-empty set X equipped with a σ-algebra $\mathcal{A}$ on X.

Definition 5. *A measure μ on a measurable space $(X, \mathcal{A})$ is a function*

$$\mu : \mathcal{A} \to [0, \infty]$$

such that

a) $\mu(\varnothing) = 0$

b) *if* $\{A_i \in \mathcal{A} : i \in \mathbb{N}\}$ *is a countable disjoint collection of sets in* $\mathcal{A}$*, then*

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$$

Definition 6. *(Metric)*

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

1. $d(x,y) = 0 \;\; iff \;\; x = y \, \forall x, y \in X$.
2. $d(x,y) = d(y,x) \;\; \forall x, y \in X$.
3. $d(x,y) \leq d(x,z) + d(z,y) \;\; \forall x, y, z \in X$.

Based on the above definitions which set the basic mathematical background needed, we are now ready to define Entropic and Divergence measures (Sections 3.2 & 3.3).

3.2 Divergence Measures

Definition 7. *(Divergence Measure)*

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:

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

Note that divergence measures are not necessarily metrics because they do not have to be symmetric or fulfil the triangular inequality (See Definition 6).

3.3 Entropy-type Measures

3.3.1 Shannon Entropy

Definition 8. *(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

$$\begin{aligned} H(X) &= E\left[\log \frac{1}{P_X(X)}\right] \\ &= \sum_{x \in S_X} p_X(x) \log \frac{1}{p_X(x)} \end{aligned}$$

This measure of uncertainty has many important properties which agree with our intuitive notion of randomness.

1. *It is always positive*
2. *It is zero if and only if X describes a certain event*
3. *It increases by adding an independent component and decreases by conditioning.*

Now, we provide the Shannon entropy for continuous distributions. A straightforward extension is given in the following definition.

Definition 9. *(Differential Entropy)*

Let a stochastic source described by a continuous random variable X with distribution F_X, support S_X and probability density function f_X. The differential entropy of X is

$$H(X) = -\int_{S_X} f(x) \log f(x) dx$$

The properties for differential entropy are the following:

1. *It is only defined for distributions with densities.*
2. *The entropy of a discrete distribution is always positive, whereas the differential entropy of a continuous variable may take any value on the extended real line.*
3. *It is "inconsistent" in the sense that the differential entropy of a uniform distribution in an interval of length α is $\log a$, which is zero if $\alpha = 1$, negative if $\alpha < 1$ and positive if $\alpha > 1$.*
4. *The differential entropy of a continuous variable decreases by conditioning.*

3.3.2 Cross Entropy

Definition 10. *(Cross Entropy)*

Consider two distributions P, Q with probability mass functions $\underset{\sim}{p} = (p_1, \ldots, p_n)^T$ and $\underset{\sim}{q} = (q_1, \ldots, q_n)^T$ respectively. Then the discrete version of Cross Entropy is defined by:

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

with properties:

1. *$H(P,Q) \neq H(Q,P)$.*
2. *The Cross Entropy equals the negative log likelihood.*

3.3.3 Relative Entropy

Definition 11. *(Discrete case)*

Consider two distributions P, Q with probability mass functions $\underset{\sim}{p} = (p_1, \ldots, p_n)^T$ and $\underset{\sim}{q} = (q_1, \ldots, q_n)^T$ respectively. Then the discrete version of Relative Entropy is defined by:

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

Definition 12. *(Continuous case)*

Consider two distributions P, Q with probability density functions $\underset{\sim}{p} = (p_1, \ldots, p_n)^T$ and $\underset{\sim}{q} = (q_1, \ldots, q_n)^T$ respectively. Then the continuous version of Relative Entropy is defined by:

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

As we can see, the Relative entropy is not symmetric. For this case, we can define a useful type of distance, called Jeffrey's Distance (Jeffreys (1946)) which is symmetric and also related to the Relative Entropy.

3.3.4 Jeffrey's Distance

Definition 13. *(Discrete case)*

Consider two distributions P, Q with probability mass functions $\underset{\sim}{p} = (p_1, \ldots, p_n)^T$ and $\underset{\sim}{q} = (q_1, \ldots, q_n)^T$ respectively. Then the discrete version of Jeffrey's Distance is defined by:

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

Definition 14. *(Continuous case)*

Consider two distributions P, Q with probability density functions $\underset{\sim}{p} = (p_1, \ldots, p_n)^T$ and $\underset{\sim}{q} = (q_1, \ldots, q_n)^T$ respectively. Then the continuous version of Jeffrey's Distance is defined by:

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

3.4 Statistical Conjectures on Entropy-type measures

A few of the many uses of Entropy-type measures appear in the subsections that follow.

3.4.1 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 method that we easily distribute two 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. We focus on the clustering with relative entropy. The relative entropy as we saw is the notion that quantifies the distance between two densities functions.

The Relative entropy technique is quite similar with the Mahalanobis distance (1936). The main difference is 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. The use of Relative entropy is 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 (SOM) (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.

3.4.2 Goodness of Fit based on Entropy-type measures

The Goodness of Fit tests are important tools to 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 $D(\cdot, \cdot)$.

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

Let A a category, O_i the frequency of results belongs to A_i and $e_i = E(O_i),\ i = 1, \ldots, k$

then the maximum likelihood (ML) estimator of q_i is $\widehat{q_i} = \dfrac{O_i}{n}$ and $p_i = \dfrac{e_i}{n}$.
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 $O = (O_1, \ldots, O_k) \sim M_k\left(n, (q_1, \ldots, q_k)\right)$, where M_k is k-dimensional multinomial distribution. For the very big sample size the vector O an asymptotic multivariate normal distribution $N_k\left(nq, n\left(D_p - qq'\right)\right)$ D_q is a diagonal matrix with diagonal elements q_i $i = 1, \ldots, k$ and $q = (q_1, \ldots, q_k)$. Thus

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

Simple algebra shows that:

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

where,

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

Now we can see from above and the asymptotic distribution of Z that we can reject $H_0 : D(Q, P) = 0$ in favor of $H_1 : D(Q, P) > 0$ 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 $1 - a$ 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, Pearson 1933) but in simulations (Sharifdoost et al., 2009) it appears to be more sensitive in G^2. The goodness of fit test based on Relative entropy are more sensitive than the usual methods for rejecting distributions which are close to the distribution we want as (Sharifdoost et al., 2009).

3.4.3 Model Selection based on Entropy-type measures

Model Selection is the field of statistics that allows us to select the right statistical model from a set of candidate models. As we know, Model Selection plays important role in Mathematical Statistics. The first scientist who studied deeply this sense was Akaike (1973) who proposed the Akaike Information Criterion (AIC) by constructing an unbiased estimator of the expected Relative entropy.

Let f be the "reality" (i.e. the true model) and g a model used to estimate f. The Relative entropy (in continuous case) 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:

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

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

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

By computing $E_f[\log(g(x|\theta))]$ we easily obtain the relative distance $D(f,g) - z$ between f and g. Instead of this quantity which can not be computed Akaike found that its expectation:

$$E_f[E_f[\log(g(x|\theta))]]$$

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:

$$\log(\mathcal{L}(\hat{\theta}|\underset{\sim}{x})) - p$$

where p is the dimension of the parameter θ and $\hat{\theta}$ is a consistent estimate of θ. Then the AIC is:

$$AIC = -2\log(\mathcal{L}(\hat{\theta}|\underset{\sim}{x})) + 2p$$

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. based on the same methodology as AIC criterion used the BHHJ Divergence (Basu et al., 1998) for developing a new criterion.
Suppose 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.

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

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 gave us an unbiased estimator is:

$$EW_\theta = E\left(W_\theta \mid \theta = \hat{\theta}\right)$$

where $\hat{\theta}$ is asymptotically normal estimator of θ. We can also say that the previous formula is the average distance between f and g_θ.

Now, we present an unbiased estimator of the expected overall discrepancy

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

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

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

The adjusted DIC model is given below (Mantalos et al., 2010).

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

$$DIC_C = nQ_{\widehat{\theta}} + (2\pi)^{-\frac{a}{2}}(1+a)^{2+\frac{p}{2}} p.$$

We can easily observe that MLE method is faster in computations than the Basu method. Also, the DIC criterion has highly performance of accuracy in simulations. It could be used in applications with outlier and contaminated observations. All these give us a powerful criterion for model selection.

Chapter 4

Weighted Entropy-type measures

4.1 Weighted Entropy-type measures

Sometimes the entropy is not as useful as expected. For example if we wish to focus on a specific characteristic (for instance the tail part) of a distribution more than others then, we have to give different weights on different parts of the distribution. The same occurs if we wish to compare specific characteristics of two populations. This desire drives scientists to re-build the original Shannon entropy to a Weighted one. Guiasu (1971) was the first who proposed the Weighted entropy.

4.1.1 Weighted Shannon Entropy

Definition 15. *(Weighted Shannon Entropy)*

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

$$H^w(X) = \sum_{i=1}^{n} w_i p_i \log \frac{1}{p_i}. \tag{4.1}$$

The standard properties of the Weighted Shannon Entropy are:

1. $H^w(X) \geq 0$.

2. If $w_1 = w_2 = ... = w_n = w$, then $H^w(X) = wH(X)$, where $H(X)$ is the Shannon entropy.

3. If $p_i = 1$ for some $i = 1, ..., n$ then $H^w(X) = 0$ irrespectively of the values of the weights $\underset{\sim}{w}$.

This property stresses that if only one event is possible then there is no uncertainty and does not provide any information. So the weighted Shannon entropy is equal to zero.

4. If $p_i = 0,\ w_i \neq 0\ \ \forall i \in I$ and $p_j \neq 0,\ w_j = 0\ \ \forall j \in J$ where $I \cup J = \{1, 2, ..., n\}$, $I \cap J = \emptyset$, then $H^w(X) = 0$.

5. $H^w(w_1, ..., w_{n+1}; p_1, ..., p_n, 0) = H^w(w_1, ..., w_n; p_1, ..., p_n) = H^w(X)$, for any w_{n+1}.

6. For every non-negative, real number λ we have $H^w(\lambda \underset{\sim}{w}; \underset{\sim}{p}) = \lambda H^w(\underset{\sim}{w}, \underset{\sim}{p}) = \lambda H^w(X)$.

$$w(E \cup F) = \frac{p(E)w(E) + p(F)w(F)}{p(E) + p(F)} \tag{4.2}$$

where $w(F)$ is the weight of event F and $p(F)$ the probability of the same event.

In addition if E, F are complementary events, then:

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

7. If the rule (4.2) for the weights holds, then:

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

where $w_n = \dfrac{p'w' + p''w''}{p' + p}$, $p_n = p' + p''$.

In following two Sections we introduce the weighted version of the Relative entropy and its symmetric counter part, Jeffrey's Distance.

4.1.2 Weighted Relative Entropy

Definition 16. *(Weighted Relative Entropy)*

Consider two probability mass functions $\underset{\sim}{p} = (p_1, \ldots, p_n)^T$, $\underset{\sim}{q} = (q_1, \ldots, q_n)^T$ *and* $\underset{\sim}{w} = (w_1, \ldots, w_n)^T$ *a vector of weights. Then the discrete version of weighted Relative entropy is defined by:*

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

This form of relative entropy is not a proper distance measure (i.e. ≥ 0*), because it can take negative values.*

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 way to prove that the Relative entropy is non-negative. We take

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

Now, we take the sum with weights $\underset{\sim}{w} = (w_1, \dots, w_n)^T$

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

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, $w_i = \dfrac{q_i}{p_i}$.
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)

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

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

Relative entropy = Cross entropy - Shannon entropy =

$$= \sum_{i=1}^{n} q_i \log\left(\frac{q_i}{p_i}\right) = E_q\left[\log\left(\frac{q_i}{p_i}\right)\right] \geq 0.$$

i.e. the subjective-objective measure of uncertainty is greater than the 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 $p_i's$.
Of course if $p_i = q_i$ then we have equality.

Similarly, for $E_p\left[\log\left(\dfrac{p_i}{q_i}\right)\right]$ where p are assumed to be the theoretical probabilities and q_i is the estimate of p_i.

□

4.1.3 Weighted Jeffrey's Distance

Definition 17. *(Weighted Jeffrey's Distance)*

Consider two probability mass functions $\underset{\sim}{p} = (p_1, \ldots, p_n)^T$, $\underset{\sim}{q} = (q_1, \ldots, q_n)^T$ *and* $\underset{\sim}{w} = (w_1, \ldots, w_n)^T$ *a vector of weights. Then the discrete version of weighted Jeffrey's Distance is defined by:*

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

Proposition 2. *At least one term of Weighted Jeffrey's Distance is non-negative.*

Proof. Weighted Jeffrey's Distance: $D_J^w(P,Q) = \sum_{i=1}^{n} w_i p_i \log\left(\frac{p_i}{q_i}\right) + \sum_{i=1}^{n} w_i q_i \log\left(\frac{q_i}{p_i}\right)$

The left part of the Jeffrey's Distance is non-negative. We prove it in the previous proposition. The right part is a complementary term of Relative entropy.

Thus, the Weighted Jeffrey's Distance:

$$D_J^w(P,Q) \geq 0 \quad \forall x \in S.$$

□

Observe that some of the terms in the expressions of D^w and D_J^w may be negative and therefore will not appropriate as distance measure if the researcher wishes to focus exclusively on them. The issue is resolved below with the proposal of the Absolute Weighted measures.

4.2 Absolute Weighted Entropy-type measures

4.2.1 Absolute Weighted Relative Entropy (A.W.R.E)

Definition 18. *(A.W.R.E)*

Consider two probability mass functions $\underset{\sim}{p} = (p_1, \ldots, p_n)^T$, $\underset{\sim}{q} = (q_1, \ldots, q_n)^T$ *and* $\underset{\sim}{w} = (w_1, \ldots, w_n)^T$ *a vector of weights. Then the discrete version of Absolute Weighted Relative Entropy is defined by:*

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

This modification makes the Absolute Weighted Relative Entropy (A.W.R.E) being always non-negative and provides the researcher with a useful tool. By using A.W.R.E, we ensure the non-negativity and we are able to give more "attention" to special parts of the distributions that we measure. If we wish to have a measure that satisfies the symmetric property then we could generalize the Jeffrey's Distance by introducing the Absolute Weighted Jeffreys Distance (A.W.J.D) defined below.

4.2.2 Absolute Weighted Jeffrey's Distance (A.W.J.D)

Definition 19. *(A.W.J.D)*

Consider two probability mass functions $\underset{\sim}{p} = (p_1, \ldots, p_n)^T$, $\underset{\sim}{q} = (q_1, \ldots, q_n)^T$ *and* $\underset{\sim}{w} = (w_1, \ldots, w_n)^T$ *a vector of weights. Then the discrete version of Absolute Weighted Jeffrey's Distance is defined by:*

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

The standard as well as the absolute version of the Relative Entropy and Jeffrey's Distance will be implemented in various scenarios in the following 2 chapters and their performance will be evaluated via simulations and 2 case studies in Financial Mathematics and Applied Sciences.

Chapter 5

Simulations

In this Chapter, we analyse the theoretic content that we presented previously via simulations. More specifically, we present the relation (i.e. the "distance") between the standard normal distribution $N(0,1)$ and t-student distribution with various degrees of freedom. For illustrative purposes we will be using $k = 2, 5, 10, 30$ degrees of freedom. The results could be extended to any value.

For doing this, we used both the Weighted Relative Entropy and the Weighted Jeffrey's Distance. Furthermore, we apply the Absolute Weighted Relative Entropy (A.W.R.E) and Absolute Weighted Jeffrey's Distance (A.W.J.D) defined in the previous Chapter.

Our intention is to show that by using appropriate weights we could reveal the differences on specific parts of the distributions under investigation. First of all, we present a graph with the distributions to be examined.

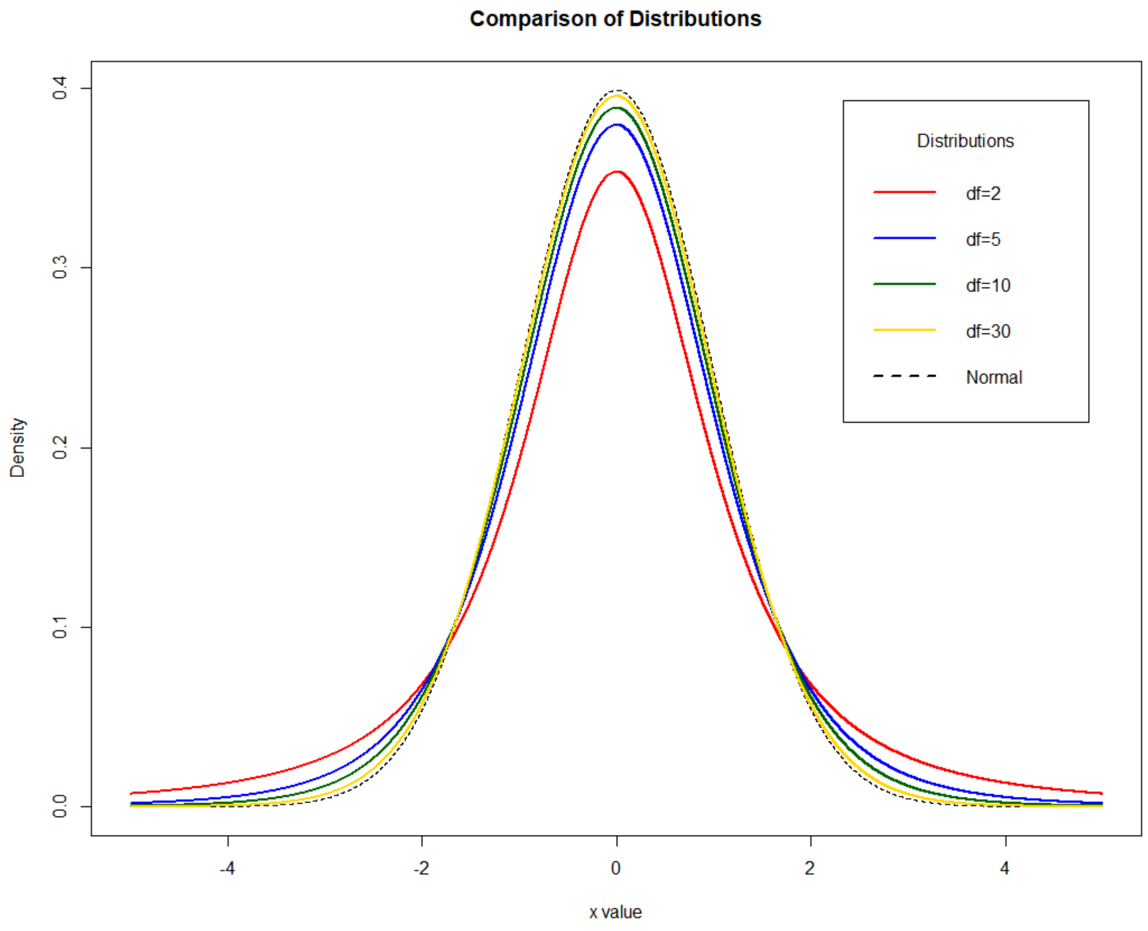

Figure 5.1: Comparison of Distributions

For the implementation of the proposed methodology the support of the distribution is divided into n number of intervals (in this case $n = 10$).

The subintervals of the support used are:

$$[-\infty,-3)\cup[-3,-2)\cup[-2,-1)\cup[-1,-0.5)\cup[-0.5,0)\cup[0,0.5)\cup[0.5,1)\cup[1,2)\cup[2,3)\cup[3,\infty]$$

Now, we will present the methods.

5.1 Standard normal - t2

We generate 10000 random numbers from standard normal and t-student with 2 degrees of freedom respectively. The probabilities that emerged in each interval for distribution comparison is given by the following table.

Probabilities by interval										
Intervals	$(-\infty,-3)$	$[-3,-2)$	$[-2,-1)$	$[-1,-0.5)$	$[-0.5,0)$	$[0,0.5)$	$[0.5,1)$	$[1,2)$	$[2,3)$	$[3,\infty]$
Std normal	0.001	0.022	0.133	0.153	0.188	0.190	0.151	0.137	0.019	0.001
t2	0.049	0.041	0.120	0.122	0.167	0.166	0.123	0.119	0.042	0.048

Table 5.1: Probabilities by interval $N(0,1)$ vs t_2

5.1.1 Middle method

The method that we present now is called "Middle method" and uses the probabilities from the previous table for calculations. The method will be applied to the four Entropy-type measures mentioned above and compare them.

Middle method algorithm

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

   (c) Absolute Jeffrey's Distance

Now, we furnish the following graph for the comparison of different Entropy-type measures for the Standard normal and t_2 distributions.

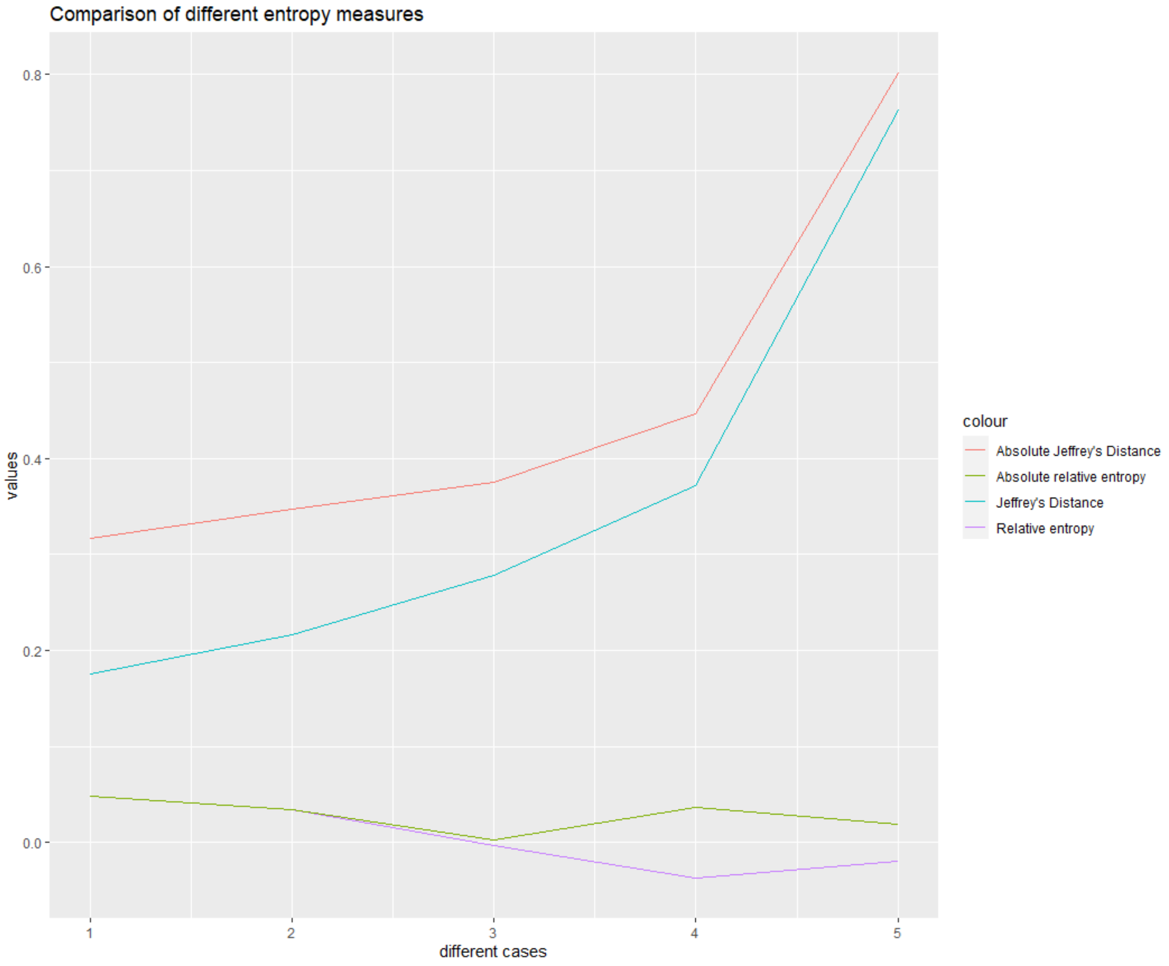

Figure 5.2: Middle method $N(0,1)$ vs t_2

5.1.2 Left to the Right (LR) method

The method that we present now is called ”Left to the right method” or ”LR method”. Using the probabilities of Table 5.1 we calculate the same Weighted Entropy-type measures as before.

LR method algorithm

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

2. Remove the first interval from the left and use the remaining by calculating the Relative entropy (weights: $w_i = \frac{n}{n-1}$ for each interval).

3. Repeat by removing the first two intervals from the left and use the remaining by calculating the Relative entropy. (weights: $w_i = \frac{n}{n-2}$ for each interval).

4. Repeat the algorithm by removing 1 more interval from the left at each step and increasing accordingly by an equal amount the weights so that $\sum w_i = n$.

5. Repeat the steps 1-4 for

    (a) Jeffrey’s Distance

    (b) Absolute Relative Entropy

    (c) Absolute Jeffrey’s Distance

Now, we furnish the following graph for the comparison of different Entropy-type measures for the Standard normal and t_2 distributions.

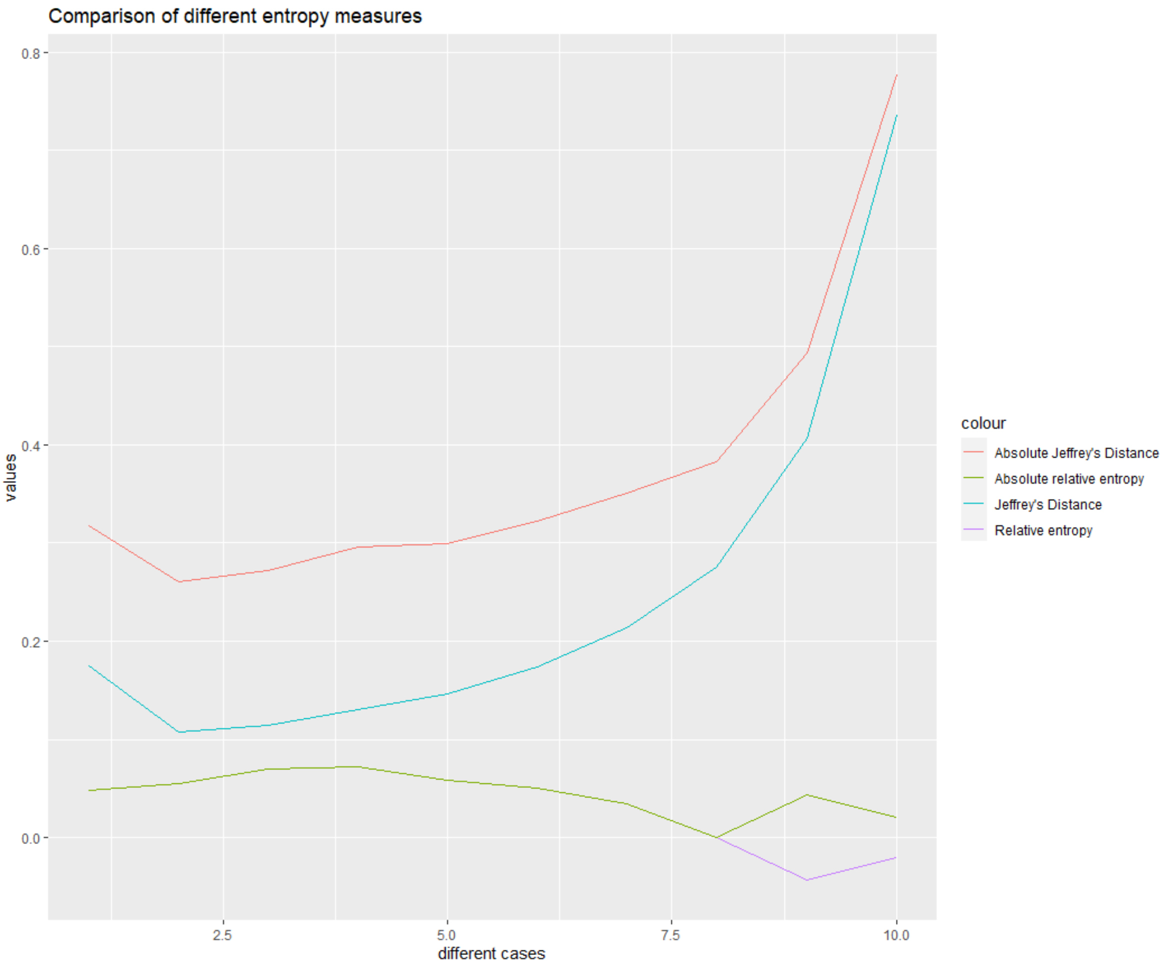

Figure 5.3: LR method $N(0,1)$ vs t_2

5.1.3 Right to the Left (RL) method

This method is called ”Right to the Left method” or ”RL method”. The process of this method is similar to ”LR” method but the difference is that we remove intervals from Right to the Left by using the probabilities of the Table 5.1. for calculating Entropy-type measures.

RL method algorithm

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

2. Remove the last interval from the right and use the remaining by calculating the Relative entropy (weights: $w_i = \frac{n}{n-1}$ for each interval).

3. Repeat by removing the first two intervals from the right and use the remaining by calculating the Relative entropy. (weights: $w_i = \frac{n}{n-2}$ for each interval).

4. Repeat the algorithm by removing 1 more interval from the right at each step and increasing accordingly by an equal amount the weights so that $\sum w_i = n$.

5. Repeat the steps 1-4 for

   (a) Jeffrey's Distance

   (b) Absolute Relative Entropy

   (c) Absolute Jeffrey's Distance

Now, we furnish the following graph for the comparison of different Entropy-type measures for the Standard normal and t_2 distributions.

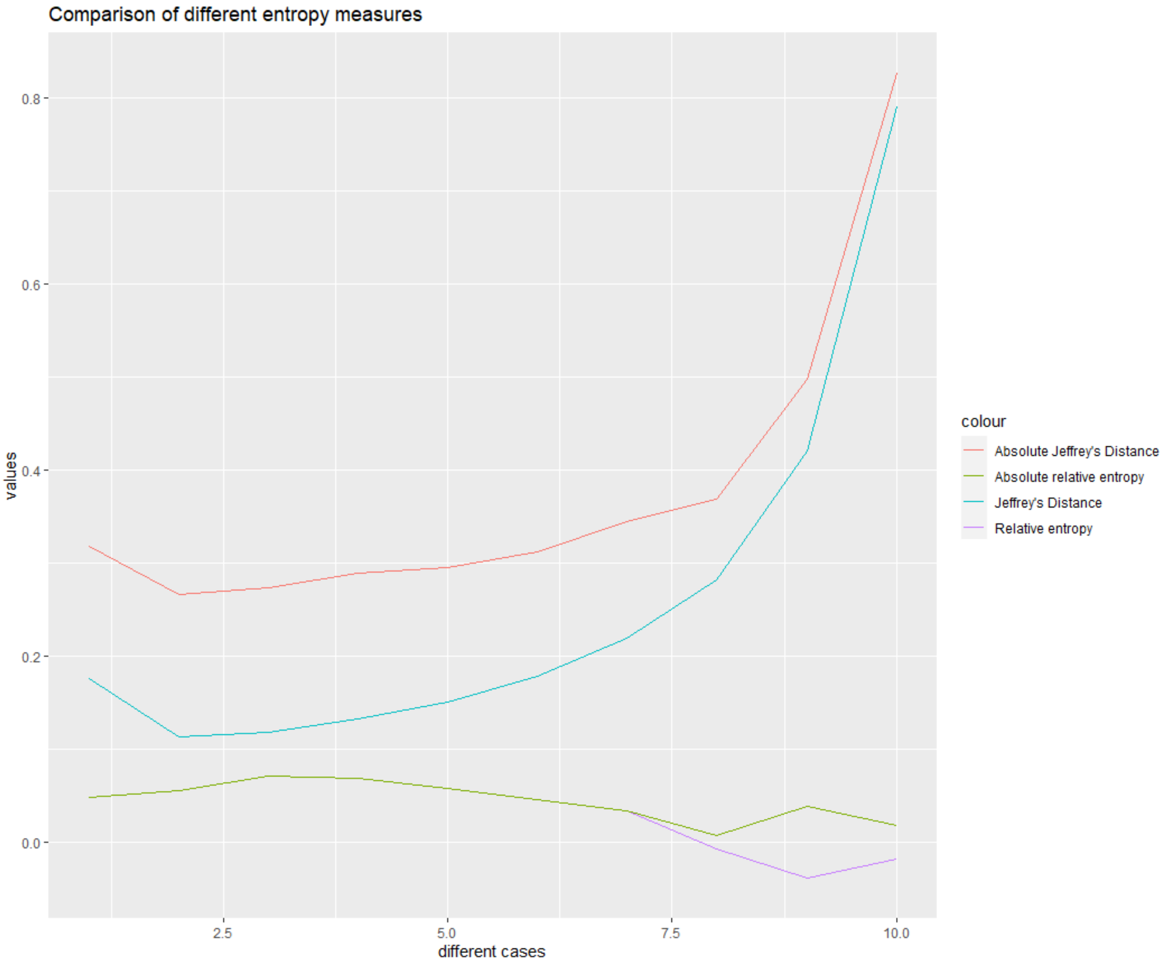

Figure 5.4: RL method $N(0,1)$ vs t_2

5.2 Standard normal - t5

We generate 10000 random numbers from standard normal and t-student with 5 degrees of freedom respectively. In this Section we compare the Standard normal and t_5 based on the data on Table 5.2. The graphs for the three methods have been obtained in exactly same fashion as in Section 5.1.

Probabilities by interval										
Intervals	$(-\infty,-3)$	$[-3,-2)$	$[-2,-1)$	$[-1,-0.5)$	$[-0.5,0)$	$[0,0.5)$	$[0.5,1)$	$[1,2)$	$[2,3)$	$[3,\infty]$
Std normal	0.001	0.022	0.133	0.153	0.188	0.190	0.151	0.137	0.019	0.001
t5	0.014	0.033	0.128	0.134	0.178	0.182	0.143	0.129	0.036	0.018

Table 5.2: Probabilities by interval $N(0,1)$ vs t_5

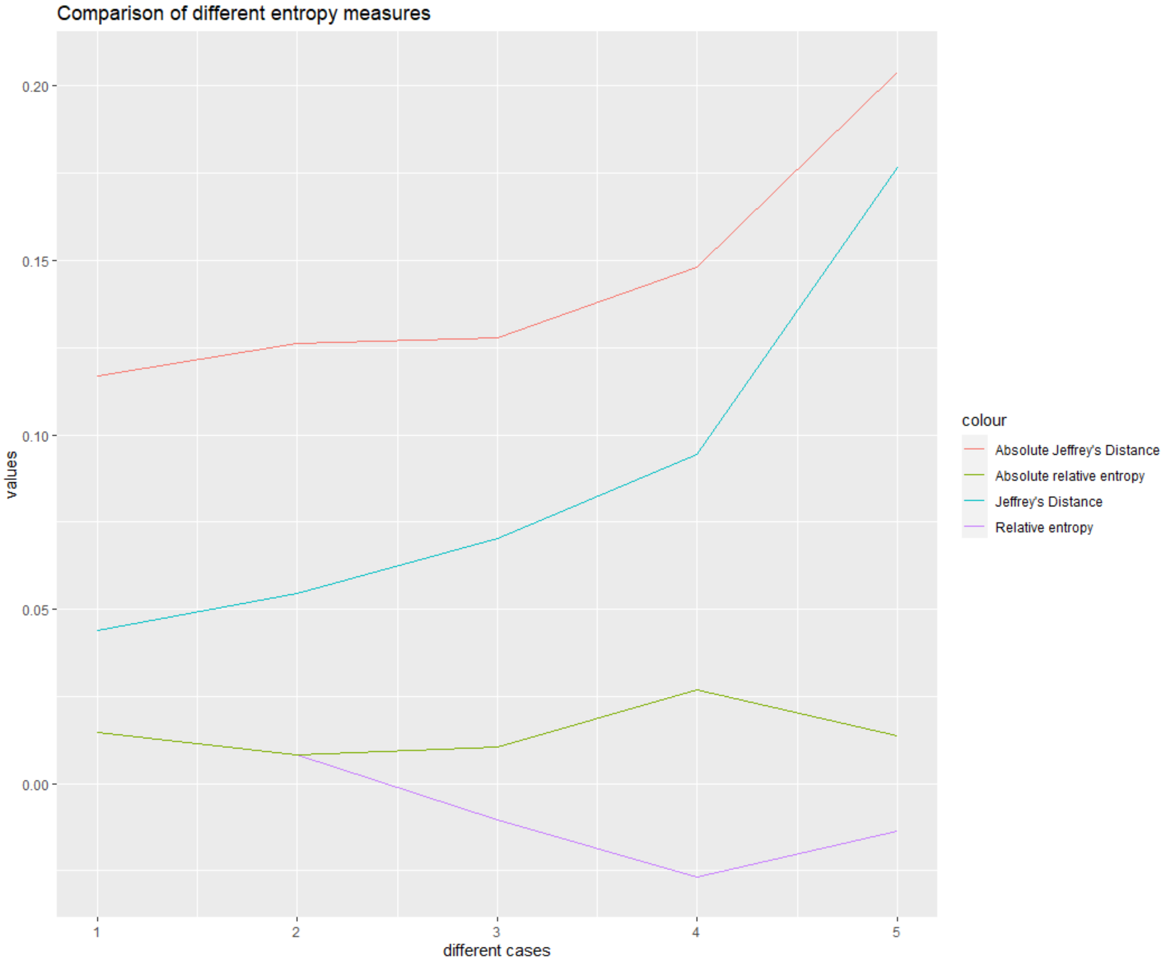

Figure 5.5: Middle method $N(0,1)$ vs t_5

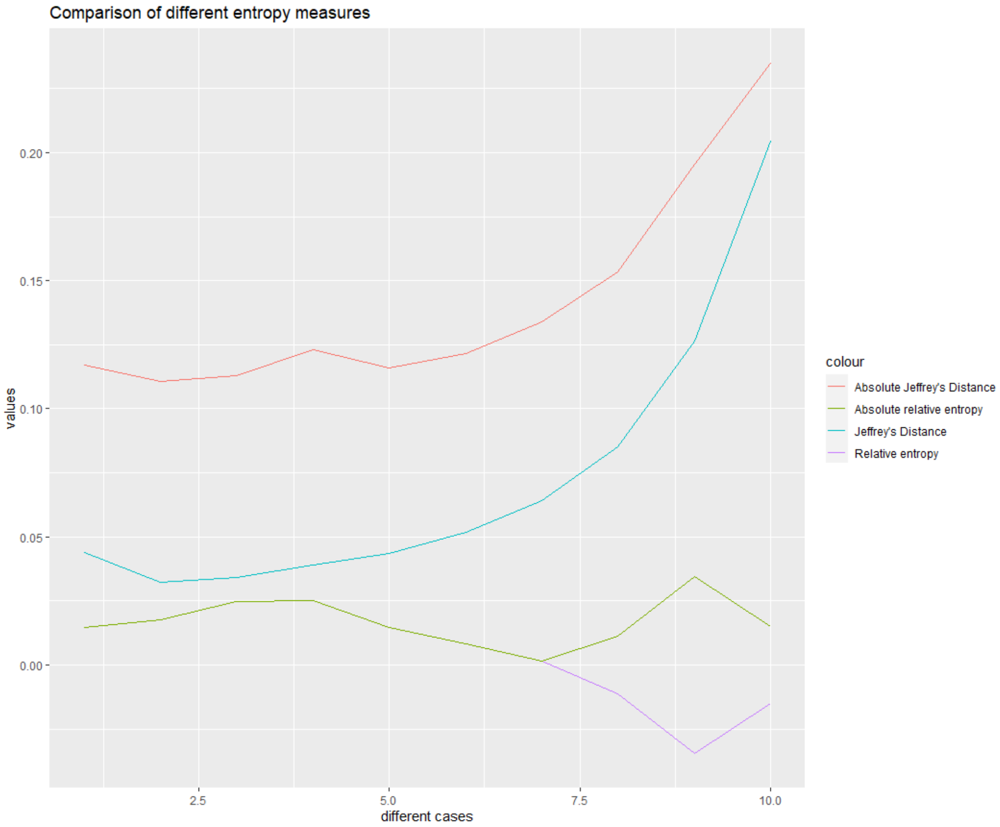

Figure 5.6: LR method $N(0,1)$ vs t_5

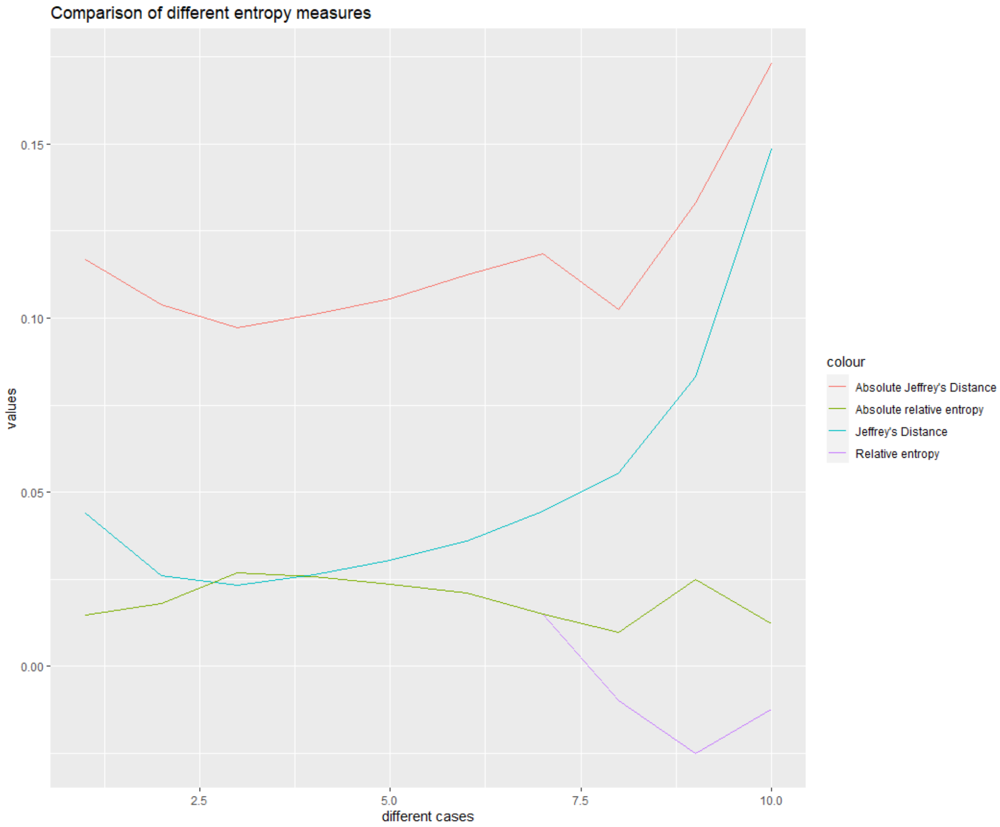

Figure 5.7: RL method $N(0,1)$ vs t_5

5.3 Standard normal - t10

We generate 10000 random numbers from standard normal and t-student with 10 degrees of freedom respectively. In this Section the comparison is for Standard normal and t_{10} based on Table 5.3. The results are presented in Figures 5.8, 5.9 and 5.10.

Probabilities by interval										
Intervals	$(-\infty,-3)$	$[-3,-2)$	$[-2,-1)$	$[-1,-0.5)$	$[-0.5,0)$	$[0,0.5)$	$[0.5,1)$	$[1,2)$	$[2,3)$	$[3,\infty]$
Std normal	0.001	0.022	0.133	0.153	0.188	0.190	0.151	0.137	0.019	0.001
t10	0.007	0.026	0.131	0.140	0.188	0.187	0.147	0.132	0.032	0.005

Table 5.3: Probabilities by interval $N(0,1)$ vs t_{10}

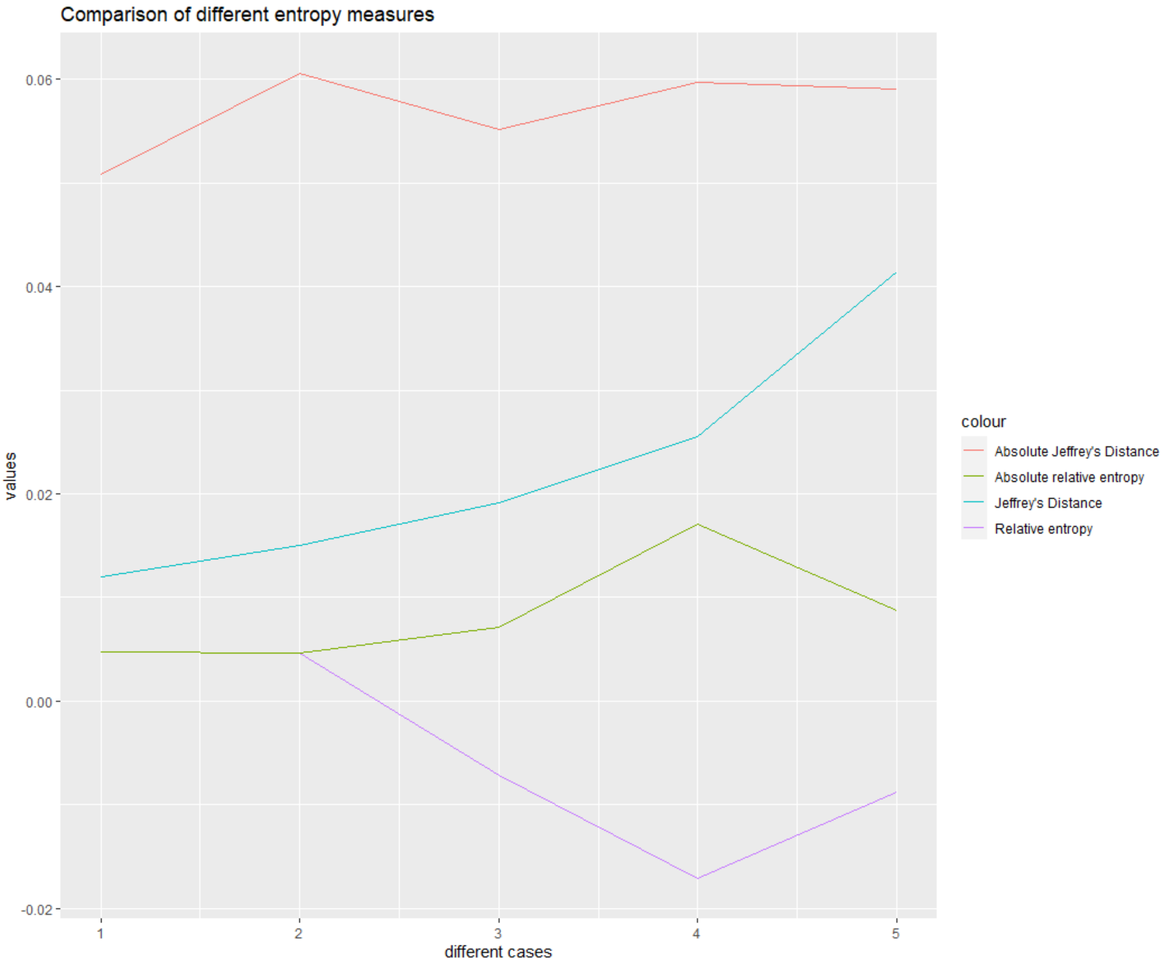

Figure 5.8: Middle method $N(0,1)$ vs t_{10}

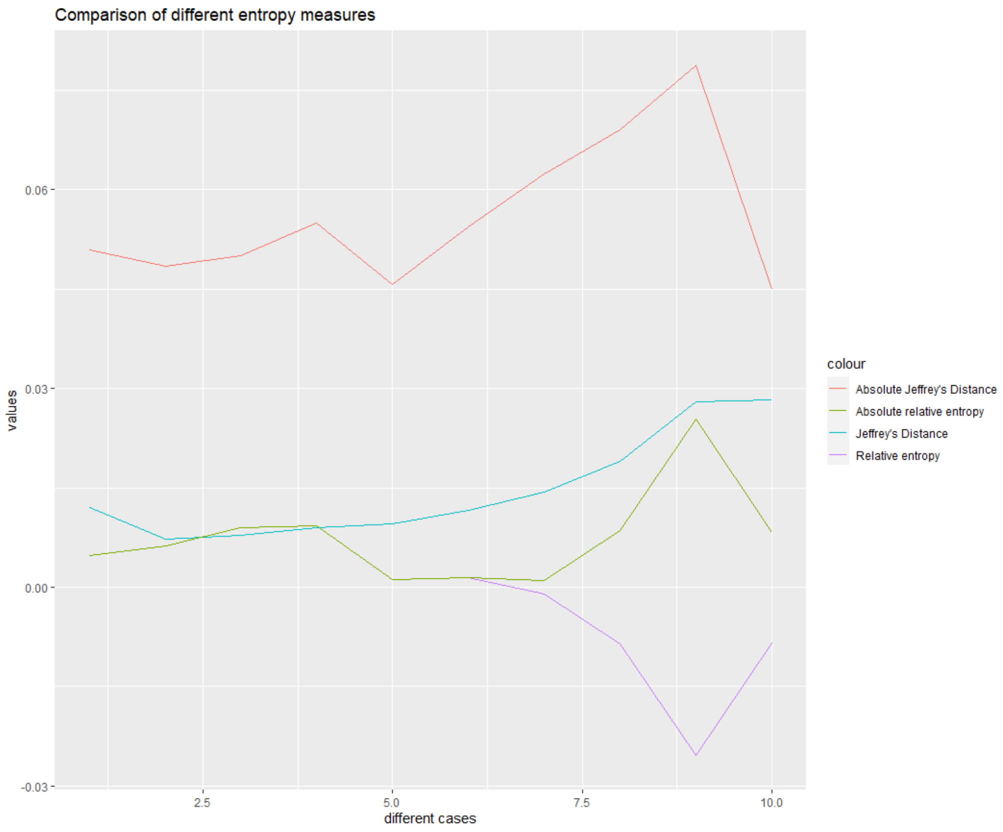

Figure 5.9: LR method $N(0,1)$ vs t_{10}

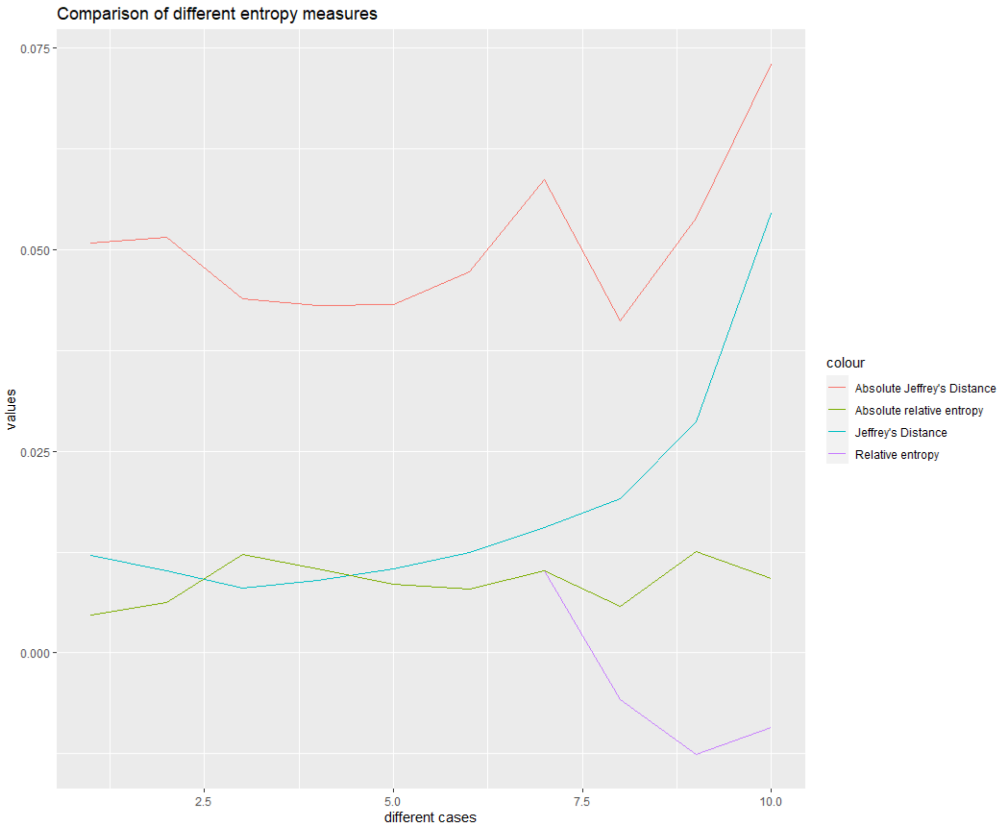

Figure 5.10: RL method $N(0,1)$ vs t_{10}

5.4 Standard normal - t30

We generate 10000 random numbers from standard normal and t-student with 30 degrees of freedom respectively. In this Section the comparison is for Standard normal and t_{30} based on Table 5.4. The results are presented in Figures 5.11, 5.12 and 5.13.

Probabilities by interval										
Intervals	$(-\infty,-3)$	$[-3,-2)$	$[-2,-1)$	$[-1,-0.5)$	$[-0.5,0)$	$[0,0.5)$	$[0.5,1)$	$[1,2)$	$[2,3)$	$[3,\infty]$
Std normal	0.001	0.022	0.133	0.153	0.188	0.190	0.151	0.137	0.019	0.001
t30	0.002	0.021	0.130	0.144	0.192	0.193	0.146	0.139	0.027	0.002

Table 5.4: Probabilities by interval $N(0,1)$ vs t_{30}

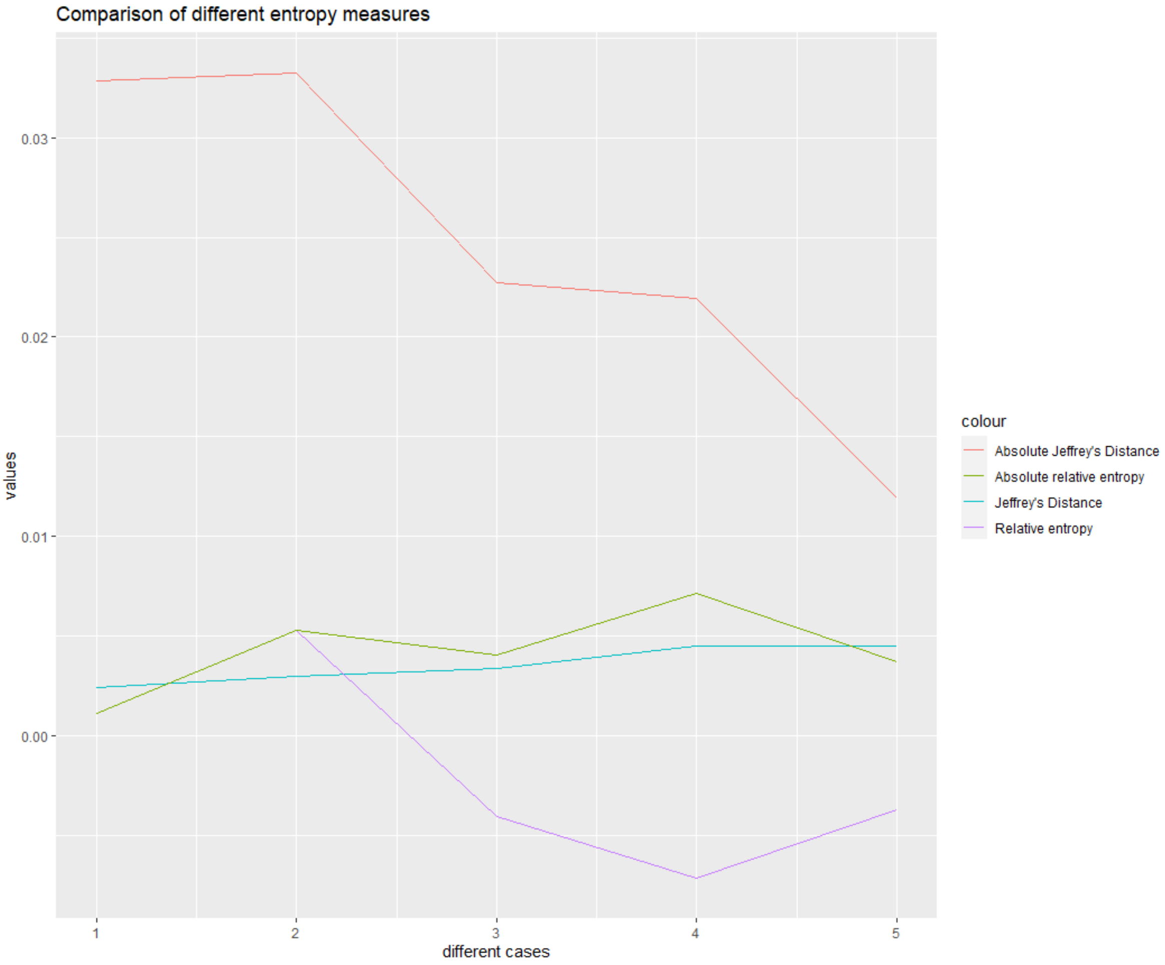

Figure 5.11: Middle method $N(0,1)$ vs t_{30}

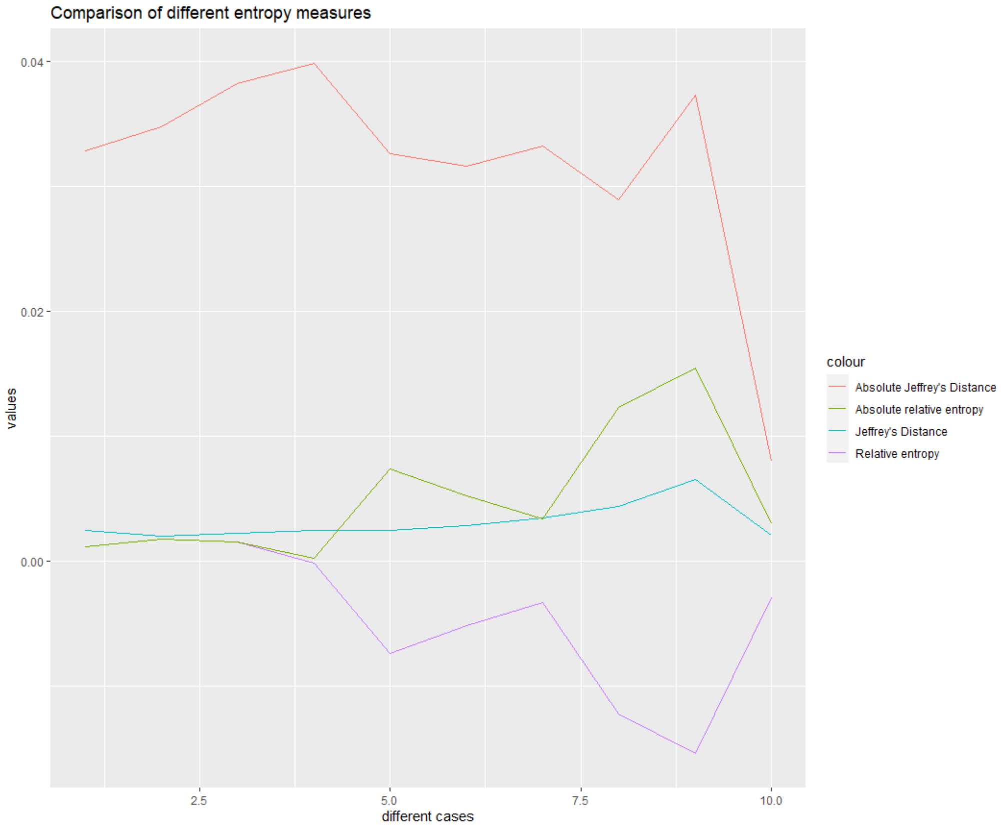

Figure 5.12: LR method $N(0,1)$ vs t_{30}

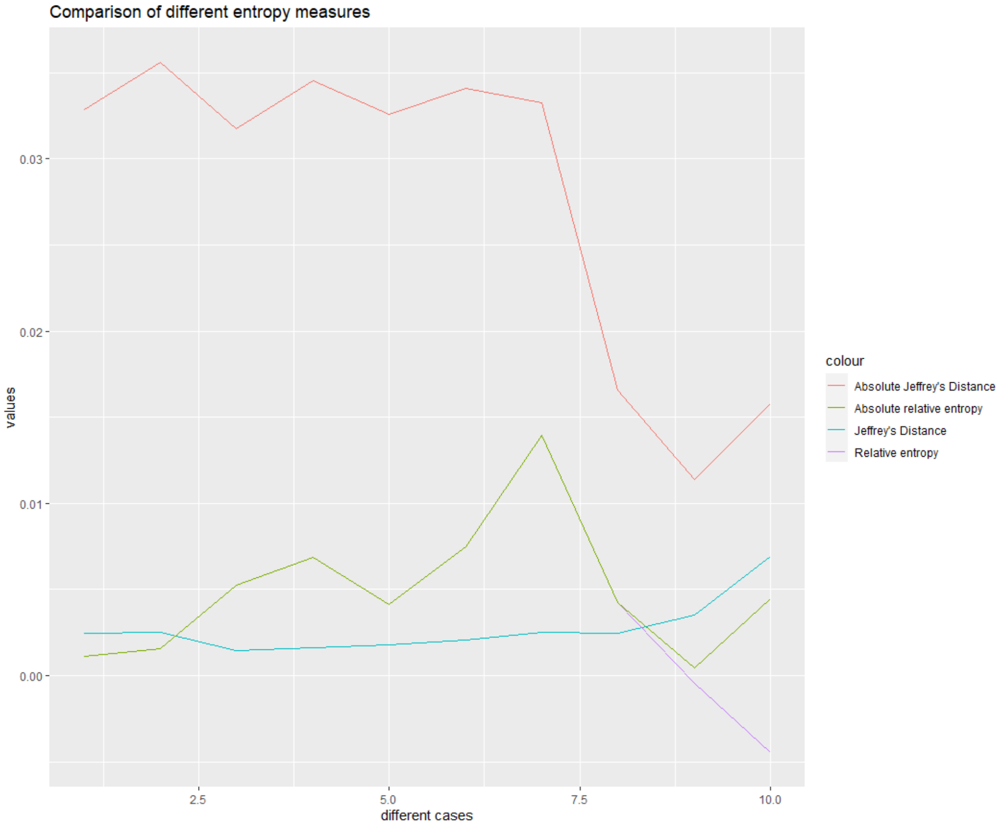

Figure 5.13: RL method $N(0,1)$ vs t_{30}

Conclusions

In Chapter 5 we used different Weighted Entropy-type measures for comparison of distributions. More specifically we focused on the tails of distributions. With the "Middle method", we remove the middle intervals of the support and we calculated four Weighted Entropy-type measure techniques. As we saw the Absolute Weighted Jeffrey's Distance (A.W.J.D) was the technique that gave us the larger distances among all distributions. The Weighted Relative Entropy was the worst of all because it takes negative values and contradicts the sense of distance.

The "LR method" which is the method by removing from left to the right the intervals show that the Absolute Weighted Jeffrey's Distance (A.W.J.D) had the greater distance values among all distributions. The Weighted Relative Entropy with the "LR method" takes also negative values which is meaningless for distribution comparison.

The "RL method" which is the method by removing intervals from right to left possesses the benefits of the (A.W.J.D) technique and avoid the disadvantages of Weighted Relative Entropy as a distance measure for all distributions examined.

Chapter 6

Applications

In this Chapter we will present two case studies in Financial Mathematics and Geosciences to clearly see the advantages of the Absolute Weighted Entropy-type measures as well as the Weighted Entropy-type measures.

6.1 Case Study - Financial Mathematics

We collected 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. We try to find the relationship between these two stocks via Absolute Weighted and Weighted Entropy-type measures. The dataset contains 1258 observations for each stock.

The reason that we used logarithmic returns against closing values is because in the Portfolio Theory the logarithmic returns follow the Normal Distribution. Thus, it has more sense to compare the returns of the stocks via Weighted Entropy-type measures.

Firstly, we present the histograms of our data separately.

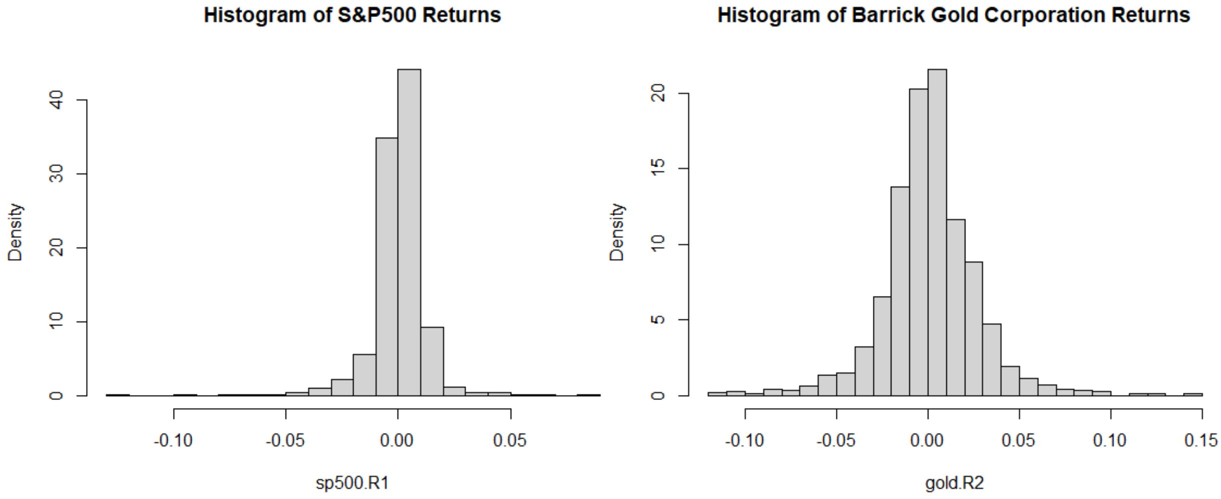

Figure 6.1: Histograms of S&P500 and GOLD

For S&P500 returns the average is 0.000483 and the standard deviation is 0.012184876. For Barrick Gold Corporation returns the average is 0.000929 and the standard deviation is 0.025566474. Now, for our method we will divide the support of the dataset in the following way:

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

$$\cup [0.02, 0.05) \cup [0.05, 0.08) \cup [0.08, \infty)$$

The following table provides the percentage of data in every interval of the dataset.

Percentages by interval										
Intervals	$(-\infty, -0.08)$	$[-0.08, -0.05)$	$[-0.05, -0.02)$	$[-0.02, -0.01)$	$[-0.01, 0)$	$[0, 0.01)$	$[0.01, 0.02)$	$[0.02, 0.05)$	$[0.05, 0.08)$	$[0.08, \infty)$
SP500	0.001	0.003	0.034	0.054	0.348	0.441	0.092	0.019	0.002	0.005
Gold	0.008	0.023	0.112	0.137	0.202	0.214	0.116	0.154	0.022	0.007

Table 6.1: Percentages by interval S&P500 vs GOLD

Recall that the main idea for cutting the support of the data is to add specific weights on each interval. At this point we introduce our three methods for the dataset. We want to see the relation between the stocks S&P500 and GOLD via Weighted Entropy-type measures. Each from the three methods of the previous chapter will be applied and examined analytically.

6.1.1 Middle method

The first method that we will present is ”The Middle method”. The method will be applied to four Entropy-type measures defined in Chapter 4 (Definitions 16, 17, 18 and 19) and the results will be compared. The steps of the algorithm are similar as in Chapter 5.

Now, we present the graph with the four Weighted Entropy-type techniques.

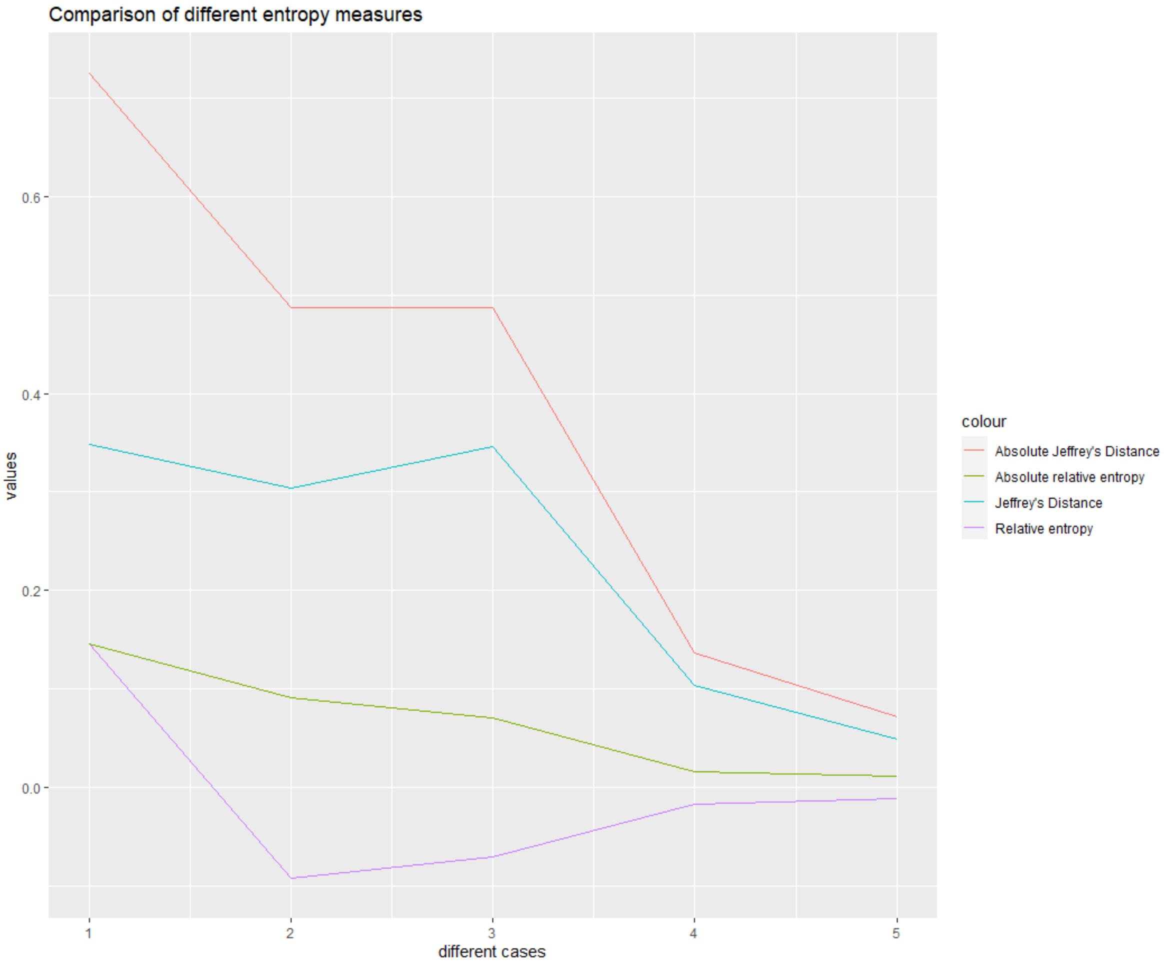

Figure 6.2: Middle method S&P500 vs GOLD

We can see that the Relative entropy takes negative values. The reason is that in such instances, the numerator of the logarithm is smaller than the denominator and when this happens the measure is negative. Secondly, note that the Absolute Relative entropy is non-negative but it takes smaller values than Jeffrey's Distance. Finally, the biggest 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 two stocks. Observe that if the Relative entropy is used the two stocks appear to be almost equidistant.

6.1.2 Left to the Right (LR) method

The second method is the ”LR method”. By using the appropriate theory in Chapter 4 and the presentation of the algorithm in Chapter 5 the results will be shown in the next graph.

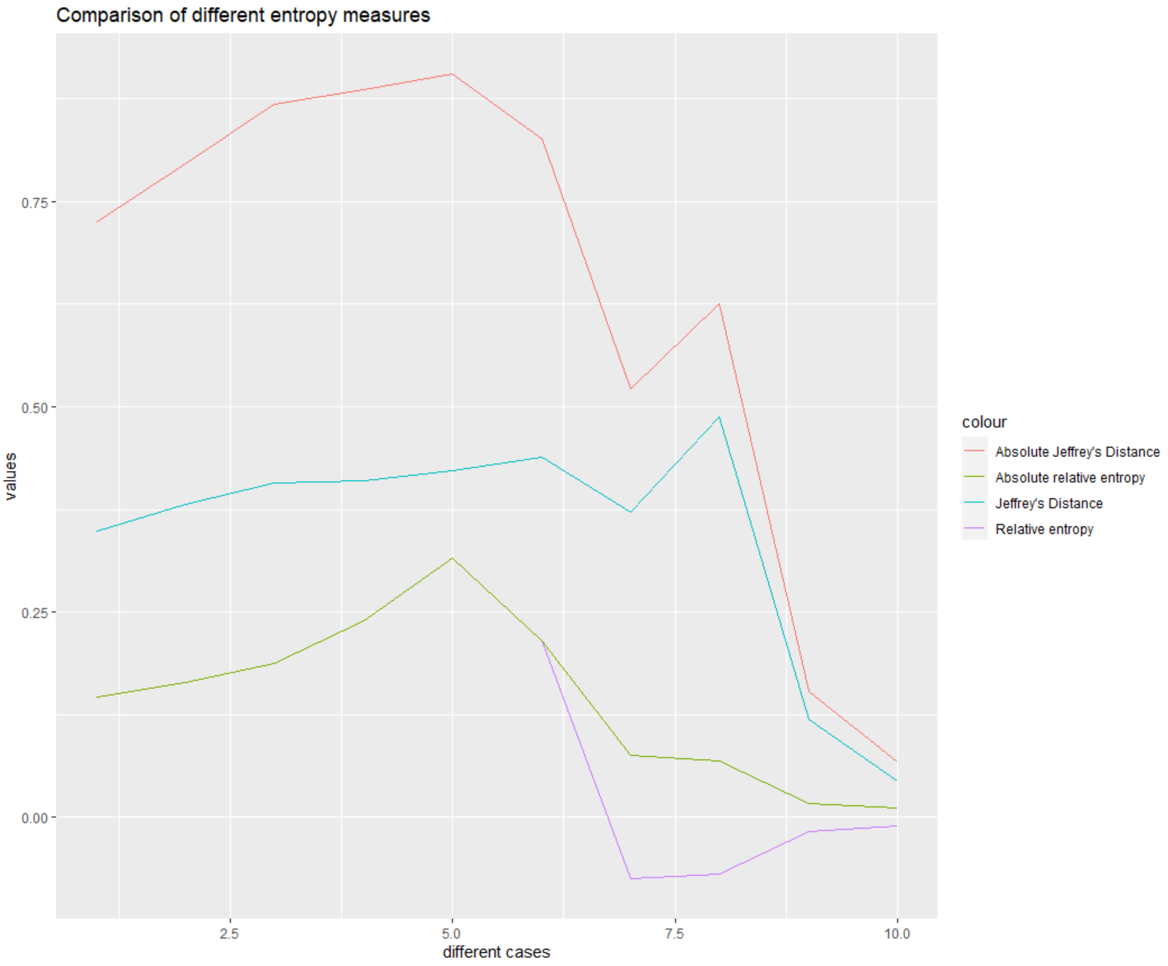

Figure 6.3: LR method S&P500 vs GOLD

We can see that, the Relative entropy takes again negative values. The reason is again that the numerator of the logarithm is smaller than the denominator and when this happens the measure is negative. Secondly, note that the Absolute Relative entropy is non-negative but it is lower on the Y-axis than the Jeffrey's Distance. Finally, the biggest differences remarked on Absolute Jeffrey's Distance method.

6.1.3 Right to the Left (RL) method

The third method that we present is called "RL method". By using the theory from Chapter 4 and the same algorithmic steps from Chapter 5 the results will be shown in the following graph.

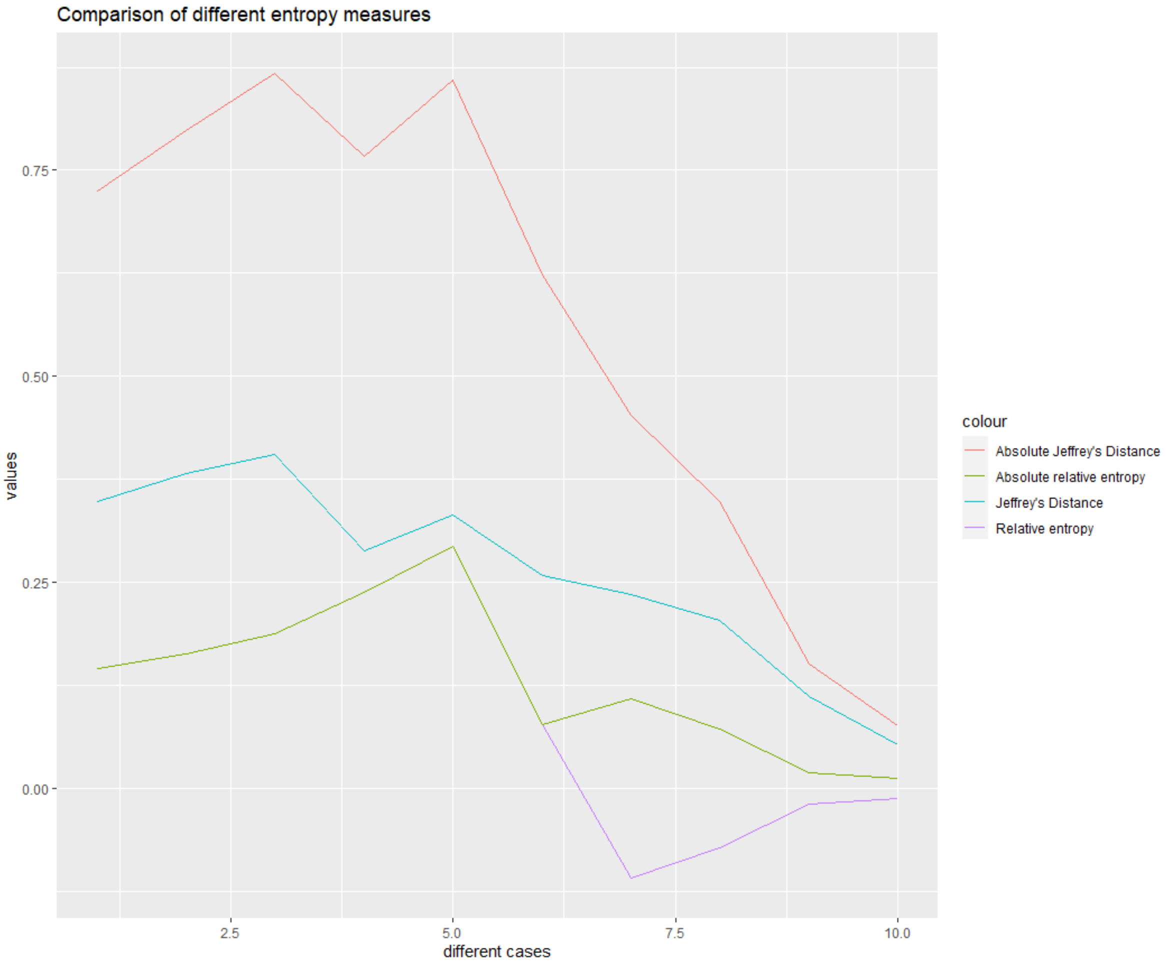

Figure 6.4: RL method S&P500 vs GOLD

We can see that, the Relative entropy takes again negative values. The reason is again that the numerator of the logarithm is smaller than the denominator and when this happens the measure is negative. Secondly, note that the Absolute Relative entropy is non-negative but it is lower than the Jeffrey's Distance.
Finally, the biggest differences are observed for the Absolute Jeffrey's Distance method.

Observe that there are only minor differences among the three proposed techniques. We can say that with three techniques when we focus on the tails the distance reduces that means the biggest dissimilarities are observed in the middle of the datasets. The results appear to be quite similar implying that there may be not significant differences when applied to symmetric distributions (like Standard Normal and t-student or two Normals as they appear in this case).

6.2 Case Study - Geosciences

For the second example, we collected data from the Institute of Geodynamics (National Observatory of Athens) *www.gein.noa.gr.* Our data concern earthquakes from 1973 to 2004 in Greece We have 5384 observations by taking the earthquakes that are below or equal 4 (≥ 4) in Richter scale.
In this part of our work, we try to see the relationship between our data and the Shifted Exponential Distribution which is a displacement of Exponential Distribution to the right by 4 units. Firstly, we will present the histogram of our data and the histogram of data with the Shifted exponential line. The histograms are the following.

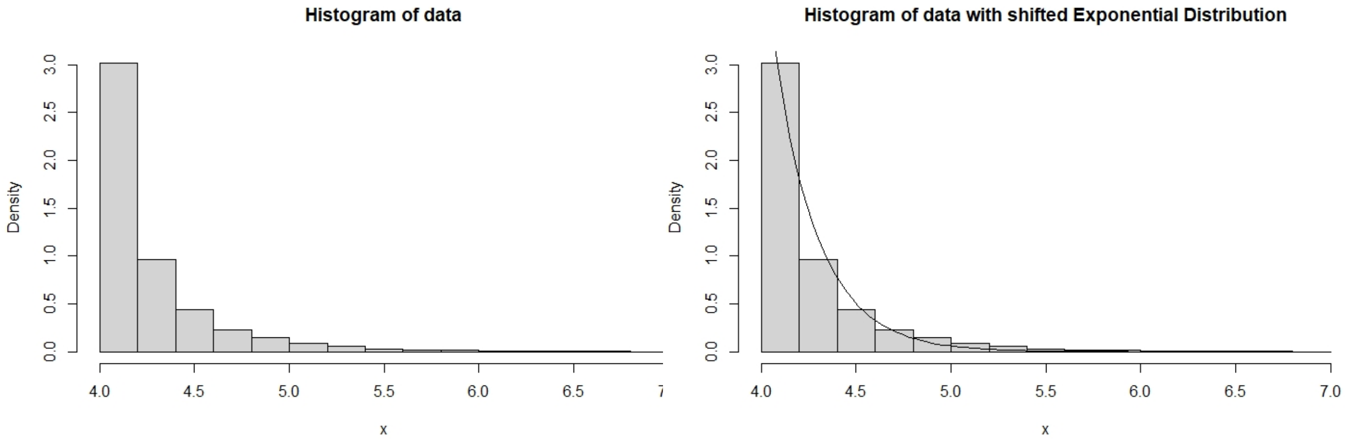

Figure 6.5: Histograms of Dataset and Shifted Exponential Distribution

The average of the data is 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 $\lambda = 4.280758$.

Now, for our method we will divide the support of the dataset with the following way:

$$[4, 4.25) \cup [4.25, 4.5) \cup [4.5, 4.75) \cup [4.75, 5) \cup [5, 5.25) \cup [5.25, 5.5) \cup [5.5, 6) \cup [6, 6.25) \cup [6.25, 6.5) \cup [6.5, 7]$$

The main idea for cutting the support of the data is to add specific weights on each interval. After this addition, we will calculate some Entropy-type techniques. The following Table provides the percentages of data in every interval for the real data and the Shifted Exponential distribution respectively.

Percentages by 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	0.602	0.242	0.062	0.051	0.017	0.012	0.008	0.001	0.0005	0.001
Shifted Exp	0.657	0.225	0.077	0.026	0.009	0.003	0.001	0.0001	0.00004	0.00001

Table 6.2: Percentages by interval Dataset vs Shifted Exponential Distribution

The three methods "Middle method", "LR method" and "RL method" have been applied in exactly the same way as in Section 6.1 and the three graphs are given below.

6.2.1 Middle method

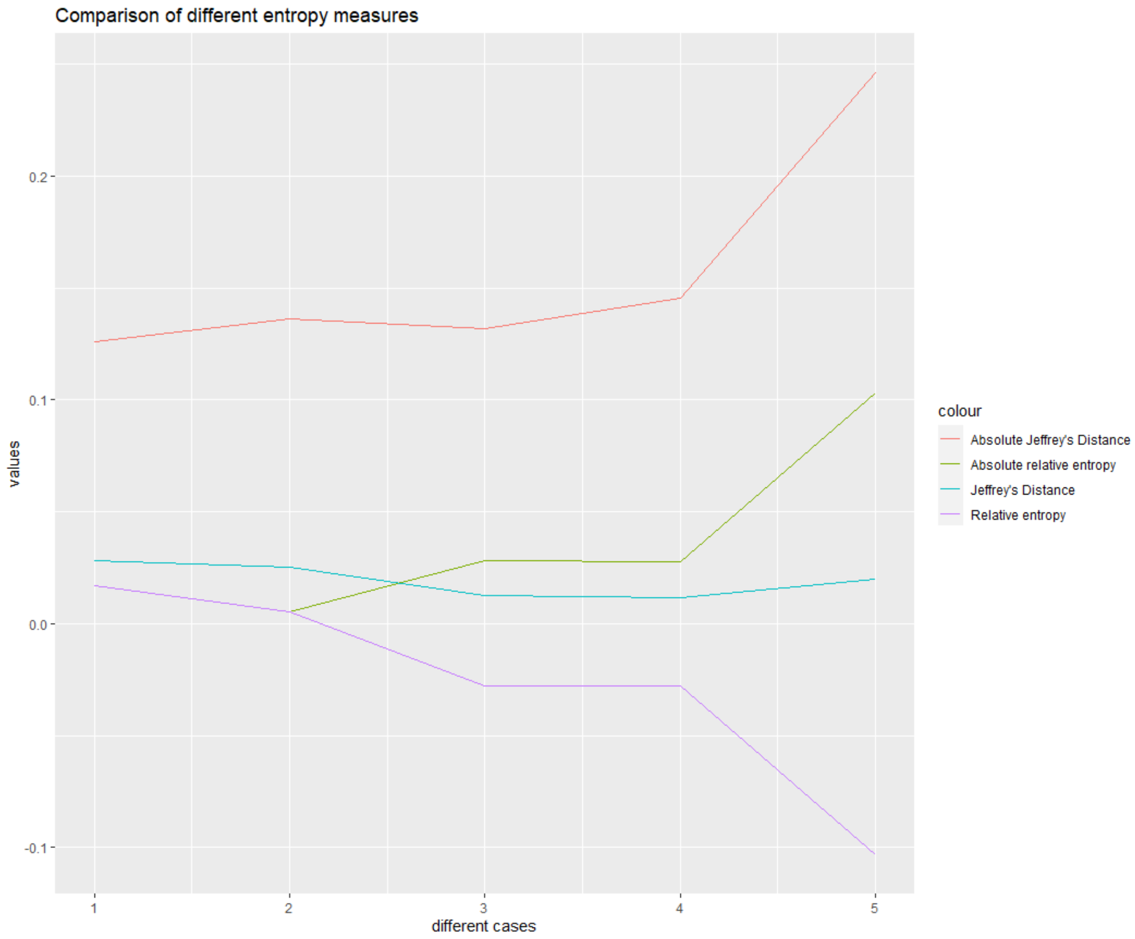

Figure 6.6: Midle method Dataset vs Shifted Exponential Distribution

The first case with Relative entropy is not useful at all. Firstly, it is not symmetric and secondly it gives negative values. The later 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. The defect stated above can be resolved if one uses the case of Absolute Jeffrey's Distance where we combine the advantages of Jeffrey's Distance and 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 where 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.

6.2.2 LR and RL method

The results of these two methods are given in graphs 6.7 and 6.8.

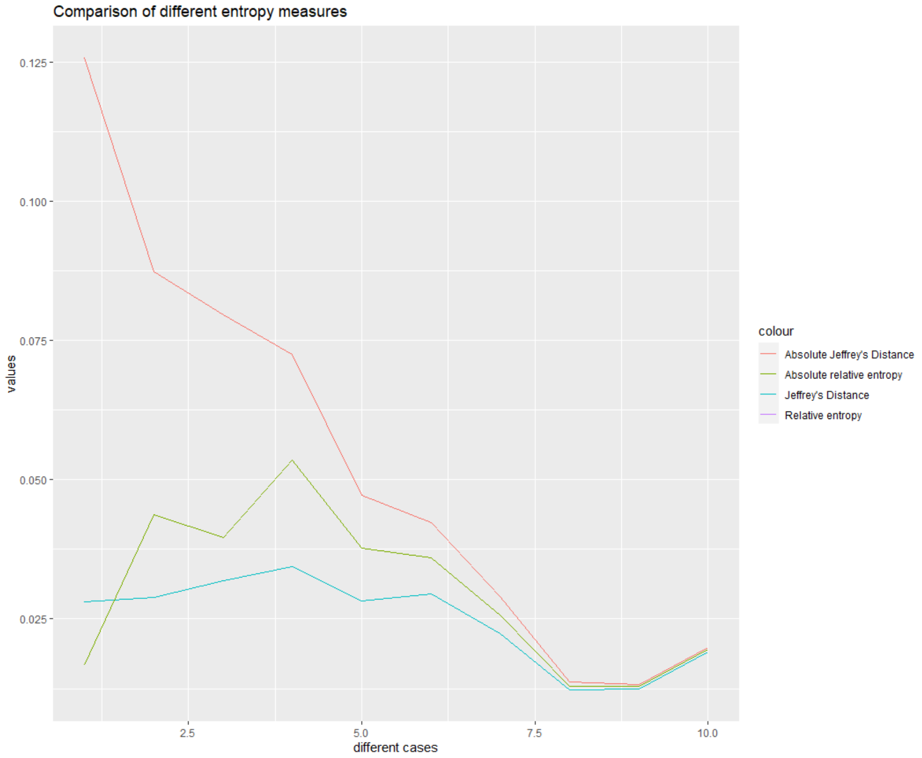

Figure 6.7: LR method Dataset vs Shifted Exponential Distribution

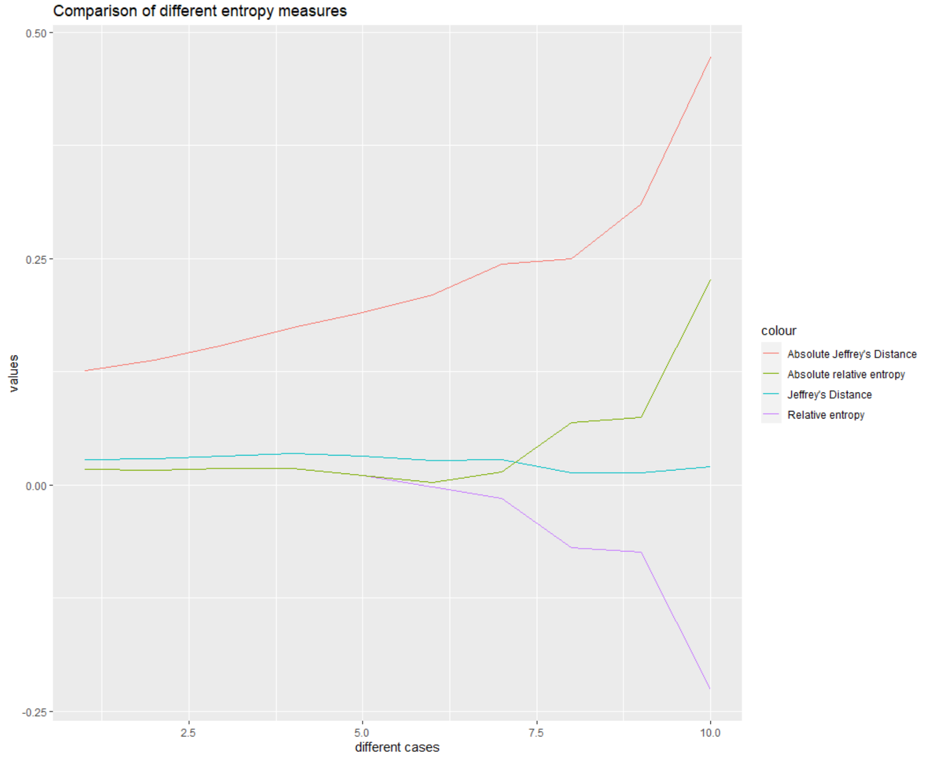

Figure 6.8: RL method Dataset vs Shifted Exponential Distribution

Observe that for the LR method, the Relative entropy and Absolute Relative Entropy are the same. In conclusion we observed that the three proposed methods behave differently with the higher values obtained for the RL method (values up to ≈ 0.5 for the Absolute Jeffrey's Distance) followed by the Middle method (values as high as ≈ 0.25 for the two Absolute measures).

Chapter 7

Conclusions

The main purpose of this thesis was the comparison between Weighted Entropy-type measures. After the presentation of the necessary theory on Divergences and Entropy, we applied all previous theoretical results in a series of simulations and in two experiments. The first experiment was to observe the relation (distance) between two stocks and the other experiment was for observing the distance and the relation between earthquakes and one fitted distribution. By introducing the Absolute Weighted Entropy-type methods we observed that the Absolute Jeffrey's Distance gives the best results (higher values) among all methods considered.

As we observed, the Weighted Relative Entropy technique is less accurate because it takes negative values which violates the main idea of distance. After this we presented the Weighted Jeffrey's Distance which is symmetric but not accurate. The two final techniques that we used are based on the previous two techniques with modifications. More specifically, we introduced the Absolute Weighted Relative Entropy (A.W.R.E) and the Absolute Weighted Jeffrey's Distance (A.W.J.D). These two techniques and especially the last one gave larger distance values between two datasets and at the same time fulfilled the properties of symmetricity and non-negativity.

In conclusion based on the simulations and 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 the 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.

Bibliography

[Abbas et al., 2017] Abbas, A. E., H Cadenbach, A., and Salimi, E. (2017). A kullback–leibler view of maximum entropy and maximum log-probability methods. *Entropy*, 19(5):232.

[Akaike, 1973] Akaike, H. (1973). Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, 60(2):255–265.

[Amigó et al., 2018] Amigó, J. M., Balogh, S. G., and Hernández, S. (2018). A brief review of generalized entropies. *Entropy*, 20(11):813.

[Aoki and Hiraide, 1994] Aoki, N. and Hiraide, K. (1994). *Topological theory of dynamical systems: recent advances*. Elsevier.

[Basu et al., 1998] Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559.

[Bhattacharyya, 1943] Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–109.

[Boltzmann, 1866] Boltzmann, L. (1866). *Über die mechanische Bedeutung des zweiten Hauptsatzes der Wärmetheorie:(vorgelegt in der Sitzung am 8. Februar 1866)*. Staatsdruckerei.

[Boltzmann, 1872] Boltzmann, L. (1872). Sitzungsberichte akad. *Wiss. Wien*, 66(275):1872.

[Clausius, 1865] 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, 1936] Fisher, R. (1936). Statistical methods for research workers (rev. & enl, pp. xiii, 339). *Edinburgh and London: Oliver & Boyd.*

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

[Guiaşu, 1971] Guiaşu, S. (1971). Weighted entropy. *Reports on Mathematical Physics*, 2(3):165–179.

[Hartley, 1928] Hartley, R. V. (1928). Transmission of information 1. *Bell System technical journal*, 7(3):535–563.

[Havrda and Charvát, 1967] Havrda, J. and Charvát, F. (1967). Quantification method of classification processes. concept of structural a-entropy. *Kybernetika*, 3(1):30–35.

[Hunter, 2011] Hunter, J. K. (2011). Measure theory. *University Lecture Notes, Department of Mathematics, University of California at Davis. http://www. math. ucdavis. edu/˜ hunter/measure_theory.*

[Jaynes, 1957] Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical review*, 106(4):620.

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

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

[Kesavan and Kapur, 1989] Kesavan, H. K. and Kapur, J. N. (1989). The generalized maximum entropy principle. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(5):1042–1052.

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

[Kullback, 1997] Kullback, S. (1997). *Information theory and statistics.* Courier Corporation.

[Kullback et al., 2013] Kullback, S., Keegel, J. C., and Kullback, J. H. (2013). *Topics in statistical information theory*, volume 42. Springer Science & Business Media.

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

[Lehmann, 2012] Lehmann, E. L. (2012). Parametric versus nonparametrics: two alternative methodologies. In *Selected works of EL Lehmann*, pages 437–445. Springer.

[Mager et al., 2004] Mager, D. E., Merritt, M. M., Kasturi, J., Witkin, L. R., Urdiqui-Macdonald, M., Sollers 3rd, J., Evans, M. K., Zonderman, A. B., Abernethy, D. R., and Thayer, J. F. (2004). Kullback-leibler clustering of continuous wavelet transform measures of heart rate variability. *Biomedical Sciences Instrumentation*, 40:337–342.

[Mahalanobis, 1936] Mahalanobis, P. C. (1936). On the generalized distance in statistics. National Institute of Science of India.

[Mantalos et al., 2010] Mantalos, P., Mattheou, K., and 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.

[Martins et al., 2004] Martins, A. M., Neto, A. D., de Melo, J. D., and Costa, J. A. F. (2004). Clustering using neural networks and kullback-leibler divergency. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)*, volume 4, pages 2813–2817. IEEE.

[Mattheou et al., 2009] Mattheou, K., Lee, S., and Karagrigoriou, A. (2009). A model selection criterion based on the bhhj measure of divergence. *Journal of Statistical Planning and Inference*, 139(2):228–235.

[Mora and Walczak, 2016] Mora, T. and Walczak, A. M. (2016). Rényi entropy, abundance distribution, and the equivalence of ensembles. *Physical Review E*, 93(5):052418.

[Murphy, 2012] Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

[Neyman and Pearson, 1933] Neyman, J. and 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, 1956] Pierce, J. F. (1956). *Raw material inventory control at the Boston Woven Hose and Rubber Company: the information system, control quantities*. PhD thesis, Massachusetts Institute of Technology. School of Industrial Management.

[Rathie and Da Silva, 2008] Rathie, P. N. and Da Silva, S. (2008). Shannon, lévy, and tsallis: a note. *Applied Mathematical Sciences*, 2(28):1359–1363.

[Rényi et al., 1961] Rényi, A. et al. (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.

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

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

[Shannon and Weaver, 1949] Shannon, C. E. and Weaver, W. (1949). *The Mathematical Theory of Communication, by CE Shannon (and Recent Contributions to the Mathematical Theory of Communication), W. Weaver*. University of Illinois Press.

[Sharifdoost et al., 2009] Sharifdoost, M., Nematollahi, N., and Pasha, E. (2009). Goodness of fit test and test of independence by entropy. *Journal of Mathematical Extension*, 3(2):43–59.

[Song, 2002] Song, K.-S. (2002). Goodness-of-fit tests based on kullback-leibler discrimination information. *IEEE Transactions on Information Theory*, 48(5):1103–1117.

[Tsallis, 1988] Tsallis, C. (1988). Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52(1):479–487.

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

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

[Wiener and BeckenBach, 1956] Wiener, N. and BeckenBach, E. (1956). The theory of prediction. *Modern Mathematics for Engineers, McGraw-Hill.*

[Yang et al., 2019] Yang, J., Grunsky, E., and 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.

[Zhou et al., 2013] Zhou, R., Cai, R., and Tong, G. (2013). Applications of entropy in finance: A review. *Entropy*, 15(11):4909–4931.