Recent Advances in Weighted Divergence Measures

Thomas Gkelsinis

Supervisor:
Prof. Karagrigoriou Alexandros
Co-Supervisors:
Prof. Konstantinides Dimitrios
Assoc. Prof. Tachtsis Eleftherios

Department of Statistics and Actuarial-Financial Mathematics
University of the Aegean
Greece
January 2020

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

Supervisors

Karagrigoriou Alexandros

Professor at Department of Statistics and Actuarial-Financial Mathematics, University of the Aegean, Greece.
email: alex.karagigoriou@aegean.gr

Konstantinides Dimitrios

Professor at Department of Statistics and Actuarial-Financial Mathematics, University of the Aegean, Greece.
email: konstant@aegean.gr

Tachtsis Eleftherios

Associate Professor at Department of Statistics and Actuarial-Financial Mathematics, University of the Aegean, Greece.
email: ltah@aegean.gr

Dedicated to Panagiotis, Anthi, Konstantina, Vasiliki, Ntinos and Rania.

Abstract

In statistics and other fields one of the main goals is the investigation and the comparison of the probabilistic behaviour of random processes. A very important tool in the 'quiver' of a researcher, for that purpose, is the concept of Divergence Measures. This type of measures quantify the dissimilarities between two random processes based only on their probabilistic behaviour.

Many times in statistics, we do not only want to emphasize in the quantitative dissimilarities but also in the qualitative ones. For example, in financial risk analysis it is common to take under consideration the existence of fat tails in the distribution of returns of an asset (especially the left tail) and in biostatistics to use robust statistical methods to trim extreme values. Motivated by these needs, through this thesis, we will present and study the concept of Weighted Divergence Measures. These measures, quantify the dissimilarities between two random processes with greater significance in specific parts (or events) of their probability distribution, so they take under consideration both their probabilistic behaviour and their qualitative characteristics.

Abstract

Στη στατιστική όπως και σε άλλες επιστήμες ένας από τους βασικούς στόχους είναι η διερεύνηση και η σύγκριση της πιθανοθεωρητικής συμπεριφοράς τυχαίων διαδικασιών. Ένα πολύ σημαντικό εργαλείο στη "φαρέτρα" του ερευνητή, για το σκοπό αυτό, είναι η έννοια των Μέτρων Απόκλισης. Αυτός ο τύπος των μέτρων ποσοτικοποιεί τις ανομοιότητες μεταξύ δύο τυχαίων διαδικασιών βασιζόμενος μόνο στη πιθανοθεωρητική συμπεριφορά τους.

Πολλές φορές στη στατιστική δεν μας ενδιαφέρει να επικεντρωθούμε μόνο στις ποσοτικές ανομοιότητες αλλά και στις ποιοτικές. Για παράδειγμα, στην χρηματοοικονομική ανάλυση ρίσκου είναι σύνηθες να μεριμνούμε για παχιές ουρές στη κατανομή των αποδόσεων ενός χρηματοοικονομικού περιουσιακού στοιχείου και στη βιοστατιστική να χρησιμοποιούμε "εύρωστες" στατιστικές μεθόδους για να περικόψουμε τις ακραίες τιμές. Εμπνευσμένοι από αυτές τις ανάγκες, στη παρούσα διπλωματική εργασία, θα παρουσιάσουμε και θα μελετήσουμε την έννοια των Σταθμισμένων Μέτρων Απόκλισης. Αυτά τα μέτρα, ποσοτικοποιούν τις ανομοιότητες μεταξύ δύο τυχαίων διαδικασιών δίνοντας μεγαλύτερη σημασία σε συγκεκριμένα υποδιαστήματα (ή γεγονότα) του στηρίγματός τους και έτσι μεριμνούν και για τη πιθανοθεωρητική συμπεριφορά αλλά και για τα ποιοτικά χαρακτηριστικά τους.

Contents

Chapter 1

Introduction

In statistics and other fields one of the most challenging aspects is to investigate the probabilistic behaviour of a random process with respect to specific events. From finance to signal processing researchers try to distinguish random processes from each other and study their behaviour. A very important tool for this distinction is the concept of Divergence Measures. This type of measures quantify the dissimilarities between random processes based only on their probabilistic behaviour. We often state that they measure the discrepancy between two probability distributions or the information needed in order to distinguish one from the other. There is plethora of estimators and hypothesis tests associated with such measures for several cases, many tests of fit defined via measures of divergence and take into account dissimilarities between the distributions involved. Also, famous model selection criteria (like Akaike Information Criterion) are based on such type of measures.

The challenge here, is to construct such measures which will take into account not only the probabilistic aspects of random processes but also the qualitative characteristics of them. These characteristics sometimes are subjective and someone would say that they are related to the significance, the relevance or the utility of the information contained which is related to a specific goal. This type of measures are called Weighted Divergence Measures. These divergences do not assume that all possible states of a random process have the same significance to a goal (like the classical ones assume), so they apply specific weights in different states or parts of these processes. For example, in financial risk analysis it is common to take under consideration the existence of heavy tails with emphasis on the left one, of the distribution associated with an asset and in biostatistics to use robust statistical methods to trim the extreme values. So, applying these weighted measures we can distinguish small dissimilarities in the probabilistic behaviour of two random processes which in other cases would be difficult to notice.

The present thesis is structured as follows. Chapter 3 is devoted to basic knowledge about measure theory, probability spaces and random variables. This background is essential to continue in the core part. In Chapter 4 we will study divergence measures, their properties, the most important families and their applications in statistical inference. In the fifth Chapter we will present the existing research on the weighted divergence measures, we will try to surpass some problems and we will try to extend this concept. In Chapter 6 there will be simulations based on the formulas presented in Chapter 5. There will be a plethora of combinations between distributions both discrete and continuous. In the last Chapter we will give a short conclusion for the above results.

Chapter 2

Literature Review

Divergence measures have their roots in the notion of entropy which is a fundamental concept in information theory. The roots of information theory can be found in the work of Shannon, a mathematician who was working in Bell labs, published his pioneer article "A mathematical theory of communication" (Shannon, 1948). In his work, Shannon proposed and studied two new concepts: the entropy and the mutual information. The entropy is a measure that quantifies the amount of uncertainty involved in the value of a random variable or the outcome of a random process. The mutual information measures the mutual dependence between two variables by quantifying the 'amount of information' (in units like "shannons" or bits) which is collected regarding one of the variable via the observation of the other. Many years earlier, from a different perspective Sir Ronald Fisher (1925) was the first who introduced the term information, known as Fisher Information, to quantify the amount of information hidden in a random variable X regarding an unknown parameter θ involved in the distribution of X. A detailed description of Information Theory has been given by Verdu (1998): "A unifying theory with profound intersections with Probability, Statistics, Computer Science, and other fields. Information Theory continues to set the stage for the development of communications, data storage and processing, and other information technologies".

Since 1948 many scientists have worked on the concept of Information Theory as well as the tools (and their usage) associated with it. The relationship between Information Theory and Statistics was established by Kullback and Leibler (1951). By extending the notion of Shannon's entropy they introduced the so called Kullback-Leibler measure of divergence which is also known as 'relative entropy'. The book "Information Theory and Statistics" (Kullback, 1959) can be viewed as the starting point of the scientific field of 'Statistical Information Theory'. At this point we have to point out that the concept of divergence measures in a more primitive form, started earlier with Mahalanobis (1936) and later Bhattacharyya (1943) but Kullback and Leibler popularized it.

The need for divergence measures lies on a plethora of subjects.In probability theory, statistics, economics and many other fields we have to identify the distribution of a random variable and the behaviour of a random process. Solutions to such issues can be provided by divergence measures. With the term divergence measure we mean a function which measures the 'distance' between two functions or in our setting between two probability distributions. Most of such measures are not metrics, from the mathematical point of view, since most of them do not fulfil the symmetry and the triangle inequality. Divergence measures establish the 'distance' between two samples or between the proposed model and the true one. There are many divergence measures in the bibliography

and some of them belong to certain families with certain characteristics. The most common families are the ϕ-divergence class of measures (Csiszar, 1964), (Morimoto, 1963), (Ali and Silvey 1966), the Cressie and Read power divergence family (Cressie and Read, 1984) and the BHHJ power divergence family (Basu et al., 1998). For a more extensive discussion on the concept of divergence measures and statistical inference see the books by Pardo (2018) and Basu et al (2011).

In statistical inference, divergence measures play a crucial role. The most common applications of them are in point estimation, hypothesis testing and goodness of fit and they are presented bellow. In point estimation it is frequently hard and occasionally impossible to obtain the Maximum Likelihood Estimator (MLE). For instance in the mixture of two normal populations with unknown proportion, mean and variance parameters, we are not able to obtain the maximum likelihood estimator in closed form. For other examples refer to Le Cam (1990). Many scientists tried to give an answer to the above estimation problem. Wolfowitz (1957) was the first who introduced the Minimum Distance Estimators, later Choi and Bulgren (1968) and MacDonald (1971) try to estimate the proportion of each known distribution by minimizing the sum of squares distance between the empirical and the theoretical distributions. For the parameter estimation, Quandt and Ramsey (1978) minimized the sum of squares distance between the empirical and the theoretical moment generating functions. Finally inferential statistics based on the minimum divergence was established and in subsequent years such estimators were proposed. In 1958 we have the first one which is the Kullback-Leibler minimum divergence estimator (an alternative to the MLE). The most important such estimators following the work of Kullback-Leibler where the minimum power divergence estimator (Cressie and Read, 1984) based on the Cressie and Read divergence and is an alternative of the MLE or the Chi-Squared estimator, the minimum ϕ-divergence estimator (Morales et al., 1995) which relies on the ϕ-divergence family and the minimum density power divergence estimator (Basu et al., 1998) which is based on the BHHJ divergence and has the invariance property.

In the domain of hypothesis testing divergence measures have a pivotal role. Traditionally for testing hypothesis like the following $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$ we use Wald test statistics, Rao test statistics and likelihood ratio tests but Kupperman (1957,1958) proposed a new ground-breaking way. He established a test statistic based on the Kullback-Leibler divergence which is asymptotically distributed as a chi-square random variable. After the pioneer work of Kupperman who "connected" divergence measure theory with hypothesis testing, several alternatives have been proposed. Later, Simon (1973) connected the likelihood ratio test statistic for hypothesis testing under the exponential distribution family with the Kullback-Leibler divergence. Nayak (1983,1985) used entropy measures to construct statistical tests for multinomial distributions. Cressie and Read (1984) proposed the family of power divergence test statistics which is based on power divergence family and includes the likelihood ratio test statistic and Pearson's chi-squared test. Sutrick (1986) used the Kullback-Leibler divergence to construct a likelihood ratio test for data from multinomial distribution and compared the asymptotic power of chi-squared test statistic and likelihood ratio test. Motivated by Kupperman, Salicru et al. (1994) introduced the ϕ-divergence test statistic, as well as its properties and asymptotic behaviour.

Following the previous if we partition the range of data in disjoint sets and test the hypothesis $H_0 : p = p^h$ about the vector of parameters of a multinomial distribution then we can construct a goodness of fit test. The most common test statistics for the previous

hypothesis are the chi-squared test and the likelihood ratio test statistics. Based on the same philosophy Cressie and Read (1984) constructed a goodness of fit test based on the power-divergence test statistics. For particular λ values in this statistic we obtain the chi-squared test statistic, the likelihood ratio test statistic, the Freeman-Tukey test statistic, the modified likelihood ratio test statistic or minimum discrimination information statistic (Gokhale and Kullback, 1978) and the Cressie-Read test statistic. A more general family of test statistics which has all the above as special cases and based on which a goodness of fit test can be constructed are the ϕ-divergence test statistics. Zografos et al. (1990) study the asymptotic distribution of the ϕ-divergence test statistic family. A more extended study in this domain can be found in Cressie and Pardo (2002).

Other applications of divergence measures are in model selection. The famous Akaike Information Criterion (Akaike, 1973) is based on the Kullback-Leibler divergence. It is an asymptotically unbiased estimator of the relative expected KL divergence. Also the Divergence Information Criterion (Mattheou et al., 2009) has the same methodology and thinking as the AIC but this is based on BHHJ divergence. Mantalos et al. (2010) proposed a modification of the DIC which is an asymptotically unbiased estimator of the expected overall discrepancy. Another model selection criterion which is based on the family of pseudodistances (Jones et al., 2001) is the Pseudodistance Information Criterion (Toma et al., 2019), this criterion has been constructed with the same thinking as the AIC.

A different concept of divergence measures is the local form of them. Avlogiaris et al. (2016) proposed the local divergence measures as a way to measure the information and study the certain characteristics in a subset of the support. They construct a method of localization for every divergence measure family. Also, they proposed the local divergence information criterion (LDiv.IC) which has the similar thinking as the AIC but it depends on the BHHJ divergence measure. In this point we want to emphasize that the concept of local divergence measures was what led us to deal with the issue to focus on a certain subset of the whole distribution without losing the whole information of the remaining part. We do that through weighted divergence measures.

The main theme of this thesis which is also an area of recent developments is the weighted form of divergence measures. The first who introduced the concept of weighting each event according to their utility to measure the information were Belis and Guiasu (1968). They claimed that the occurrence of an event has a double uncertainty: "the quantitative one which is based on the probability of the occurrence and the qualitative one which is related to it's utility for the fulfilment of the goal". Guiasu (1971) constructed the weighted form of Shannon entropy and studied the properties, the axioms and the maximum value (according to the maximum information principle) of the weighted entropy.

After that, Guiasu (1986) through weighted entropy proposed a clustering to unequal subsets of the ordered dataset based on the information balance produced by the sum of information and degree of homogeneity. More specifically we have a trade off between the information that is lost due to the partition and the increased data homogeneity. The amount of information contained in the initial dataset is decreased to the amount of information contained in the classes of the partition due to the lack of distinction is made between the observations of the same class. This method is based on the concept of weighted entropy and it is directly proportional to the importance one wishes to place on specific regions of the domain.

Later, Kapur (1994) set the rules that a weighted divergence measure must have. He

proposed the 'corrected' Kullback-Leibler divergence measure, which is positive everywhere in the support and the weighted form of it. He also gave some other weighted divergence measures. Taneja (1998) proposed the weighted generalizations of J-divergence, Jensen difference divergence and the arithmetic and geometric divergence measures. Pak and Basu (1998) introduced the minimum disparity estimator in linear regression models which is based on weighted Hellinger distance between a weighted kernel density estimator of the errors and a smoothed model density of errors. Also, they have shown that if the weights chosen appropriately then the estimators would be asymptotically normally distributed and sufficient. Barbu et al. (2018) gave the weighted form of the generalizations of Alpha divergence measures and Beta divergence measures for Markov chains. They also studied their asymptotic behaviour.

Chapter 3

Basic Background

3.1 Measure Theory

We will start with a short introduction of basic concepts of measure theory needed for the mathematical foundation of divergence measures. This introduction of measure theory is not complete and its only scratches the surface. For the more in depth and complete introduction to measure theory see 'Measure Theory' by John K. Hunter (2011) on whose notes we where heavily based or 'Measure Theory' by Halmos (2013).

Some of the concepts that could be useful in this thesis are those of a 'topological space', 'σ-Algebra', 'Borel σ-Algebra', 'measurable space', 'measure', 'measure space', 'measurable function' etc. The proper definitions of all the above notions are provided in the Appendix A1 for the convenience of the reader.

3.2 Probability Spaces

We shall define the probability space $(\Omega, \mathcal{F}, P)$ using the terminology of measure theory. The sample space Ω is a set of all possible outcomes $\omega \in \Omega$ of some random experiment. Probabilities are assigned by a probability measure P: $A \longmapsto P(A)$ where $A \subseteq \mathcal{F}$ of all possible sets of outcomes. The event space $\mathcal{F}$ represents both the amount of information available as a result of the experiment conducted and the collection of all subsets of possible interest to us, where we denote elements of $\mathcal{F}$ as events. A pleasant mathematical framework results by imposing on $\mathcal{F}$ the structural conditions of a σ-algebra.

Some of the notions that could be used here are 'probability measure', 'random variable', 'distribution', 'probability density function', 'probability mass function' etc. The proper definitions of all the above notions are provided in the Appendix A2 for the convenience of the reader.

3.3 Classification of Random Variables

Regarding the classification of random variables some of the useful notions are the 'absolute continuity', 'singular continuity', 'discrete' and 'continuous measures' etc. The proper definitions of all the above notions are provided in the Appendix A3 for the convenience of the reader.

Chapter 4

Divergence Measures

Distance, as a general notion of closeness, plays an important role in statistics and sets the fundamentals for statistical inference. In order to determine the structural form of the data we use the position of the values of a population and their distance from a reference point, often the mean. For this reason there are a plethora of functions to determine the 'distance', its one with its own properties. Still they have to fulfil some conditions. In the following, we will give the definition of distance (or metric).

Definition 1. *(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. $d(x,y) \leq d(x,z) + d(z,y) \;\; \forall x, y, z \in X$

Someone would think that if we measure the mean distance between the values of a population and a reference point we will take the diversity within the population around this point. If this point is the mean of population and as a distance we take the square difference then we have the variance of the population. The diversity from this point of view is closely related to the variation of the population.

A different concept related to the diversity within a population is the entropy. Entropy is connected with uncertainty through the probability of each event to occur. Shannon (1948) proposed the information entropy to measure the information that a stochastic source contains. The following definition gives the entropic formula proposed by Shannon.

Definition 2. *(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

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

The maximum value of entropy is taken if all possible values in the support are equiprobable (or the population is discretely uniformly distributed) and the least value if one has a probability close to one. Also, we have that the diversity in the first case is bigger than the second case (like entropies). From this point of view we can express the diversity within a population as the uncertainty of the outcome of a sampling process.

From the concept of diversity within the population arises the following question. If we can determine the diversity within population, then can we determine the dissimilarities between different populations through the same concept?

Pearson (1900) was the first who tried to answer this question by measuring the dissimilarities between two statistical populations. A completely different approach, also one of the most famous, was given by Kolmogorov (1933) who measure the dissimilarities using a statistical metric. He proposed the distance called 'Kolmogorov metric' to compare two cumulative distributions as follows.

Definition 3. *(Kolmogorov metric)*

Consider a space S of all one dimensional distributions, d_K is the Kolmogorov metric between cumulative distributions $P, Q \in S$ and is defined as

$$d_K(P, Q) = \sup_{x \in \mathbb{R}} |P(x) - Q(x)|$$

Based on this metric Smirnov (1948) construct the famous Kolmogorov-Smirnov test of goodness of fit. It was the first time that a statistical distance used to construct this type of non-parametric tests to stand out two distributions.

Slightly later, Mahalanobis (1936) from a completely different point of view had the initial idea to set the distance between two distributions by taking into account the variance-covariance matrix of the data. But, a disadvantage of this method is that for complex variance-covariance matrix it is difficult to find the inverse form.

For many years after the pioneer works by Kolmogorov and Mahalanobis mathematicians didn't get involved with the concept of measure the similarity between two statistical populations until the cornerstone work by Kullback and Leibler (1951). They were motivated by the concept of measuring the amount of information needed to discriminate one distribution from an other. For this reason they use entropy, as proposed by Shannon, to set a measure that is not a metric, from the mathematical point of view,

but it measures the similarity between two statistical populations. With this work, the concept of measuring the discrepancy between two distributions has come to the front. These type of measures are called divergence measures.

Definition 4. *(Divergence Measure)*

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

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

As we can see divergence measures are not metrics because they does not have to be symmetric and fulfil the triangular inequality.

After the work by Kullback and Leibler many researchers proposed different divergences with different properties and usage. They used them as a tool for statistical inference. The main difference between them is focused in the functional form of divergence. Different divergences have been proposed for different reasons, such as robust parameter estimation, multinomial goodness of fit or model selection criteria. In 4.1 we provide the continuity of divergence measures (based on Kullback-Leibler divergence). Then, in 4.2 we present the most important divergence families and we focus in ϕ-divergence and it's properties. Finally, in 4.3 we study the main concepts in statistical inference based on divergence measures.

Before we proceed further we have to define some notations. We have a random variable X which takes values on a space $\mathcal{X}$ and F_X is the distribution function of X. With $\underset{\sim}{\theta} \in \Theta \subset \mathbb{R}^k, \ k \geq 1$ we denote the vector of unknown parameters. Let $(\mathcal{X}, \mathcal{F}, P_\theta)_{\theta \in \Theta}$ be the probability space associated with X. With $\mathcal{F}$ we denote the σ-field of Borel subsets $B \in \mathcal{X}$ and $\{P_\theta\}_{\theta \in \Theta}$ a family of probability distributions defined on the measurable space $(\mathcal{X}, \mathcal{F})$ with Θ an open subset of $\mathbb{R}^k, \ k \geq 1$. The support of probability distribution P_θ is denoted by S_X. We assume a σ-finite measure μ on $(\mathcal{X}, \mathcal{F})$, this measure can be the Lebesgue measure or a counting measure. Following we will make a distinction between probability density ($f_\theta(x)$) and mass ($p_\theta(x)$) function according to the measure μ:

$$\frac{dP_\theta}{d\mu}(x) = \begin{cases} f_\theta(x), \ \ x \in S_X, & \text{if } \mu \text{ is the Lebesgue measure} \\ P_\theta(X = x) = p_\theta(x), \ \ x \in S_X, & \text{if } \mu \text{ is a counting measure} \end{cases}$$

Definition 4 of divergence measures can be expressed in several ways according to the problem. For example, if we want to emphasize in the divergence between two distributions $P_{\theta_1}, P_{\theta_2}$ related to the parameters θ_1, θ_2, then the formula of divergence measure would be:

$$D(P_{\theta_1}, P_{\theta_2}) \equiv D(\theta_1, \theta_2)$$

As we can see this formula computes the divergence between two distributions from the same family but with different parameters. This form of divergence is usually used in statistical inference via divergence measures, so we will extensively use it in the following.

4.1 Continuity and Properties of Divergence Measures

One of the most important issues we have to mention before providing properties, types and applications of divergence measures is the concept of continuity. In reality, if P and Q are complex, it is often difficult to compute directly $D(P,Q)$. Instead of this, we discretized and compute them numerically. We want continuity of divergence to guarantee that this procedure converges to a real value. Following, Theorem 1 shows that divergence on a general support (e.g. infinite) can be defined as a divergence on finite support, after discretization. As the partition becomes thinner the descretized divergence approaches the true value of divergence.

Theorem 1. *(Gelfand-Yaglom-Perez, Pinsker (1960))*

Let P, Q be two probability measures on S_X with σ-algebra $\mathcal{F}$. Then

$$D(P,Q) = \sup_{\{E_1,\dots,E_n\}} \sum_{i=1}^{n} P[E_i] \log \frac{P[E_i]}{Q[E_i]}$$

where the supremum is over all finite $\mathcal{F}$-measurable partitions: $\bigcup_{j=1}^{n} E_j = S_X, E_j \cap E_i = \emptyset$, and $0 \log \frac{1}{0} = 0$ and $\log \frac{1}{0} = \infty$.

The proof of divergence continuity in discrete random variables is easier than the proof in continuous case. The proposition bellow provide the continuity in discrete case of Kullback-Leibler divergence, a particular and very important measure that in the following sections we will study in depth.

Definition 5. *(Kullback-Leibler Divergence)*

Consider two distributions P, Q with probability mass functions $\underset{\sim}{p} = (p_1, ..., p_n)'$ and $\underset{\sim}{q} = (q_1, ..., q_n)'$ respectively. Then the discrete version of Kullback-Leibler divergence (or relative entropy) is the following:

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

Proposition 1. *(Continuity of discrete Kullback-Leibler divergence measure)*

Let S_X be finite. Fix a distribution Q on S_X with $Q(x) > 0 \; \forall x \in S_X$. Then the map

$$P \mapsto D(P,Q)$$

is continuous, where P is a distribution.

Proof. The discrete Kullback-Leibler divergence is

$$D(P,Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

Each term is a continuous function of $P(x)$. □

Following, the next task is to work in general supports to study the continuity properties. As we have already seen divergence measures depend on the σ-algebra $\mathcal{F}$ of the space of probability measures. So, this dependence will be clearly denoted by $D(P_{\mathcal{F}}, Q_{\mathcal{F}})$.

Note also that divergence is continuous under monotone limits, which is verified by the following finding.

Finding 1. *(Theoretic properties of divergence measures)*

Let P, Q be probability measures on the measurable space $(\mathcal{X}, \mathcal{H})$. Assume all algebras below are sub-algebras of $\mathcal{H}$, $\mathcal{F}, \mathcal{G} \subset \mathcal{H}$. Then:

1. *If $\mathcal{F} \subseteq \mathcal{G}$ then*

$$D(P_{\mathcal{F}}, Q_{\mathcal{F}}) \leq D(P_{\mathcal{G}}, Q_{\mathcal{G}})$$

2. *Let $\mathcal{F}_1 \subseteq \mathcal{F}_2 ...$ be an increasing sequence of algebras and let $\mathcal{F} = \bigcup_n \mathcal{F}_n$ be their limit, then*

$$D(P_{\mathcal{F}_n}, Q_{\mathcal{F}_n}) \nearrow D(P_{\mathcal{F}}, Q_{\mathcal{F}})$$

   where $\nearrow$ defines the continuity from bellow.

3. *If $(P+Q)$-dense in $\mathcal{G}$ then*

$$D(P_{\mathcal{F}}, Q_{\mathcal{F}}) = D(P_{\mathcal{G}}, Q_{\mathcal{G}})$$

   Note that $\mathcal{F}$ is μ-dense in $\mathcal{G}$ if $\forall E \in \mathcal{G}, \epsilon > 0 \; \exists E' \in \mathcal{F}$ such that $\mu[E \Delta E'] \leq \epsilon$.

4. *Let $\mathcal{F}_1 \subseteq \mathcal{F}_2 ...$ be an increasing sequence of algebras and let $\mathcal{F} = \vee_n \mathcal{F}_n$ be the σ-algebra generated by them, then*

$$D(P_{\mathcal{F}_n}, Q_{\mathcal{F}_n}) \nearrow D(P_{\mathcal{F}}, Q_{\mathcal{F}})$$

5. *If $P_n \to P$ and $Q_n \to Q$ pointwise on the algebra $\mathcal{F}$, then*

$$D(P_{\mathcal{F}}, Q_{\mathcal{F}}) \leq lim_{n\to\infty} D(P_{n,\mathcal{F}}, Q_{n,\mathcal{F}})$$

   Where $P_n \to P$ pointwise on some algebra $\mathcal{F}$ is $\forall E \in \mathcal{F} : P_n[E] \to P[E]$.

So if:

$$\mathcal{F}_n \nearrow \mathcal{F} \Rightarrow D(P_{\mathcal{F}_n}, Q_{\mathcal{F}_n}) \nearrow D(P_{\mathcal{F}}, Q_{\mathcal{F}}) \tag{4.1}$$

$$\mathcal{F}_n \searrow \mathcal{F} \Rightarrow D(P_{\mathcal{F}_n}, Q_{\mathcal{F}_n}) \searrow D(P_{\mathcal{F}}, Q_{\mathcal{F}}) \tag{4.2}$$

We have to note that if $\mathcal{F}$ is not a σ-algebra and P, Q are not σ-additive then Radon-Nikodym theorem is not applicable so the original definition of divergence does not exist.

Finally we will prove the continuity of decreasing σ-algebra as in (4.2).

Proposition 2. *Let $\mathcal{F}_n \searrow \mathcal{F}$ be a sequence of decreasing σ-algebras and P, Q two probability measures on $\mathcal{F}_0$. If $D(P_{\mathcal{F}_0}, Q_{\mathcal{F}_0}) < \infty$ then we have*

$$D(P_{\mathcal{F}_n}, Q_{\mathcal{F}_n}) \searrow D(P_{\mathcal{F}}, Q_{\mathcal{F}})$$

The condition $D(P_{\mathcal{F}_0}, Q_{\mathcal{F}_0}) < \infty$ must be fulfilled.

Proof. (Polyanskiy-Wu, 2014)

Let $X_{-n} = \frac{dP}{dQ}|_{\mathcal{F}_n}$. Since $X_{-n} = E\left[\frac{dP}{dQ}|\mathcal{F}_n\right]$, we have that $(..., X_{-1}, X_0)$ is a uniformly integrable martingale. By the martingale convergence theorem in reversed time we have almost surely

$$X_{-n} \to X_{-\infty} \triangleq \frac{dP}{dQ}\,|_{\mathcal{F}_n} \tag{4.3}$$

We need to prove that

$$E_Q[X_{-n} \log X_{-n}] \to E_Q[X_{-\infty} \log X_{-\infty}].$$

We will do so by decomposing $x \log x$ as follows

$$x \log x = x \log^+ x + x \log^- x,$$

where $\log^+ x = max(\log x, 0)$ and $\log^- x = min(\log x, 0)$. Since $x \log^- x$ is bounded, the bounded convergence theorem ensures that:

$$E_Q[X_{-n} \log^- X_{-n}] \to E_Q[X_{-\infty} \log^- X_{-\infty}].$$

To prove a similar convergence for $\log^+$ we need to notice first that the function $x \mapsto x \log^+ x$ is convex. Furthermore, for any non-negative convex function ϕ such that $E[\phi(X_0)] < \infty$ the collection $\{Z_n = \phi(E[X_0|\mathcal{F}_n]), n \geq 0\}$ is uniformly integrable. Indeed, we have from Jensen's inequality

$$P[Z_n > c] \leq \frac{1}{c}\phi(E[X_0|\mathcal{F}_n]) \leq \frac{\phi(E[X_0|\mathcal{F}_n])}{c}$$

and thus $P[Z_{n>c}] \to 0$ as $c \to \infty$. Therefore, we have again by Jensen's inequality,

$$E[Z_n 1_{\{Z_n > c\}}] \leq E[\phi(X_0) 1_{\{Z_n > c\}}] \to 0, \quad c \to \infty.$$

Finally, since $X_{-n} \log^+ X_{-n}$ is uniformly integrable, we have from (4.3)

$$E_Q[X_{-n} \log^+ X_{-n}] \to E_Q[X_{-\infty} \log^+ X_{-\infty}]$$

and this concludes the proof. □

4.2 Divergence Measures and their Families

Divergence measures is a general label covering the general concept of measuring the similarity of two random processes. In this section we will study some of the most popular divergence families and some of their most important properties. Different types with different functional form have been introduced over the years and they are available in the literature.

4.2.1 ϕ-Divergence Family and its Properties

One of the most important divergence families is the ϕ-divergence family of measures proposed separately by Csiszar (1963), Morimoto (1963) and Ali-Silvey (1966). The main family is characterized by the ϕ-function which plays a key role while its properties are provided in the definition below.

Definition 6. *(ϕ-divergence measure)*

The ϕ-divergence measure between two continuous probability distributions $P_{\theta_1}, P_{\theta_2}$ with $\theta_1, \theta_2 \in \Theta$ associated with densities $f_{\theta_1}, f_{\theta_2}$ is defined by

$$D_\phi(P_{\theta_1}, P_{\theta_2}) = D_\phi(\theta_1, \theta_2) = \int_{\mathcal{X}} f_{\theta_2}(x)\phi\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right) d\mu(x) = E_{\theta_2}\left[\phi\left(\frac{f_{\theta_1}(X)}{f_{\theta_2}(X)}\right)\right],\ \phi \in \Phi^* \tag{4.4}$$

where Φ^* is the class of all convex functions $\phi(x)$, $x \geq 0$ such that at $x = 1,\ \phi(1) = 0$, at $x = 0,\ 0\phi(\frac{0}{0}) = 0$ and $0\phi(\frac{p}{0}) = lim_{u\to\infty}\frac{\phi(u)}{u}$.

As mentioned above the convex ϕ-function plays an important role as in the case of the increasing rate of divergence as the two distributions move apart, which can be expressed by the second derivative of the ϕ-function. This concept will be studied further in subsequent chapters.

Remark 1. *(Pardo, 2018)*

Suppose $\phi \in \Phi^*$ be differentiable at $x = 1$, then the fuction

$$\psi(x) \equiv \phi(x) - \phi^{'}(1)(x - 1)$$

also belongs to Φ^* and has the additional property that $\psi^{'}(1) = 0$. This property plus convexity, implies that $\psi(x) \geq 0,\ for\ any\ x \geq 0$.

Proof. Observe that

$$\psi^{'}(x) = \phi^{'}(x) - \phi^{'}(1) \Rightarrow \psi^{'}(1) = 0$$

and

$$\psi^{''}(x) = \phi^{''}(x) \geq 0$$

We prove that ψ function is convex.

So, from Jensen inequality we have: $E\left[\psi(X)\right] > \psi\left[E(X)\right]$. For $x = \dfrac{f(x)}{g(x)}$, where f, g are two densities, we have $x \in \mathbb{R}_+$, so $E(x) \in \mathbb{R}_+$. Since ψ has minimum at 1 we have that $\psi\left[E(X)\right] \geq \psi(1) = 0 \Rightarrow E\left[\psi(X)\right] \geq 0$. We prove that ψ is positive on average $\forall x \in \mathbb{R}_+$.

Finally, note that $\psi \geq 0,\ \forall\, x \in \mathbb{R}_+$.

□

Further, we will prove that $D_\psi(\theta_1, \theta_2) = D_\phi(\theta_1, \theta_2)$

$$D_\psi(\theta_1, \theta_2) = \int_X f_{\theta_2}(x) \left(\phi\left(\frac{f_{\theta_1}(x)}{\theta_2(x)} \right) - \phi^{'}(1) \left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} - 1 \right) \right) d\mu(x) =$$

$$= \int_X f_{\theta_2}(x) \phi\left(\frac{f_{\theta_1}(x)}{\theta_2(x)} \right) d\mu(x) - \int_X f_{\theta_2}(x) \phi^{'}(1) \left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} - 1 \right) d\mu(x) =$$

$$= \int_X f_{\theta_2}(x) \phi\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} \right) d\mu(x) = D_\phi(\theta_1, \theta_2)$$

Since the two divergence measures coincide, we can consider the set Φ^* to be equivalent to the set:

$$\Phi \equiv \Phi^* \cap \left\{ \psi : \psi^{'}(1) = 0 \right\}$$

For example observe that Kullback-Leibler divergence is obtained for $\phi(x) = x \log(x)$ or equivalently $\psi(x) = x \log(x) - x + 1$.

In the following table (Pardo, 2018) we present some of the most important divergence measures which are particular cases of the ϕ-divergence family (according to ψ-function):

$\psi-$function	Divergence
$x\log(x)-x+1$	Kullback-Leibler (1959)
$-\log(x)+x-1$	Minimum Discrimination Information
$(x-1)\log(x)$	J - Divergence
$\frac{1}{2}(x-1)^2$	Pearson (1900), Kagan (1963)
$\frac{(x-1)^2}{(x+1)^2}$	Balakrishnan and Sanghvi (1968)
$\frac{-x^s+s(x-1)+1}{1-s},\ s\neq 1$	Rathie and Kannappan (1972)
$\frac{1-x}{2}-(\frac{1+x^-r}{2})^{\frac{-1}{r}},\ r>0$	Harmonic Mean (Mathai and Rathie (1975))
$\frac{(1-x)^2}{2(\alpha+(1-\alpha)x)},\ 0\leq\alpha\leq 1$	Rukhin (1994)
$\frac{\alpha x\log(x)-(\alpha x+1-\alpha)\log(\alpha x+1-\alpha)}{\alpha(1-\alpha)},\ \alpha\neq 0,1$	Lin (1991)
$\frac{x^{\lambda+1}-x-\lambda(x-1)}{\lambda(\lambda+1)},\ \lambda\neq 0,-1$	Cressie and Read (1984)
$\lvert 1-x^{\alpha}\rvert^{\frac{1}{\alpha}},\ 0<\alpha<1$	Matusita (1964)
$\lvert 1-x\rvert^{\alpha},\ \alpha\geq 1$	$\begin{cases}\chi-\text{divergence of order }\alpha\text{ (Vajda, 1973)}\\ \text{Total Variation if }\alpha=1\text{ (Saks, 1937)}\end{cases}$

In the following propositions we will present some of the most important properties of the ϕ-divergence family. Proposition 3 gives an upper bound of the divergence and the increasing property when two distributions move apart. This is one of the most important properties because sets the fundamentals.Proposition 4 stress that any transformation of the data decreases the divergence except that if the transformation is sufficient with respect to the probability distributions. Lastly, Proposition 5 provides the layout property of divergence with respect to the arguments (parameters). These propositions and their proofs are based on Vajda (1989), where the interested reader could referred to for an in depth study.

Proposition 3. *Suppose P_{θ_1} and P_{θ_2} be two probability distributions and let $\phi\in\Phi^*$ be differentiable at $t=1$. Then*

$$0\leq D_{\phi}(\theta_1,\theta_2)\leq\phi(0)+\lim_{r\to\infty}\frac{\phi(r)}{r}$$

where

$$D_{\phi}(\theta_1,\theta_2)=0\ \ if\ \ P_{\theta_1}=P_{\theta_2} \tag{4.5}$$

and

$$D_\phi(\theta_1, \theta_2) = \phi(0) + \lim_{r\to\infty} \frac{\phi(r)}{r} \quad if \quad S_1 \cap S_2 = \emptyset. \tag{4.6}$$

If ϕ is also strictly convex at $t = 1$, then (4.5) holds iff $P_{\theta_1} = P_{\theta_2}$. If moreover,

$$\phi(0) + \lim_{r\to\infty} \frac{\phi(r)}{r} < \infty$$

then (4.6) holds iff $S_1 \cap S_2 = \emptyset$, where S_i, $i = 1, 2$, is the support of the probability distribution P_{θ_i}, $i = 1, 2$.

The proof can be found in Appendix B1.

Remark 2. *Proposition 3 consists of two consequences and when they are equivalences. Mainly:*

We have that if $P_{\theta_1} = P_{\theta_2}$ $(S_1 = S_2) \Rightarrow D_\phi(\theta_1, \theta_2) = 0$. Also if ϕ is strictly convex for $x = 1$ then we have $P_{\theta_1} = P_{\theta_2} \Leftrightarrow D_\phi(\theta_1, \theta_2) = 0$.

We have that if $S_1 \cap S_2 = \emptyset \Rightarrow D_\phi(\theta_1, \theta_2) = \phi(0) + \lim_{r\to\infty} \frac{\phi(r)}{r}$. If moreover $\phi(0) + \lim_{r\to\infty} \frac{\phi(r)}{r} < +\infty$ then we have $P_{\theta_1} = P_{\theta_2} \Leftrightarrow D_\phi(\theta_1, \theta_2) = 0$.

Let $X_1, ..., X_n$ be a sample from P_θ, $\theta \in \Theta$. For μ being the Lebesgue measure or a counting measure, let $f_\theta(x) = \frac{dP_\theta}{d\mu}(x)$ where $x = (x_1, ..., x_n)$. Suppose that T is a measurable transformation from $(\mathcal{X}, \mathcal{F}_\mathcal{X})$ onto a measurable space $(\mathcal{Y}, \mathcal{F}_\mathcal{Y})$. We denote

$$Q_{\theta_i}(A) = P_{\theta_i}(T^{-1}(A)), \quad i = 1, 2 \tag{4.7}$$

with $A \in \mathcal{F}_\mathcal{Y}$ and

$$g_{\theta_i}(t) = \frac{dQ_{\theta_i}}{d\mu}(t), \; f_{\theta_i}(\frac{x}{t}) = \frac{dP_{\theta_i}}{dQ_{\theta_i}}, \; = 1, 2 \tag{4.8}$$

with t denoting the values of T. In this context we have the following property.

Proposition 4. *Assume that $\theta_1, \theta_2 \in \Theta \subset \mathbb{R}$. Let $\phi \in \Phi^*$ and Q_{θ_i}, P_{θ_i}, $i = 1, 2$, be probability measures defined by (4.7) and (4.8). Then we have*

$$D_\phi(Q_{\theta_1}, Q_{\theta_2}) \leq D_\phi(P_{\theta_1}, P_{\theta_2})$$

The equality holds if T is sufficient for the probability distributions P_{θ_1} and P_{θ_2}.

Proof. (Vajda, 1995)

We have

$$D_\phi(P_{\theta_1}, P_{\theta_2}) = \int_X f_{\theta_2}(x)\phi\left(\frac{f_{\theta_1}}{f_{\theta_2}}\right) d\mu(x) =$$

$$= \int_X \int_Y f_{\theta_2}(x/t) g_{\theta_2}(t) \phi\left(\frac{f_{\theta_1}}{f_{\theta_2}}\right) d\mu(t) d\mu(x) =$$

$$= \int_Y g_{\theta_2}(t) \left(\int_X f_{\theta_2}(x/t) \phi\left(\frac{f_{\theta_1}}{f_{\theta_2}}\right) d\mu(x) \right) d\mu(t).$$

Applying Jensen's inequality we obtain

$$D_\phi(P_{\theta_1}, P_{\theta_2}) \geq \int_Y g_{\theta_2}(t) \left(\phi \left(\int_X f_{\theta_2}(x/t) \frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} d\mu(x) \right) \right) d\mu(t).$$

But

$$\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} = \frac{\frac{dP_{\theta_1}}{d\mu}}{\frac{dP_{\theta_2}}{d\mu}} = \frac{g_{\theta_1}(t) f_{\theta_1}(\frac{x}{t})}{g_{\theta_2}(t) f_{\theta_2}(\frac{x}{t})} \tag{4.9}$$

then,

$$D_\phi(P_{\theta_1}, P_{\theta_2}) \geq \int_Y g_{\theta_2}(t) \phi\left(\frac{g_{\theta_1}(t)}{g_{\theta_2}(t)}\right) d\mu(t) = D_\phi(Q_{\theta_1}, Q_{\theta_2}).$$

If ϕ is strictly convex, the equality holds iff

$$\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} = \int_X f_{\theta_2}(\frac{x}{t}) \frac{f_{\theta_1(x)}}{f_{\theta_2(x)}} d\mu(x), \quad \forall x.$$

The second term in the previous inequality is equal to $\dfrac{g_{\theta_1}(t)}{g_{\theta_2}(t)}$ by (4.9). Then, using the Factorization Theorem, the equality holds if T is sufficient for the probability distributions P_{θ_1} and P_{θ_2}. □

In the following proposition $\{P_\theta\}_{\theta\in\Theta}$, $\Theta \subset \mathbb{R}$, is a family of probability measures defined on the σ-field of Borel subsets of the real line with monotone likelihood ratio in x, i.e., if for any $\theta_1 < \theta_2$, $f_{\theta_1}(x)$ and $f_{\theta_2}(x)$ are distinct and the ratio $\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}$ is a non decreasing function of x. It is also possible to define families of densities with non increasing monotone likelihood ratio in x, but such families can be treated by symmetry.

Proposition 5. *Assume that $\theta_1, \theta_2 \in \Theta \subset \mathbb{R}$. Suppose that the probability distributions $\{P_\theta\}_{\theta\in\Theta}$ are on the real line, $\theta \in (a, b) \subset \mathbb{R}$ and let P_θ be absolutely continuous with respect to a σ-finite measure μ (Lebesgue measure or counting measure). Suppose also that the corresponding density functions or probability mass functions have monotone likelihood ratio in x. If $a < \theta_1 < \theta_2 < \theta_3 < b$ and the function ϕ is continuous, it holds*

$$D_\phi(\theta_1, \theta_2) \leq D_\phi(\theta_1, \theta_3),\ \phi \in \Phi^* \tag{4.10}$$

The proof can be found in Appendix B2.

Remark 3. *If we consider a function $\phi \in \Phi^*$ which is strictly convex at $x = 1$, the corresponding ϕ-divergence is a reflexive distance on the space $P = \{P\}_{\theta\in\Theta}$. It is possible to define a new measure of divergence, based on a given ϕ-divergence, in such a way that the new measure of divergence will be not only reflexive but also symmetric. This is possible if we consider the measure of divergence associated with the function $\phi(t) = \phi(t) + t\phi(1/t)$. For more details see Vajda (1995).*

4.2.2 Cressie and Read Power Divergence Family

Another important divergence family is the Cressie and Read Power Divergence (1984). This family of measures leads to the family of power divergence test statistics which are used in multinomial goodness of fit tests. This family of test statistics include as special cases Pearson's chi-squared test, log likelihood ratio test, the Freeman-Tukey test and the Neyman modified chi-squared test statistic.

Definition 7. *(Cressie-Read Power Divergence Family)*

The Cressie and Read power divergence measure between two continuous probability distributions $P_{\theta_1}, P_{\theta_2}$ with $\theta_1, \theta_2 \in \Theta$ associated with densities $f_{\theta_1}, f_{\theta_2}$ is defined by

$$D^\lambda_{CR}(P_{\theta_1}, P_{\theta_2}) = \frac{1}{\lambda(\lambda+1)}\left(\int_{\mathcal{X}} \frac{f_{\theta_1}(x)^{\lambda+1}}{f_{\theta_2}(x)^\lambda} d\mu(x) - 1\right) = \frac{1}{\lambda(\lambda+1)}\left(E_{\theta_1}\left[\left(\frac{f_{\theta_1}(X)}{f_{\theta_2}(X)}\right)^\lambda\right] - 1\right),$$

for $-\infty < \lambda < \infty$.

As particular cases for $\lambda \to 0$ we have the Kullback-Leibler divergence and for $\lambda = 1$ we have the Pearson's chi squared divergence.

4.2.3 BHHJ Power Divergence Family

A more recent family of measures proposed by Basu et al. (1998) is the BHHJ (Basu-Harris-Hjort-Jones) or density power divergence family which is characterized by a non-negative index α. This index plays an important role in the trade-off between robustness and asymptotic efficiency of the BHHJ estimator of the parameter involved. This estimator is that which minimizes the BHHJ divergence.

Definition 8. *(BHHJ Power Divergence Family)*

The BHHJ density power divergence measure between two continuous probability distributions P_{θ_1} and P_{θ_2} with $\theta_1, \theta_2 \in \Theta$ associated with densities $f_{\theta_1}, f_{\theta_2}$, corresponding to index α is defined by

$$D_\alpha(P_{\theta_1}, P_{\theta_2}) = \int_{\mathcal{X}} \left\{ f_{\theta_2}^{1+\alpha}(x) - \left(1 + \frac{1}{a}\right) f_{\theta_1}(x) f_{\theta_2}^{\alpha}(x) + \frac{1}{\alpha} f_{\theta_1}^{1+\alpha}(x) \right\} dx, \quad \alpha > 0.$$

The divergence for $\alpha = 1$ reduces to the Euclidean distance. When $\alpha = 0$ the divergence $D_0(P_{\theta_1}, P_{\theta_2})$ is undefined but the limit is the Kullback-Leibler divergence,

$$\lim_{\alpha \longrightarrow 0} D_\alpha(P_{\theta_1}, P_{\theta_2}) = D_{KL}(P_{\theta_1}, P_{\theta_2}) = \int_{\mathcal{X}} f_{\theta_1}(x) \log \frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} dx.$$

4.3 Statistical Inference based on Divergence Measures

4.3.1 Point Estimation

According to ϕ-divergence family Morales et al. (1995) propose the least ϕ-divergence estimator.

Suppose a statistical space $(\mathcal{X}, \mathcal{F}, P_{\underset{\sim}{\theta}\in\Theta})$ associated with the random variable X, $\{P_{\underset{\sim}{\theta}}\}_{\underset{\sim}{\theta}\in\Theta}$ is a family of probability distributions with $\Theta \subset \mathbb{R}^k$, $k \geq 1$. Let $\mathcal{P} = \{E_i\}_{i=1,...,L}$ be a partition of the sample space $\mathcal{X}$ in L intervals. Then $P_{\underset{\sim}{\theta}}(E_i) = p_i(\underset{\sim}{\theta}),\ i = 1, ..., L$ defines a random sample from the population described by the random variable X.

Let $Y_1, ..., Y_N$ be a random sample from distribution F, also $n_i = \sum_{j=1}^{n} I_{E_i}(Y_j)$ and $\hat{p}_i = \frac{n_i}{N}$.

First of all we will show that the maximum likelihood estimator (MLE) of a parameter $\underset{\sim}{\theta}$ equals the minimum Kullback-Leibler divergence estimator.

Theorem 2. *(Equivalence between maximum likelihood and minimum Kullback-Leibler divergence estimators)*

Estimating an unknown parameter $\underset{\sim}{\theta} \in \Theta$ by maximum likelihood method, under the discrete statistical model, equals to minimizing Kullback-Leibler divergence on $\underset{\sim}{\theta} \in \Theta$.

Proof. (Pardo, 2018)

$$
\begin{aligned}
P_{\underset{\sim}{\theta}}(N_1 = n_1, ..., N_L = n_L) &= \frac{n!}{n_1! \cdots n_L!} p_1(\underset{\sim}{\theta})^{n_1} \cdots p_L(\underset{\sim}{\theta})^{n_L} \\
l(\underset{\sim}{\theta}) &= \log \frac{n!}{n_1! \cdots n_L!} + n \sum_{i=1}^{L} \hat{p}_i \log p_i(\underset{\sim}{\theta}) = \\
&= \log \frac{n!}{n_1! \cdots n_L!} - n \sum_{i=1}^{L} \hat{p}_i \log \frac{1}{p_i(\underset{\sim}{\theta})} + n \sum_{i=1}^{L} \hat{p}_i \log \hat{p}_i - n \sum_{i=1}^{L} \hat{p}_i \log \hat{p}_i = \\
&= \log \frac{n!}{n_1! \cdots n_L!} - n \sum_{i=1}^{L} \hat{p}_i \log \frac{\hat{p}_i}{p_i(\theta)} + n \sum_{i=1}^{L} \hat{p}_i \log \hat{p}_i = \\
&= k - n \sum_{i=1}^{L} \hat{p}_i \log \frac{\hat{p}_i}{p_i(\theta)} = \\
&= k - n D_{KL}(\underset{\sim}{\hat{p}}, \underset{\sim}{p}(\underset{\sim}{\theta})),
\end{aligned}
$$

where $k = \log \frac{n!}{n_1! \cdots n_L!} - n \sum_{i=1}^{L} \hat{p}_i \log \hat{p}_i$, $\underset{\sim}{\hat{p}} = (\hat{p}_1, .., \hat{p}_L)^T$ and $\underset{\sim}{p}(\underset{\sim}{\theta}) = (p_1(\underset{\sim}{\theta}), ..., p_L(\underset{\sim}{\theta}))^T$.

So we can easily see that maximizing the log-likelihood $l(\underset{\sim}{\theta})$ is equivalent to minimizing the $D_{KL}(\underset{\sim}{\hat{p}}, \underset{\sim}{p}(\underset{\sim}{\theta}))$ with respect to $\underset{\sim}{\theta}$. □

With the same procedure we can easily show that we can choose an estimator of $\underset{\sim}{\theta}$ that:

$$D(\underset{\sim}{\hat{p}}, \underset{\sim}{p}(\underset{\sim}{\hat{\theta}})) = \inf_{\underset{\sim}{\theta}\in\Theta} D(\underset{\sim}{\hat{p}}, \underset{\sim}{p}(\underset{\sim}{\theta})),$$

where D is any divergence measure.

Following we will see some of the most important minimum divergence estimators.

With $\underset{\sim}{\theta}_0 \in \Theta$ we denote a value that exists when a given model is correct, so $\underset{\sim}{\pi} = \underset{\sim}{p}(\underset{\sim}{\theta}_0)$, where π is the true value of the multinomial probability distribution.

Definition 9. *(Minimum ϕ-divergence estimator)*

Let $Y_1, ..., Y_n$ be a random sample from a population described by the random variable X with associated statistical space $(\mathcal{X}, \mathcal{F}, P_{\underset{\sim}{\theta}})_{\underset{\sim}{\theta}\in\Theta}$. The minimum ϕ-divergence estimator of $\underset{\sim}{\theta}_0$ is any $\underset{\sim}{\hat{\theta}}_\phi \in \Theta$ verifying

$$D_\phi(\underset{\sim}{\hat{p}}, \underset{\sim}{p}(\underset{\sim}{\hat{\theta}}_\phi)) = \inf_{\underset{\sim}{\theta}\in\Theta} D_\phi(\underset{\sim}{\hat{p}}, \underset{\sim}{p}(\underset{\sim}{\theta})),$$

or the minimum ϕ-divergence estimator satisfies the following condition

$$\underset{\sim}{\hat{\theta}}_\phi = arg \inf_{\underset{\sim}{\theta}\in\Theta} D_\phi(\underset{\sim}{\hat{p}}, \underset{\sim}{p}(\underset{\sim}{\theta}))$$

Theorem 3. *(Cressie-Read minimum power divergence estimator)*

In the family of power divergence measures we get the power divergence estimator studied by Cressie and Read (1984). This is given by the condition

$$\underset{\sim}{\hat{\theta}}_{(\lambda)} = arg \inf_{\underset{\sim}{\theta}\in\Theta} D^\lambda_{CR}(\underset{\sim}{\hat{p}}, \underset{\sim}{p}(\underset{\sim}{\theta})),$$

where

$$D^\lambda_{CR}(\underset{\sim}{\hat{p}}, \underset{\sim}{p}(\underset{\sim}{\theta})) = \frac{1}{\lambda(\lambda+1)} \sum_{i=1}^{L} \hat{p}_i \left(\left(\frac{\hat{p}_i}{p_i(\underset{\sim}{\theta})} \right)^\lambda - 1 \right).$$

For particular values of λ we have various estimators. For example, for $\lambda \longrightarrow 0$ we have the maximum likelihood estimator, for $\lambda = 1$ the minimum chi-square estimator and for $\lambda = \frac{2}{3}$ the Cressie-Read estimator.

In the following we will denote as $\hat{\underset{\sim}{\theta}}_D$ the maximum likelihood estimator in the discretized model and $\hat{\underset{\sim}{\theta}}$ the maximum likelihood estimator of the original data.

Before presenting the properties of the minimum ϕ-divergence estimator, we provide some regularity conditions about parameter $\underset{\sim}{\theta}_0$ proposed by Birch (1964). We assume that the model is correct, so $\underset{\sim}{\pi} = \underset{\sim}{p}(\underset{\sim}{\theta}_0)$ and $k \leq L - 1$.

Remark 4. *(Birch's regularity conditions)*

1. $\underset{\sim}{\theta}_0$ *is an interior point of* Θ.

2. $\pi_i = p_i(\underset{\sim}{\theta}_0) > 0$ *for* $i = 1, ..., L$. *Thus* $\underset{\sim}{\pi} = (\pi_1, ..., \pi_L)^T$ *is an interior point of the set* Δ_L.

3. *The mapping* $\underset{\sim}{p} : \Theta \to \Delta_L$ *is totally differentiable at* $\underset{\sim}{\theta}_0$ *so that partial derivatives of* $p_i(\underset{\sim}{\theta}_0)$ *with respect to each* θ_j *exist at* $\underset{\sim}{\theta}_0$ *and* $p_i(\underset{\sim}{\theta})$ *has a linear approximation at* $\underset{\sim}{\theta}_0$ *given by:*

$$p_i(\underset{\sim}{\theta}) = p_i(\underset{\sim}{\theta}_0) + \sum_{j=1}^{L}(\theta_j - \theta_{0j})\frac{\partial p_i(\underset{\sim}{\theta}_0)}{\partial \theta_j} + o(||\underset{\sim}{\theta} - \underset{\sim}{\theta}_0||)$$

where $o(||\underset{\sim}{\theta} - \underset{\sim}{\theta}_0||)$ *denotes a function verifying*

$$\lim_{\underset{\sim}{\theta} \to \underset{\sim}{\theta}_0} \frac{o(||\underset{\sim}{\theta} - \underset{\sim}{\theta}_0||)}{(||\underset{\sim}{\theta} - \underset{\sim}{\theta}_0||)} = 0.$$

4. *The Jacobian matrix*

$$\mathbf{J}(\underset{\sim}{\theta}_0) = \left(\frac{\partial \underset{\sim}{p}(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}}\right)_{\underset{\sim}{\theta} = \underset{\sim}{\theta}_0}$$

is full ranked.

5. *The inverse mapping* $\underset{\sim}{p}^{-1} : T \to \Theta$ *is continuous at* $\underset{\sim}{p}(\underset{\sim}{\theta}_0) = \underset{\sim}{\pi}$.

6. *The mapping* $\underset{\sim}{p} : \Theta \to \Delta_L$ *is continuous at every point* $\underset{\sim}{\theta} \in \Theta$.

Where, Δ_L *is a convex set of probability measures on* $\mathcal{X}$ *defined as*

$$\Delta_L = \left\{ \underset{\sim}{p} = (p_1, ..., p_L)^T : p_i \geq 0, i = 1, ..., L, \sum_{i=1}^{L} p_i = 1 \right\}.$$

The following theorems set that under Birch's conditions the minimum ϕ-divergence estimator is a best asymptotically normal estimator.

Theorem 4. *(Minimum ϕ-divergence Estimator)*

Let $\phi \in \Phi^*$ be a twice continuously differentiable function in $x > 0$ with $\phi^{''}(1) > 0$ and $\underset{\sim}{\pi} = \underset{\sim}{p}(\underset{\sim}{\theta}_0)$. Under the Birch regularity conditions and assuming that the function $\underset{\sim}{p} : \Theta \to \Delta_L$ has continuous second partial derivatives in a neighbourhood of $\underset{\sim}{\theta}_0$, it holds

$$\hat{\underset{\sim}{\theta}}_{\phi} = \underset{\sim}{\theta}_0 + \mathbf{I}_F(\underset{\sim}{\theta}_0)^{-1}\mathbf{B}(\underset{\sim}{\theta}_0)^T diag(\underset{\sim}{\pi}^{-1/2})(\hat{\underset{\sim}{p}} - \underset{\sim}{\pi}) + o(||\hat{\underset{\sim}{p}} - \underset{\sim}{\pi}||),$$

where $\hat{\underset{\sim}{\theta}}_{\phi}$ is unique in a neighbourhood of $\underset{\sim}{\theta}_0$, $\mathbf{I}_F(\underset{\sim}{\theta}_0)$ is the $k \times k$ Fisher information matrix associated with the multinomial model and $\mathbf{B}(\underset{\sim}{\theta}_0) = diag(\underset{\sim}{\pi}^{-1/2})\mathbf{J}(\underset{\sim}{\theta}_0)$.

Theorem 5. *(Asymptotic Distribution of Minimum ϕ-divergence Estimator)*

Under the assumptions of Theorem 4, it holds

$$\sqrt{n}(\hat{\underset{\sim}{\theta}}_{\phi} - \underset{\sim}{\theta}_0) \xrightarrow[n\to\infty]{L} N\left(\underset{\sim}{0}, \mathbf{I}_F(\underset{\sim}{\theta}_0)^{-1}\right).$$

4.3.2 Hypothesis testing in general populations

After that, we can use divergence measures for hypothesis testing in general populations.

Assume the probability measures P_θ associated with densities $f_\theta(x) = \frac{dP_\theta}{d\mu(x)}$ with μ is a σ-finite measure on $(\mathcal{X}, \mathcal{F})$.

Kupperman (1957, 1958) was the first who proposed the idea of a test statistic based on the Kullback-Leibler divergence to test the **simple** null hypothesis $H_0 : \underset{\sim}{\theta} = \underset{\sim}{\theta}_0$ against the **composite** one $H_1 : \underset{\sim}{\theta} \neq \underset{\sim}{\theta}_0$. The proposed test statistic given by

$$T_n^{KL}(\hat{\underset{\sim}{\theta}}, \underset{\sim}{\theta}_0) \equiv 2nD_{KL}((\hat{\underset{\sim}{\theta}}, \underset{\sim}{\theta}_0))$$

is found to be asymptotically chi-squared distributed with k degrees of freedom (as the number of unknown parameters).

In the same philosophy as Kupperman, Salicru et al. (1994) introduced the ϕ-divergence test statistic as follows:

$$T_n^{\phi}(\hat{\underset{\sim}{\theta}}, \underset{\sim}{\theta}_0) = \frac{2n}{\phi''(1)} D_\phi(\hat{\underset{\sim}{\theta}}, \underset{\sim}{\theta}_0)$$

This statistic is an interesting alternative to Wald test statistics and Rao test statistics.

Before we proceed with the asymptotic distribution of ϕ-divergence test statistic we have to make the following regularity assumptions:

1. $\forall \underset{\sim}{\theta}_1 \neq \underset{\sim}{\theta}_2 \in \Theta \subset \mathbb{R}^k,\ k \geq 1$

$$\mu(\{x \in \mathcal{X} : f_{\underset{\sim}{\theta}_1}(x) \neq f_{\underset{\sim}{\theta}_2}(x)\}) > 0.$$

2. The set $S_{\mathcal{X}} = \{x \in \mathcal{X} : f_{\underset{\sim}{\theta}}(x) > 0\}$ is independent of $\underset{\sim}{\theta}$.

3. The first, second and third partial derivatives

$$\frac{\partial f_{\underset{\sim}{\theta}}(x)}{\partial \theta_i}, \frac{\partial^2 f_{\underset{\sim}{\theta}}(x)}{\partial \theta_i \partial \theta_j}, \frac{\partial^3 f_{\underset{\sim}{\theta}}(x)}{\partial \theta_i \partial \theta_j \partial \theta_k} \quad i, j, k = 1, ..., k$$

   exist everywhere $\forall 1 \leq i, j, k \leq k$.

4. The first, second and third partial derivatives of $f_{\underset{\sim}{\theta}}(x)$ with respect to $\underset{\sim}{\theta}$ are absolutely bounded by functions α, β and γ with integrals

$$\int_{\mathcal{X}} \alpha(x) d\mu(x) < \infty,\ \int_{\mathcal{X}} \beta(x) d\mu(x) < \infty,\ \int_{\mathcal{X}} \gamma(x) d\mu(x) < \infty.$$

5. For each $\underset{\sim}{\theta} \in \Theta$, the Fisher information matrix

$$I_{\mathcal{F}}(\underset{\sim}{\theta}) = \left(\int_{\mathcal{X}} \frac{\partial \log f_{\underset{\sim}{\theta}}(x)}{\partial \theta_i} \frac{\partial \log f_{\underset{\sim}{\theta}}(x)}{\partial \theta_j} f_{\underset{\sim}{\theta}}(x) d\mu(x) \right)_{i,j=1,\ldots k}$$

exists and is positive definite, with elements continuous in the variable $\underset{\sim}{\theta}$.

The function ϕ used in ϕ-divergence test statistic has to satisfy the following assumptions:

- $\Phi(1)$ The function $\phi \in \Phi^*$ is twice continuously differentiable, with $\phi^{''}(1) > 0$

- $\Phi(2)$ For each $\underset{\sim}{\theta}_0 \in \Theta$ there exists an open neighbourhood $N(\underset{\sim}{\theta}_0)$ such that for all $\underset{\sim}{\theta} \in N(\underset{\sim}{\theta}_0)$ and $1 \leq i, j \leq k$ it holds:

$$\frac{\partial}{\partial \theta_i} \int_{\mathcal{X}} f_{\underset{\sim}{\theta}_0}(x) \phi \left(\frac{f_{\underset{\sim}{\theta}}(x)}{f_{\underset{\sim}{\theta}_0}(x)} \right) d\mu(x) = \int_{\mathcal{X}} \frac{\partial}{\partial \theta_i} \left(f_{\underset{\sim}{\theta}_0}(x) \phi \left(\frac{f_{\underset{\sim}{\theta}}(x)}{f_{\underset{\sim}{\theta}_0}(x)} \right) \right) d\mu(x),$$

$$\frac{\partial^2}{\partial \theta_i \partial \theta_j} \int_{\mathcal{X}} f_{\underset{\sim}{\theta}_0}(x) \phi \left(\frac{f_{\underset{\sim}{\theta}}(x)}{f_{\underset{\sim}{\theta}_0}(x)} \right) d\mu(x) = \int_{\mathcal{X}} \frac{\partial^2}{\partial \theta_i \partial \theta_j} \left(f_{\underset{\sim}{\theta}_0}(x) \phi \left(\frac{f_{\underset{\sim}{\theta}}(x)}{f_{\underset{\sim}{\theta}_0}(x)} \right) \right) d\mu(x),$$

and these expressions are continuous on $N(\underset{\sim}{\theta}_0)$.

The following theorem presents the asymptotic distribution of the ϕ-divergence test statistic given in Salicru et al. (1994).

Theorem 6. *(Asymptotic Distribution of ϕ-divergence test statistic)*

Let the model $(\mathcal{X}, \mathcal{F}, P_{\underset{\sim}{\theta}})_{\underset{\sim}{\theta} \in \Theta}$. Suppose ϕ satisfies the assumptions 1)-5) and $\Phi(1) - \Phi(2)$. Under the null hypothesis

$$H_0 : \underset{\sim}{\theta} = \underset{\sim}{\theta}_0$$

the asymptotic distribution of the ϕ-divergence test statistic is chi-square with k degrees of freedom:

$$T_n^{\phi}(\hat{\underset{\sim}{\theta}}, \underset{\sim}{\theta}_0) \xrightarrow[n \to \infty]{L} X_k^2$$

Later Morales et al. (1997) studied the ϕ-divergence test statistic for the **composite null hypothesis** $H_0 : \underset{\sim}{\theta} \in \Theta_0 \subset \Theta$:

$$T_n^{\phi}(\hat{\underset{\sim}{\theta}}, \underset{\sim}{\theta}^*) = \frac{2n}{\phi''(1)} D_{\phi}(\hat{\underset{\sim}{\theta}}, \underset{\sim}{\theta}^*)$$

Where $\hat{\underset{\sim}{\theta}}$ is the MLE of $\underset{\sim}{\theta}$ in Θ and $\underset{\sim}{\theta}^*$ is the MLE under the null hypothesis Θ_0.

4.3.3 Goodness of Fit based on Divergence Measures

Goodness of fit tests play a crucial role in statistical theory. Through goodness of fit tests we can identify the behaviour of a random process. **The simple null hypothesis $H_0 : F = F_h$ can be tested if we partition the range of data in disjoint intervals and then testing the hypothesis $H_0 : \underset{\sim}{p} = \underset{\sim}{p}^h$ for the vector of parameters of a multinomial distribution using various divergence test statistics.**

Suppose $\mathcal{P} = \{E_i\}_{i=1,...,L}$ be a partition of $\mathbb{R}$ in L intervals. Let $\underset{\sim}{p} = (p_1, ..., p_L)^T$ and $\underset{\sim}{p}^h = (p_1^h, ..., p_L^h)^T$ be the true and hypothetical probabilities of intervals $E_i,\ i = 1, ..., L$ respectively. Then $p_i = P_F(E_i)\ i = 1, ..., L$ and $p_i^h = P_{F_h}(E_i) = \int_{E_i} dF_h,\ i = 1, ..., L.$

Let $Y_1, ..., Y_N$ be a random sample from F, also $n_i = \sum_{j=1}^{n} I_{E_i}(Y_j)$, where $I_{E_i}(Y_j) = 1$ if $Y_j \in E_i$ and 0 otherwise, $\underset{\sim}{\hat{p}} = (\hat{p_1}, ..., \hat{p_L})^T$ with $\hat{p}_i = \dfrac{n_i}{N},\ i = 1, ..., L.$

Now if we want to test the simple null hypothesis:

$$H_0 : \underset{\sim}{p} = \underset{\sim}{p}^h$$

we can use the various divergence test statistics about them we talked in the previous subsection.

The first who introduce this concept was Cressie and Read (1984) with the family of Cressie-Read power divergence test statistics:

$$T_n^{\lambda}(\underset{\sim}{\hat{p}}, \underset{\sim}{p}^h) = \frac{2N}{\lambda(\lambda+1)} \sum_{i=1}^{L} \hat{p}_i \left(\left(\frac{\hat{p}_i}{p_i^h} \right)^{\lambda} - 1 \right),\ \ -\infty < \lambda < \infty,\ \lambda \neq 0, -1.$$

The Cressie-Read power divergence test statistics is a more general family and includes the chi-square test ($\lambda = 1$) introduced by Pearson (1900) and is defined as:

$$X^2 = \sum_{i=1}^{L} \frac{(n_i - Np_i^h)^2}{Np_i^h}$$

and the likelihood ratio test ($\lambda \to 0$) given by:

$$G^2 = 2 \sum_{i=1}^{L} n_i \log \left(\frac{n_i}{Np_i^h} \right).$$

For ($\lambda = \frac{2}{3}$) we have the Cressie-Read test statistic . Also, Cressie and Read (1984) prove that

$$T_n^{\lambda}(\underset{\sim}{\hat{p}}, \underset{\sim}{p}^h) \xrightarrow[n\to\infty]{L} X^2_{L-1} \ ,$$

under $H_0 : \underset{\sim}{p} = \underset{\sim}{p}^h, \ \forall\lambda \in \mathbb{R}$.

Another option would be to use the family of ϕ-divergence test statistics:

$$T_n^{\phi}(\underset{\sim}{\hat{p}}, \underset{\sim}{p}^h) = \frac{2N}{\phi''(1)} \sum_{i=1}^{L} p_i^h \phi\left(\frac{\hat{p}_i}{p_i^h}\right), \ \ \phi \in \Phi^*$$

In the following theorem we will study the asymptotic distribution of ϕ-divergence test statistic under multinomial testing with a fixed number of classes, as proposed by Zografos et al. (1990).

Theorem 7. *(Asymptotic Distribution of ϕ-divergence test statistic under multinomial testing with a fixed number of classes)*

Under the null hypothesis $H_0 : \underset{\sim}{p} = \underset{\sim}{p}^h = (p_1^h, ..., p_L^h)^T$, the asymptotic distribution of the ϕ-divergence test statistic, $T_n^{\phi}(\underset{\sim}{\hat{p}}, \underset{\sim}{p}^h)$, is chi-square with $L-1$ degrees of freedom.

Based on Theorem 7 if the sample size is large enough someone would propose the following testing 'rule':

'Reject H_0, with a significance level α, if $T_n^{\phi}(\underset{\sim}{\hat{p}}, \underset{\sim}{p}^h) > X^2_{L-1,\alpha}$',

where $\alpha = P(X^2_{L-1} \geq X^2_{L-1,\alpha})$.

4.3.4 Model Selection based on Divergence Measures

As we mention in introduction of this chapter, divergence measures have major applications in model selection criteria. The first who make this relationship clear was the famous statistician Akaike (1973). He proposed the Akaike Information Criterion (AIC) by constructing an unbiased estimator of the expected Kullback-Leibler divergence.

Let f be the 'reality' (or the true model) and g a model which is used to estimate f. The Kullback-Leibler divergence (in continuous case) between f and g, as we saw earlier, is:

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

Here, the D_{KL} represents the information lost when g is used to estimate f. Equivalently we can write:

$$D_{KL}(f,g) = \int_X f(x) \log f(x) dx - \int_X f(x) \log(g(x|\theta)) dx = E_f[\log f(x)] - E_f[\log(g(x|\theta))].$$

The first expectation is constant (z) across model, so

$$D_{KL}(f,g) = z - E_f[\log(g(x|\theta))] \Rightarrow D_{KL}(f,g) - z = -E_f[\log(g(x|\theta))].$$

Following the computation of quantity $E_f[\log(g(x|\theta))]$ is the key to find the relative $D_{KL}(f,g) - z$ distance between f, g. But, this quantity can not be computed. So Akaike found that the expectation:

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

can be computed. Finally, he proposed the asymptotically unbiased estimator of the relative expected Kullback-Leibler information:

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

where p is the number of estimating parameters in model g. 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 Kullback-Leibler divergence estimator). Selecting the model with the smallest AIC value is related to the model with the least Kullback-Leibler divergence between the true one and the estimated.

Chapter 5

Weighted Divergence Measures

A common issue in statistics and other fields dealing with random events is to take into account both the probabilistic aspect and the qualitative characteristics of them. Let us to determine what we mean with the term 'qualitative characteristics'. This is a subjective term and someone would say that it is related to the significance, the relevance or the utility of the information they carry with respect to a goal. But, a question arises. How we will measure the information or the uncertainty with respect to certain characteristics of such events? The foundations of the answer lye on the work of Belis and Guiasu (1968) while the answer itself was given by Guiasu (1971) who proposed the weighted entropy. He explicitly defined the axioms, the properties and the maximum value of weighted entropic formula. After this pioneer work Guiasu (1986) used the weighted entropy to group data with respect to the importance of specific regions of the domain. Later, Narowcki and Harding (1986) proposed the use of weighted entropy as a measure of investment risk, Di Crescenzo and Longobardi (2007) propose the weighted residual and past entropies and Suhov and Zohren (2014) proposed the quantum version of weighted entropy and its properties in quantum statistical mechanics.

Following these, the concept of measuring the dissimilarities between random processes with a higher significance in certain regions of them arise. Several times in statistics we want to emphasize and study a random process with respect to certain characteristics. For example, in financial risk analysis it is common to take care of fat tails in the distribution of an asset (especially the left tail) and in biostatistics to use robust statistical methods to trim the extreme values. These concepts create the need for special statistical methods to study their behaviour. Motivated by these, through this chapter, we will present and study the concept of weighted divergence measures. In contrast with weighted entropy, the weighted form of divergence measures has not extensively studied by researchers. In the previous chapter we stress that divergence measures is a concept to measure the probabilistic dissimilarities between two statistical populations. Correspondingly, the weighted form of divergences measure the probabilistic dissimilarities between two statistical populations while taking into account the qualitative characteristics of each region of the support.

The Chapter 5 will be divided in two sections. In Section 5.1 we will present, in discrete case, the weighted form of Shannon entropy and the weighted ϕ-divergence family, then we will extend it to the weighted Kullback-Leibler divergence. Also, we will underline some problems arising from this concept and we will try to surpass them. In Section 5.2 the continuous case will be presented.

5.1 Discrete Case

5.1.1 Weighted Shannon Entropy

As we mentioned earlier in the introduction of chapter, the first who proposed the weighted form of Shannon entropy and its properties was Guiasu (1971). The relevant definition is presented bellow.

Definition 10. *(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:

$$H^w(X) = H^w(\underset{\sim}{w}, \underset{\sim}{p}) = \sum_{i=1}^{n} w_i p_i \log \frac{1}{p_i}. \tag{5.1}$$

We proceed below with the properties of the above weighted entropy as proposed by Guiasu.

Proposition 6. *(Weighted Shannon Entropy Properties)*

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 $H^w(X) = 0$.*

   This property stresses that if an experiment whose useless or non-significant events are possible and whose useful or significant events are impossible then the total 'weighted' information is equal to zero. We have to notice here that the corresponding Shannon entropy is different from zero (if set J has at least two elements).
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 $H^w(\lambda\underset{\sim}{w}; \underset{\sim}{p}) = \lambda H^w(\underset{\sim}{w}, \underset{\sim}{p}) = \lambda H^w(X)$.*

   Until now we did not require any restrictions about the weights (except that they are non-negative real numbers). 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.:

$$w(E \cup F) = \frac{p(E)w(E) + p(F)w(F)}{p(E) + p(F)} \tag{5.2}$$

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:*

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

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

Also, Guiasu proposed the axioms of weighted entropy and provided the conditions that maximized it. We will suppose that the following axioms hold from now on.

Axiom 1. *$H^w(w_1, w_2; p, 1 - p)$ is a continuous function of p on the interval $[0, 1]$.*

Axiom 2. *$H^w(\underset{\sim}{w}, \underset{\sim}{p})$ is a symmetric function with respect to all pairs of variables (w_i, p_i), $i = 1, ..., n$. This means that it is invariable in any permutation that keeps the pairs (w_i, p_i), $i = 1, ..., n$ unchanged.*

Axiom 3. *If* $w_n = \frac{p'w' + p''w''}{p' + p''}$, $p_n = p' + p''$ *then* $H^w(w_1, ..., w_{n-1}, w', w''; p_1, ..., p_{n-1}, p', p'') = H^w(\underset{\sim}{w}, \underset{\sim}{p}) + p_n H^w(w', w''; \frac{p'}{p_n}, \frac{p''}{p_n})$.

Axiom 4. *If all probabilities are equal* $(p_i = \frac{1}{n},\ i = 1, ..., n)$, *then:*

$$H^w(w_1, ..., w_n; \frac{1}{n}, ..., \frac{1}{n}) = \log n \frac{(w_1 + ... + w_n)}{n},$$

where $\log n > 0,\ \forall n > 1$.

The following theorem states the condition between p and w that maximizes the weighted entropy.

Theorem 8. *(Maximized Weighted Entropy)*

Consider the random variable X associated with the discrete probability distribution $p_i \geq 0$, $i = 1, ..., n$, $\sum_{i=1}^{n} p_i = 1$ and the weights $w_i \geq 0$, $i = 1, ..., n$. The weighted entropy:

$$H^w(X) = H^w(\underset{\sim}{w}, \underset{\sim}{p}) = \sum_{i=1}^{n} w_i p_i \log \frac{1}{p_i}$$

is maximum if and only if:

$$p_i = e^{-(\frac{\alpha}{w_i})-1}, \quad (i = 1, ..., n)$$

where α is the solution of the equation:

$$\sum_{i=1}^{n} e^{-(\frac{\alpha}{w_i})-1} = 1$$

and the maximum value of $H^w(X)$ is:

$$\alpha + \sum_{i=1}^{n} w_i e^{-(\frac{\alpha}{w_i})-1}.$$

Proof. We have that $-xlogx \leq \frac{1}{e}$, $\forall x \geq 0$ and $-xlogx = \frac{1}{e}$ if and only if $x = \frac{1}{e}$, we obtain, by using Lagrange multipliers method:

$$H^w(\underset{\sim}{w}; \underset{\sim}{p}) - \alpha = \sum_{i=1}^{n} w_i p_i log \frac{1}{p_i} - \alpha \sum_{i=1}^{n} p_i = \tag{5.3}$$

$$= \sum_{i=1}^{n} p_i \left(w_i log \frac{1}{p_i} - \alpha \right) = \tag{5.4}$$

$$= \sum_{i=1}^{n} w_i e^{\frac{-\alpha}{w_i}} (p_i e^{\frac{\alpha}{w_i}} log p_i e^{\frac{\alpha}{w_i}}) \leq \sum_{i=1}^{n} w_i e^{-(\frac{\alpha}{w_i})-1}. \tag{5.5}$$

The equality holds if and only if:

$$p_i = e^{-(\frac{\alpha}{w_i})-1}, \quad (i = 1, ..., n),$$

but these probabilities must verify $\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} e^{-(\frac{\alpha}{w_i})-1} = 1$. □

5.1.2 Weighted Divergence Measures

As we mentioned on Chapter 4 the relative entropy is the Kullback-Leibler divergence $D_{KL}(P,Q)$ between distributions P and Q. This is one of the most popular and extensively used divergence measures in the literature. From the statistical point of view, as a divergence, it quantifies the 'distance' between two distributions with the same support. On the other hand, in information theory Kullback-Leibler divergence $D_{KL}(P,Q)$ quantifies the amount of information gained by learning that a variable previously thought to be distributed as P is actually distributed as Q.

But, as in the case of Shannon entropy, the Kullback-Leibler divergence does not take into account the qualitative characteristics of random events. The idea here is to determine the 'distance' between two distributions with specific weight in each part of the support. In the discrete case this is resolved by putting weights on each event of the sample space. The concept of weighting is equivalent to the concept of weighted Shannon entropy. The weights are related to the significance of each event with respect to a specific goal. If one would have thought that the weighted Kullback-Leibler divergence, in discrete case, is analogous to weighted entropy then he/she would have concluded to the following.

Consider two probability mass functions $\underset{\sim}{p} = (p_1, ..., p_n)^T$, $\underset{\sim}{q} = (q_1, ..., q_n)^T$ and $\underset{\sim}{w} = (w_1, ..., w_n)^T$ be a vector of weights. Then the discrete version of weighted Kullback-Leibler divergence would be the following:

$$D^w_{KL}(\underset{\sim}{p}, \underset{\sim}{q}) = \sum_{i=1}^{n} w_i p_i \log\left(\frac{p_i}{q_i}\right) \tag{5.6}$$

This is not correct because the Kullback-Leibler divergence is not positive everywhere in the support.

The following theorem gives the average positivity of Kullback-Leibler divergence measure.

Theorem 9. *(Average positivity of Kullback-Leibler divergence)*

The Kullback-Leibler divergence $D_{KL}(P,Q)$ between two distributions P, Q is positive on average.

Proof. Firstly, we have that $f(x) = \log(x)$ is concave for $x = \dfrac{q(x)}{p(x)} \in \mathbb{R}_+$.

Let S_X be the support of the random variable X and $p(x), q(x)$ the probability mass functions associated with the distributions P, Q respectively. Then:

$$-D_{KL}(P,Q) = -\sum_{x\in S_X} p(x)\log\frac{p(x)}{q(x)} = \sum_{x\in S_X} p(x)\log\frac{q(x)}{p(x)} = E_p\left[\log\frac{q(X)}{p(X)}\right].$$

From Jensen inequality for a concave function we have that:

$$\begin{aligned} -D_{KL}(P,Q) &= E_p\left[\log\frac{q(X)}{p(X)}\right] \\ &\leq log\left[E_p\left(\frac{q(X)}{p(X)}\right)\right] \\ &= \log\sum_{x\in S_X} p(x)\frac{q(x)}{p(x)} \\ &= \log\sum_{x\in S_X} q(x) = \log 1 = 0. \end{aligned}$$

We prove that $-D_{KL}(p,q) = E_p\left[\log\frac{q(X)}{p(X)}\right] \leq 0 \Rightarrow D_{KL}(p,q) = E_p\left[\log\frac{p(X)}{q(X)}\right] \geq 0$ with equality if and only if $p(x) = q(x),\ \forall x \in S_X$. □

Even though it is positive on average, i.e. for the 'finite case of equal weights' it is not positive in general. The following example will make clear that the weighted Kullback-Leibler divergence could be negative.

Example 1. *Let P be a binomial distribution with $n = 2,\ p = 0.4$ and $P(X = 0) = 0.36,\ P(X = 1) = 0.48,\ P(X = 2) = 0.16$ and Q a discrete uniform distribution with the three possible outcomes $X = 0, 1, 2$ each with probability $p = 1/3$. The Kullback-Leibler divergence between P and Q is $D_{KL}(P,Q) = 0.08529$. Now, if we give the event $X = 2$ an enormously greater significance than the others, for example we put the following weights $\underset{\sim}{w} = (1, 1, 4)'$, then the weighted Kullback-Leibler divergence, according to the previous definition, will be $D^w_{KL}(P,Q) = -0.11342$. This is due to the fact that the logarithm in the interval $(0, 1)$ is negative.*

Due to this fact, Kapur (1994) stressed that a weighted divergence measure will be an appropriate measure of weighted directed divergence if the following **conditions** are satisfied:

1. It is a continuous function of $\underset{\sim}{p}$, $\underset{\sim}{q}$ and $\underset{\sim}{w}$.
2. It is permutationally symmetric function of $\underset{\sim}{p}$, $\underset{\sim}{q}$ and $\underset{\sim}{w}$, i.e. it does not change when the triplets (p_1, q_1, w_1), (p_2, q_2, w_2) , ... , (p_n, q_n, w_n) are permuted among themselves.
3. It is always greater than or equal to zero for all possible choices of weights $\underset{\sim}{w}$ and vanishes when $p_i = q_i$ for each $i = 1, ..., n$.
4. It is a convex function of $p_1,\ p_2,\ \ldots,\ p_n$ which has its minimum value zero when $p_i = q_i$ for each $i = 1, ..., n$.
5. It reduces to an ordinary measure of directed divergence upon ignoring weights $(w_i = c,\ c > 0,\ \forall\, i = 1, ..., n)$.

The most important of these limitations is the Condition 3, which is violated for most of the usual measures. The solution to this problem is quite simple. We just need to transform the usual measure to it's positive equivalent. In Section 4.2.1 we presented the ψ-transformation of the ϕ-function in ϕ-divergence family which makes the divergence everywhere in the support positive.

One of the most popular divergences of ϕ-family is the Kullback-Leibler divergence which is obtained for $\phi(x) = x \log x$. If we transform the ϕ-function to the $\psi(x) = x \log x + x - 1$ we have the corrected Kullback-Leibler divergence which is positive everywhere. Then we can construct the weighted form of the Kullback-Leibler divergence which meets the Condition 3.

An other approach to the previous problem was given by Kapur (1994) with the following ϕ-function he introduced:

$$\phi(x) = p(x) \log \left(\frac{p(x)}{q(x)} \right) - p(x) + q(x) \tag{5.7}$$

This function is everywhere in the support positive ($\phi(x) \geq 0, \ \forall \ x \in S_X$). So the divergence which is based on this ϕ-function is also positive for every subset of the support.

As a result for Kapur's corrected Weighted Kullback-Leibler divergence measure we have the following definition.

Definition 11. *(Corrected Weighted Kullback-Leibler Divergence)*

Consider two probability mass functions $\underset{\sim}{p} = (p_1, ..., p_n)^T$, $\underset{\sim}{q} = (q_1, ..., q_n)^T$ *and* $\underset{\sim}{w} = (w_1, ..., w_n)^T$ *be a vector of weights. Then the discrete version of Kapur's corrected weighted Kullback-Leibler divergence is the following:*

$$D^w_{CKL}(\underset{\sim}{p}, \underset{\sim}{q}) = \sum_{i=1}^{n} w_i \left(p_i \log \left(\frac{p_i}{q_i} \right) - p_i + q_i \right)$$

Where $D^w_{CKL}(\underset{\sim}{p}, \underset{\sim}{q}) \geq 0, \ \forall x \in S_X$.

According to information monotonicity we have to be very cautious with the partition of the support. To make it clear the information monotonicity states that the partition of the support may reduce the divergence between two probability distributions. Consider a probability distribution $\underset{\sim}{p} = (p_0, ..., p_n)$ with $p_i = P(X = x_i)$ and the support $S_X = \{x_0, ..., x_n\}$. Let $\{G_i \ | i = 1, ..., m, \ (m < n + 1)\}$ be a partition of S_X, where

$$S_X = \bigcup_{i=1}^{m} G_i$$

$$G_i \cap G_j = \emptyset \text{ for } i \neq j.$$

Let us assume that we do not know the exact value of x_i but we know the group G_j it belongs to, this is coarse-graining of S_X.

Thus, coarse-graining generates a new probability distribution $\underset{\sim}{p}^*$ over $G_1, ..., G_m$,

$$p_j^* = P(G_j) = \sum_{x_i \in G_j} P(X = x_i).$$

Because the coarse-graining effect of S_X summarizes some of x_i into one group, G_j, detailed information is lost. Assume a divergence measure, $D^*(p^*, q^*)$, between probability distributions p^* and q^*, then we have

$$D^*(\underset{\sim}{p}^*, \underset{\sim}{q}^*) \leq D(\underset{\sim}{p}, \underset{\sim}{q}).$$

Now we will study when the equality holds. Assume that the outcome x_i for two distributions $\underset{\sim}{p}, \underset{\sim}{q}$ belongs to group G_j, then if we want to distinguish these two distributions we want specific information inside each group. Since x_i belongs to group G_j we consider the conditional probability distributions

$$p(x_i|G_j),\ q(x_i|G_j).$$

Now if they are equal we have no further information to distinguish $\underset{\sim}{p}$ from $\underset{\sim}{q}$ by observing the outcome x_i inside G_j. So,

$$D^*(\underset{\sim}{p}^*, \underset{\sim}{q}^*) = D(\underset{\sim}{p}, \underset{\sim}{q})$$

holds if and only if

$$p(x_i|G_j) = q(x_i|G_j).$$

A divergence that fulfils the above is called an **invariant divergence** and has the property of **information monotonicity**.

5.2 Continuous Case

5.2.1 Weighted Entropy

Before we propose the continuous version of the weighted entropy, we have to present the continuous version of entropy. Here the problem arises. Let us to take the thread from the beginning. Shannon, as we told earlier in Chapter 4, proposed the entropy as a function which measures the uncertainty or the information of a random source. He proposed the famous entropic formula for a discrete random source, to describe the uncertainty (or the information) a discrete signal contains. But in statistical information theory the nature of a signal could be not only discrete but also continuous. So, there is a need to introduce a formula which will measure the uncertainty in the continuous case. In the literature the continuous entropy is described by the following definition.

Definition 12. *(Continuous Entropy)*

Let a stochastic source X which is described by the continuous probability distribution P with support S_X, μ_p be an absolutely continuous probability measure with respect to μ and p be the induced density. Then, the continuous entropy measure is defined by:

$$h(X) = -\int_{S_X} p(x)\log(p(x))d\mu(x). \tag{5.8}$$

This function, also called **differential entropy**, satisfies some of the properties of a suitable measure of uncertainty but fails to fulfil two of them, the positivity and the invariance under the change of variables. The following examples will make it clear.

Example 2. *Suppose a stochastic source X that is uniformly distributed in the interval (a, b), $X \sim U(a, b)$. The entropy of X is given by:*

$$h(X) = \log(b - a).$$

As it can be easily seen, if $b - a < 1 \Rightarrow b < a + 1$ then the entropy of X will be negative.

Example 3. *Suppose a stochastic source X that is exponentially distributed as $X \sim Exp(\lambda)$ with p.d.f. $p_X(x) = \lambda e^{-\lambda x}$. The entropy of X is given by:*

$$h(X) = 1 - \log(\lambda).$$

As it can be easily seen, if $\log\lambda > 1 \Rightarrow \lambda >$ base of logarithm, then the entropy of X will be negative.

Since results like the above are meaningless, one should focus on the elements that result in a negative uncertainty.

The problem, in our point of view, is that the probability density function (pdf) could take values greater than one. So, the $\log p(x)$ is not negative and sometimes not bounded function. A solution to this problem could be to 'normalize' the probability density function. This possible solution is categorized in two forms according to the support. Definition 13 gives the measure in a bounded support and Definition 14 in an unbounded.

Definition 13. *(Entropy in bounded Continuous Supports)*

Let a stochastic source X which is described by the continuous probability distribution P with a bounded connected support S_X, μ_p be an absolutely continuous probability measure with respect to μ and p be the induced density. Then, the entropy measure would be:

$$h_1(X) = -\int_{S_X} \frac{p(x)}{\max_{x\in S_X}(p(x))} \log_b \left(\frac{p(x)}{\max_{x\in S_X}(p(x))} \frac{1}{b^c} \right) d\mu(x),$$

where c is the range of the support and b the base of the logarithm we use.

Definition 14. *(Entropy in unbounded Continuous Supports)*

Let a stochastic source X which is described by the continuous probability distribution Q with a unbounded connected support S_X, μ_q be an absolutely continuous probability measure with respect to μ and q be the induced density. Then, the entropy measure would be:

$$h_2(X) = -\int_{S_X} \frac{q(x)}{\max_{x\in S_X}(q(x))} \log \left(\frac{q(x)}{\max_{x\in S_X}(q(x))} \right) d\mu(x).$$

In the following examples we will re-evaluate the entropy of uniform and exponential distributions, under the above mentioned measures.

Example 4. *Suppose a stochastic source X that is uniformly distributed in the interval (a, b), $X \sim U(a, b)$. The support is a bounded connected subset of $\mathbb{R}$ so we will use the first form. The entropy of X is given by:*

$$h_1(X) = (b - a)^2.$$

This result is now meaningful in the sense that it is always non-negative, zero if and only if $b = a$ and proportional to the range of the support.

Example 5. *Suppose a stochastic source X that is exponentially distributed as $X \sim Exp(\lambda)$ with p.d.f. $q_X(x) = \lambda e^{-\lambda x}$. The support is an unbounded connected subset of $\mathbb{R}$ so we will use the second form. The entropy of X is given by:*

$$h_2(X) = \frac{1}{\lambda},$$

which is inversely proportional to the parameter λ. This makes sense because it is always non-negative, zero if and only if $\lambda \to \infty$ and the measure decreases (as the variance) as the parameter increases.

Although, this could be considered as an appropriate solution we have to restrict the measure in the family of probability density functions that they have a maximum. This concept needs further investigation. For the rest of this Chapter we will assume that the continuous entropy of a random variable X is of the form (5.8) but we have to mention that this formula is inefficient in some cases.

We will present the definition of continuous weighted entropy in a similar way as Guiasu proposed the weighted Shannon entropy.

Definition 15. *(Weighted Continuous Entropy)*

Let a stochastic source X described by a continuous probability distribution P, μ_p be an absolutely continuous probability measure with respect to μ and p be the induced density. If $w(x)$ is a weighted function assumed to be measurable and positive, then the weighted continuous entropy measure is defined by:

$$H^w(X) = -\int w(x)p(x)\log(p(x))d\mu(x)$$

where $w(x)$ represents the utility or the significance function of each region of the support.

We have to mention some measure theoretic properties about the function $w(x)$, the indicator function and simple functions.

Definition 16. *(Indicator Function)*

The characteristic function (or indicator function) of a subset $E \subset X$ is the function $I_E : X \to \mathbb{R}$ defined by:

$$I_E(x) = \begin{cases} 1, & x \in E \\ 0, & x \notin E \end{cases}$$

The function I_E is measurable if and only if E is a measurable set.

Definition 17. *(Simple Function)*

A simple function $\phi : X \to \mathbb{R}$ on a measurable space $(X, \mathcal{A})$ is a function of the form

$$\phi(x) = \sum_{i=1}^{N} c_i I_{E_i}(x)$$

where $c_1, ..., c_N \in \mathbb{R}$ and $E_1, ..., E_N \in \mathcal{A}$.

Note that, according to this definition, a simple function is measurable. Also, note that the representation of ϕ is not unique. If the constants c_i are distinct and the sets E_i are disjoint the representation is called standard.

Definition 18. *(Measure of a Positive Simple Function)*

Consider a simple function $\phi : X \to [0, \infty)$, as provided by Definition 17. This function will be called positive if $c_i \geq 0$, $\forall i = 1, ..., N$. Then the integral of ϕ with respect to μ is:

$$\int \phi d\mu = \sum_{i=1}^{N} c_i \mu(E_i).$$

Applying a positive simple function as the weighting function $w(x)$ in Definition 15 we get the following result for the continuous weighted Shannon entropy.

Definition 19. *(Weighted Continuous Entropy)*

Let a stochastic source X described by a continuous probability distribution P with probability density p, $A_i \in \mathcal{A}$ be a partition of support S_X, i.e. $\bigcup_{i=1}^{n} A_i = S_X$, $A_i \cap A_j = \emptyset \; \forall \; i \neq j$. Then if $w(x) = \sum_{i=1}^{n} w_i I_{A_i}(x)$ and $\mu_{p|_{A_i}}$ are the restrictions of μ_p at A_i, where μ_p is an absolutely continuous probability measure with respect to μ, then the weighted continuous entropy measure is defined by:

$$H^w(X) = -\sum_{i=1}^{n} \int w_i p(x) \log(p(x)) d\mu_{|_{A_i}}(x) \equiv -\sum_{i=1}^{n} \int_{S_X} I_{A_i}(x) w_i p(x) \log(p(x)) d\mu(x). \tag{5.9}$$

5.2.2 Weighted Divergence Measures

According to Section 5.1.2 the continuous form of weighted divergence measures have a slightly different thinking of construction. In discrete case we multiply each data point of the support with the desirable weight. In the continuous case we have to take into account that the support is infinite, so we have to partition it and apply the appropriate weight in each interval. The thinking is the same as the continuous form of weighted entropy and therefore in the same way as in (5.6) the continuous weighted Kullback-Leibler divergence should be given by the following.

Consider two absolutely continuous probability measures μ_f and μ_g with corresponding densities f, g with respect to a certain measure μ and $A_i \in \mathcal{A}$ be a partition of support S_X, i.e. $\bigcup_{i=1}^{n} A_i = S_X, \quad A_i \cap A_j = \emptyset \ \forall \ i \neq j$. Then, if the weighting function is $w(x) = \sum_{i=1}^{n} w_i I_{A_i}(x)$ the continuous version of weighted Kullback-Leibler divergence measure would be the following:

$$D_{KL}^{w}(f,g) = \sum_{i=1}^{n} w_i \left(\int_{S_X} I_{A_i}(x) \left(f(x) \log \left(\frac{f(x)}{g(x)} \right) \right) d\mu(x) \right) \tag{5.10}$$

As expected, the same issues appear as in the discrete case. Indeed, in some regions of the support the Kullback-Leibler divergence is negative, so if by applying specific weights the divergence would be negative.

According to Kapur's restricted conditions for a weighted divergence measure we can construct the weighted Kullback-Leibler divergence by taking into account the correction given in (5.7).

Definition 20. *(Corrected Weighted Kullback-Leibler Divergence)*

Consider two absolutely continuous probability measures μ_f and μ_g with corresponding densities f, g with respect to a certain measure μ and $A_i \in \mathcal{A}$ be a partition of support S_X, i.e. $\bigcup_{i=1}^{n} A_i = S_X, \quad A_i \cap A_j = \emptyset \ \forall \ i \neq j$. Then if the weighting function is $w(x) = \sum_{i=1}^{n} w_i I_{A_i}(x)$ the continuous version of corrected weighted Kullback-Leibler (KL) divergence measure is the following:

$$D_{CKL}^{w}(f,g) = \sum_{i=1}^{n} w_i \left(\int_{S_X} I_{A_i}(x) \left(f(x) \log \left(\frac{f(x)}{g(x)} \right) - f(x) + g(x) \right) d\mu(x) \right) \tag{5.11}$$

where $D_{CKL}^{w}(f,g) \geq 0, \ \ \forall x \in S_X$.

For understanding Kapur's corrected Kullback-Leibler divergence we need to focus on its functional form. The function of Kapur's corrected divergence has the following form:

$$f(x)\log\left(\frac{f(x)}{g(x)}\right) - f(x) + g(x)$$

where, after mathematical transformations we have:

$$\begin{aligned}
&f(x)\log\left(\frac{f(x)}{g(x)}\right) - f(x) + g(x) \geq 0 \Leftrightarrow \\
&f(x)\log\left(\frac{f(x)}{g(x)}\right) + g(x) \geq f(x) \overset{f(x)\neq 0}{\Leftrightarrow} \\
&\log\left(\frac{f(x)}{g(x)}\right) - \frac{g(x)}{f(x)} \geq 1 \overset{\lambda=\frac{f(x)}{g(x)}}{\Leftrightarrow} \\
&\log(\lambda) - \frac{1}{\lambda} \geq 1.
\end{aligned}$$

The restriction $f(x) \neq 0$ it is not an irrational one, since if we work on the support of the distribution this restriction is always met. From this point on, in the function of Kapur's corrected divergence we will ignore the $g(x)$. So that the 'x-form' and the 'λ-form' forms of the measure would be:

x-form of the divergence measure: $\kappa(x) = x\log(x) - x$
λ-form of the divergence measure: $\tilde{\kappa}(\lambda) = \log(\lambda) - \frac{1}{\lambda}$.

The properties of the 'x-form' and the 'λ-form' will be studied separately. The 'x-form' is needed to be convex with respect to x since convexity between Kapur's function $\kappa(x)$ and density $f(x)$ is required (Condition 4, §5.1.2).

x-form	**λ-form**
$\kappa(x) = x\log(x) - x$	$\tilde{\kappa}(\lambda) = \log(\lambda) + \frac{1}{\lambda}$
$\kappa^{'}(x) = \log x$	$\tilde{\kappa}^{'}(\lambda) = \frac{1}{\lambda} - \frac{1}{\lambda^2} = 0 \Rightarrow \lambda = 1$
$\kappa^{''}(x) = \frac{1}{x} > 0,\ x > 0$	$\tilde{\kappa}^{''}(\lambda) = \frac{-1}{\lambda^2} + \frac{2}{\lambda^3}\ \ with\ \ \tilde{\kappa}^{''}(1) > 0$
κ convex of x	$\tilde{\kappa} \geq 0$
	$\tilde{\kappa}$ min at 1
	$\tilde{\kappa}$ faster

The convexity of $\tilde{\kappa}(\lambda),\ \forall\lambda \in \mathbb{R}_+$ is verified by:

$$
\begin{aligned}
h(\lambda) = \tilde{\kappa}^{''}(\lambda) &= \frac{-1}{\lambda^2} + \frac{2}{\lambda^3} \\
h^{'}(\lambda) &= \frac{2}{\lambda^3} - \frac{6}{\lambda^4} = 0 \Rightarrow \lambda = 3 \\
h^{''}(\lambda) &= \frac{-6}{\lambda^4} + \frac{24}{\lambda^5} \Rightarrow h^{''}(3) = \frac{4}{3} > 0
\end{aligned}
$$

which implies that $h(\lambda) > 0\ \ \forall\lambda \in \mathbb{R}_+ \Rightarrow \tilde{\kappa}^{''}(\lambda) > 0\ \ \forall\lambda \in \mathbb{R}_+$.

Chapter 6

Simulations

In this Chapter we will implement all the weighted cases from Chapter 5. More specifically, we will present both discrete and continuous examples of weighted entropy and weighted corrected Kullback-Leibler divergence. This Chapter will be divided in two sections. In Section 6.1 we will present examples of discrete cases for the weighted Shannon entropy (proposed by Guiasu) and for the weighted corrected Kullback-Leibler divergence (proposed by Kapur). In Section 6.2 we will present examples of continuous cases for the weighted differential entropy and for the weighted corrected Kullback-Leibler divergence.

The simulations have been done with R Project for Statistical Computing (https://www.r-project.org/) version 3.6.2.

6.1 Discrete Case

6.1.1 Weighted Shannon Entropy

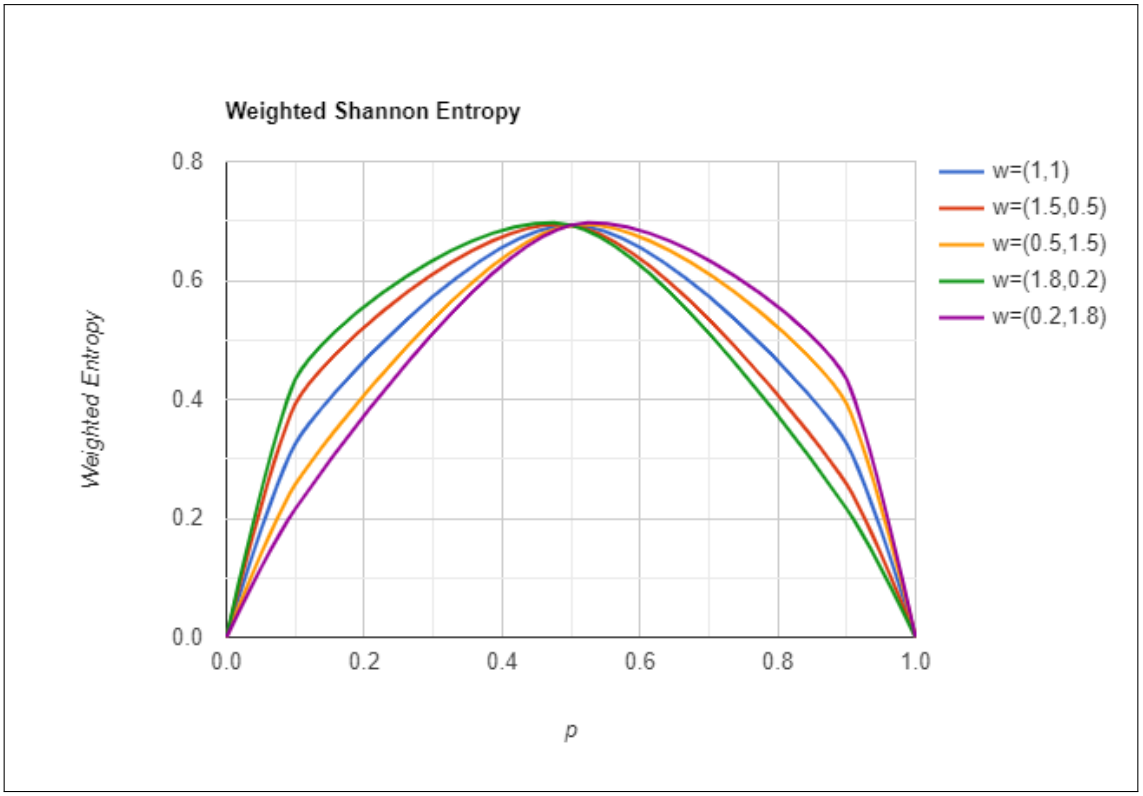
In this section we will present examples of weighted Shannon entropy based on discrete distributions.

Example 6. *(Bernoulli distribution)*

Assume a coin toss, the random variable X which enumerates the probability of heads is described by a Bernoulli distribution with probability p. In the following table we present the weighted entropy of this variable for various p and $\underset{\sim}{w}$.

$\underset{\sim}{w},\ p$	$p = 0.1$	$p = 0.5$	$p = 0.9$
$\underset{\sim}{w} = (1, 1)^T$	0.325	0.693	0.325
$\underset{\sim}{w} = (1.5, 0.5)^T$	0.392	0.693	0.257
$\underset{\sim}{w} = (0.5, 1.5)^T$	0.257	0.693	0.392
$\underset{\sim}{w} = (1.8, 0.2)^T$	0.433	0.693	0.216
$\underset{\sim}{w} = (0.2, 1.8)^T$	0.216	0.693	0.433

In the following graph we depict the above table.



Example 7. *(Binomial distribution)*

Let a random variable X described by a binomial distribution with parameters n, p. The weighted Shannon entropy for various $\underset{\sim}{w}, n, p$ will be given by the following table.

$\underset{\sim}{w}$, (n,p)	$(2,0.4)$	$(2,0.9)$	$\underset{\sim}{w}$, (n,p)	$(3,0.4)$	$(3,0.9)$
$\underset{\sim}{w}=(1,1,1)^T$	1.013	0.525	$\underset{\sim}{w}=(1,1,1,1)^T$	1.22	0.912
$\underset{\sim}{w}=(1.5,0.5,1.5)^T$	1.167	0.4794	$\underset{\sim}{w}=(1.5,0.5,0.5,1.5)^T$	1.12	1.017
$\underset{\sim}{w}=(0.5,2,0.5)^T$	1.035	0.7256	$\underset{\sim}{w}=(0.5,1.5,1.5,0.5)^T$	1.33	0.806

Example 8. *(Weighted Shannon Entropy as a Risk Measure)*

Inspired by the work of Nawrocki and Harding (1986) we extend the example they have proposed. Consider two assets, A and B. The returns and the probability of each to occur for each asset is given in the following table.

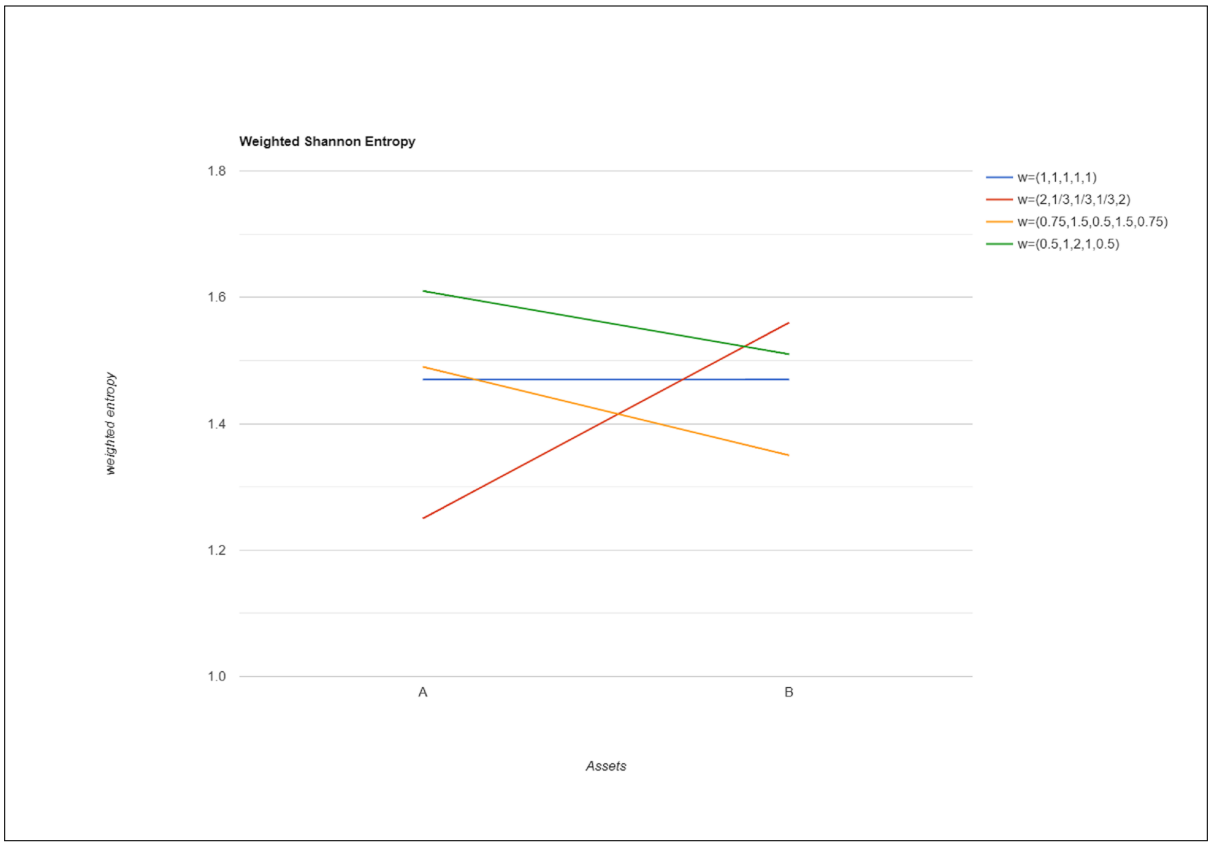
State i	**Return**	**A**(p_i)	**B**(q_i)
1	1%	0.1	0.2
2	3%	0.2	0.1
3	5%	0.4	0.4
4	8%	0.2	0.1
5	10%	0.1	0.2

The Shannon entropy of each asset is H(A)=H(B)=1.47. But, someone could easily recognize that the asset B is riskier than the asset A. Thus, using the weighted Shannon entropy with specific weights on each state (according to its significance) we can get the

uncertainty (or the riskiness) of each asset. In the following table we present the weighted Shannon entropy for various vectors of weights $\underline{w}$.

$\underline{w}$	$\mathbf{H^w(A)}$	$\mathbf{H^w(B)}$
$\underline{w} = (1, 1, 1, 1, 1)^T$	1.47	1.47
$\underline{w} = (2, 0.333, 0.333, 0.333, 2)^T$	1.25	1.56
$\underline{w} = (0.75, 1.5, 0.5, 1.5, 0.75)^T$	1.49	1.35
$\underline{w} = (0.5, 1, 2, 1, 0.5)^T$	1.61	1.51

So, if we apply larger weights on the riskier states we get bigger weighted entropy which is directly related to higher risk.

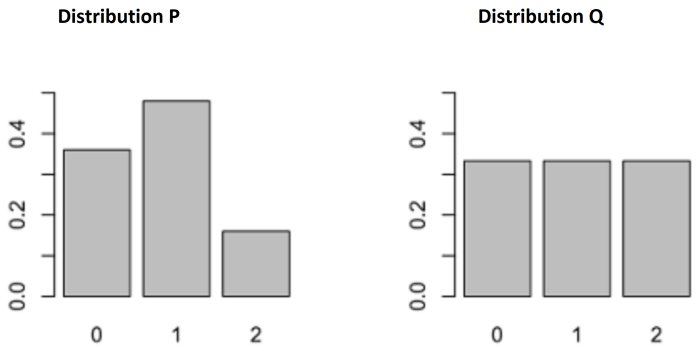


6.1.2 Weighted Corrected Kullback-Leibler Divergence

In this subsection we present examples of the corrected weighted Kullback-Leibler (CWKL) divergence between two discrete distributions. We consider be cases based on several distributions and various weights.

Example 9. *(Binomial and Discrete Uniform distribution)*

Let P be a binomial distribution with $n = 2,\ p = 0.4$ hence $P(X = 0) = 0.36,\ P(X = 1) = 0.48,\ P(X = 2) = 0.16$ and Q a discrete uniform distribution with the three possible outcomes $Y = 0, 1, 2$ each with probability $P(Y = 0, 1, 2) = 1/3$.



In the following table we present the CWKL between these distributions for various $\underset{\sim}{w}$.

$\underset{\sim}{w}$	$\mathbf{D^w_{CKL}(P, Q)}$
$\underset{\sim}{w} = (1, 1, 1)^T$	0.0852
$\underset{\sim}{w} = (1.25, 0.5, 1.25)^T$	0.0853
$\underset{\sim}{w} = (0.5, 2, 0.5)^T$	0.0851
$\underset{\sim}{w} = (0.5, 1.25, 1.25)^T$	0.106
$\underset{\sim}{w} = (2, 0.5, 0.5)^T$	0.044

It is clear that if we apply larger weights on states with the greatest 'distance' (like $(X, Y) = (1, 1)$ and $(X, Y) = (2, 2)$) we get a significant bigger CWKL divergence. On the other hand, if we apply bigger weights on states with the least 'distance' (like $(X, Y) = (0, 0)$) we get a smaller CWKL divergence.

Example 10. *(Binomial distributions)*

In the following table we present the CWKL divergence between binomial distributions P, Q with $n = 3$, different weights $\underset{\sim}{w}$ and probabilities p_1, p_2.

	$\underset{\sim}{w} = (1, 1, 1, 1)^T$	$\underset{\sim}{w} = (1.25, 0.75, 0.75, 1.25)^T$	$\underset{\sim}{w} = (0.75, 1.25, 1.25, 0.75)^T$
$(p_1 = 0.4, p_2 = 0.6)$	0.234	0.264	0.204
$(p_1 = 0.2, p_2 = 0.5)$	0.578	0.648	0.507
$(p_1 = 0.1, p_2 = 0.9)$	0.11	0.105	0.114

Example 11. *(CWKL divergence between two assets)*

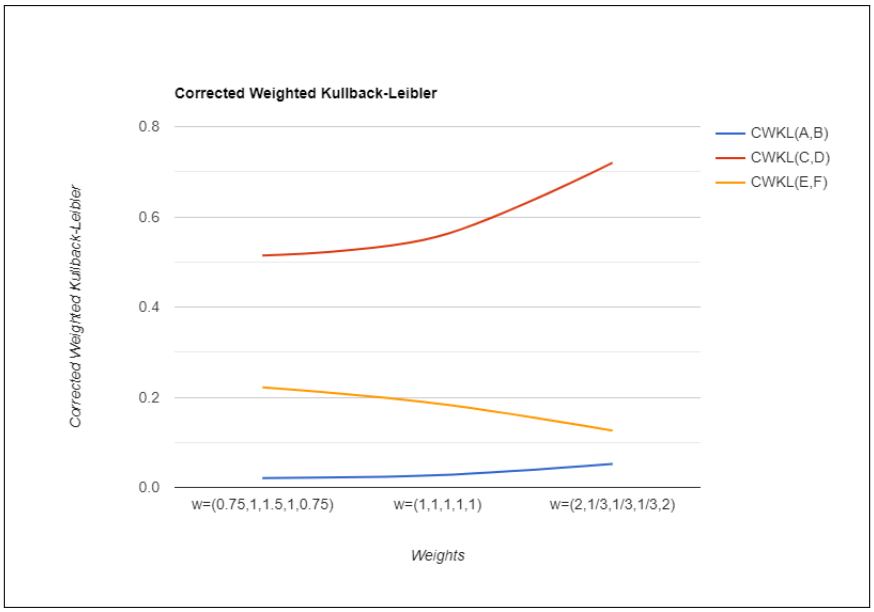
Consider the assets, A, B, C, D, E and F. The returns associated with five states and the probability of each to occur are given in the following table.

State i	**Return**	$\mathbf{A}(p_i)$	$\mathbf{B}(q_i)$	$\mathbf{C}(p_i)$	$\mathbf{D}(q_i)$	$\mathbf{E}(p_i)$	$\mathbf{F}(p_i)$
1	1%	0.1	0.14	0.05	0.3	0.1	0.05
2	3%	0.21	0.22	0.25	0.1	0.25	0.15
3	5%	0.38	0.35	0.4	0.2	0.3	0.6
4	8%	0.21	0.22	0.25	0.1	0.25	0.15
5	10%	0.1	0.07	0.05	0.3	0.1	0.05

In the following table we present the CWKL divergence between the assets A, B, C, D, E and F for various vectors of weights $\underset{\sim}{w}$.

$\underset{\sim}{w}$	$\mathbf{D^w_{CKL}(P_A, Q_B)}$	$\mathbf{D^w_{CKL}(P_C, Q_D)}$	$\mathbf{D^w_{CKL}(P_E, Q_F)}$
$\underset{\sim}{w} = (1,1,1,1,1)^T$	0.027	0.556	0.186
$\underset{\sim}{w} = (2,0.333,0.333,0.333,2)^T$	0.052	0.72	0.126
$\underset{\sim}{w} = (0.75,1,1.5,1,0.75)^T$	0.021	0.514	0.222

As we can easily see the probability distributions of the assets above, in some cases like assets A and B, are quite similar. In this case, someone would observe that the greater differences are in states 1 *and* 5*, the extreme 'states'. So, when we apply bigger weights in these states then the $D^w_{CKL}(P_A, Q_B)$ increases significantly (from* 0.027 *to* 0.052*). On the other hand, when we apply bigger weights in the more 'similar' states like* 2,3,4 *the $D^w_{CKL}(P_A, Q_B)$ decreases (from* 0.027 *to* 0.021*). In the case of assets C and D the probability distributions are definitely different, so among all cases here we have the biggest value for the divergence $D^w_{CKL}(P_C, Q_D)$ for bigger weights in the 'extreme' states. In the last case of assets E and F the probability distributions are quite similar in the 'extreme' states* $1, 5$ *and they distinct in the more frequent states* $2, 4$ *and especially* 3*. As a consequence the $D^w_{CKL}(P_E, Q_F)$ is higher when we apply larger weights in the 'middle' states* $2, 4$ *and* 3*. All the above are visualized in the following graph.*



6.2 Continuous Case

6.2.1 Weighted Entropy

In this subsection we present examples of the weighted differential entropy based on different continuous distributions.

Example 12. *(Normal distribution)*

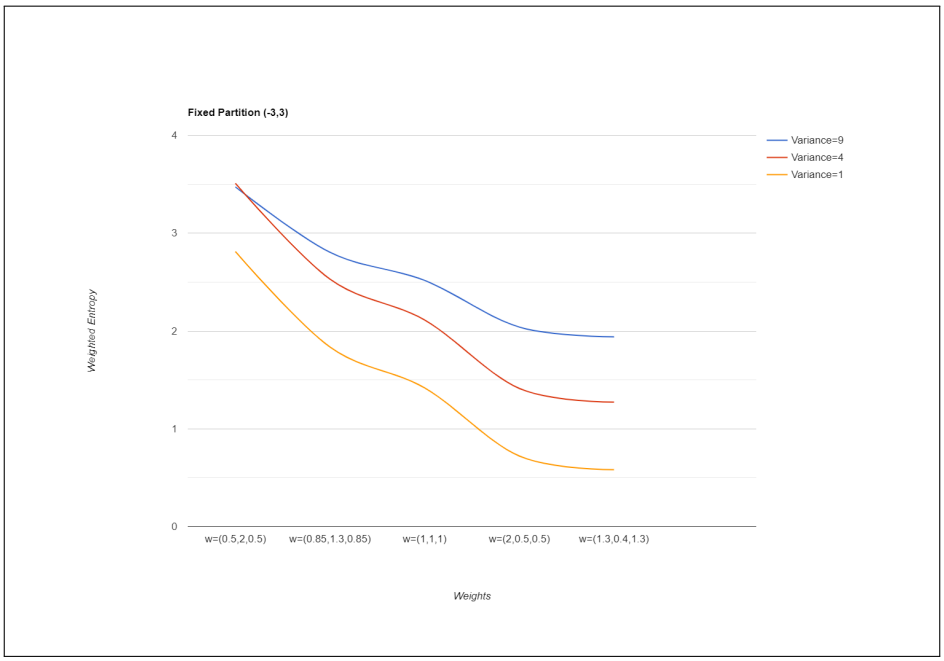
Let a stochastic source X described by a Normal distribution with parameters $\mu = 0$ and $\sigma^2 = 1$. In the table below we present the weighted entropy of this source for various weights $\underset{\sim}{w}$ applied on different parts of the support.

$\underset{\sim}{w}$	$\mathbf{S_1} = (-\infty, \mathbf{-3}), \mathbf{S_2} = (\mathbf{-3}, \mathbf{3}), \mathbf{S_3} = (\mathbf{3}, \infty)$	$\mathbf{S_1} = (-\infty, \mathbf{-1}), \mathbf{S_2} = (\mathbf{-1}, \mathbf{1}), \mathbf{S_3} = (\mathbf{1}, \infty)$
$\underset{\sim}{w} = (1, 1, 1)^T$	1.418	1.418
$\underset{\sim}{w} = (1.3, 0.4, 1.3)^T$	0.582	1.191
$\underset{\sim}{w} = (0.85, 1.3, 0.85)^T$	1.836	1.533
$\underset{\sim}{w} = (2, 0.5, 0.5)^T$	0.722	1.228
$\underset{\sim}{w} = (0.5, 2, 0.5)^T$	2.812	1.799

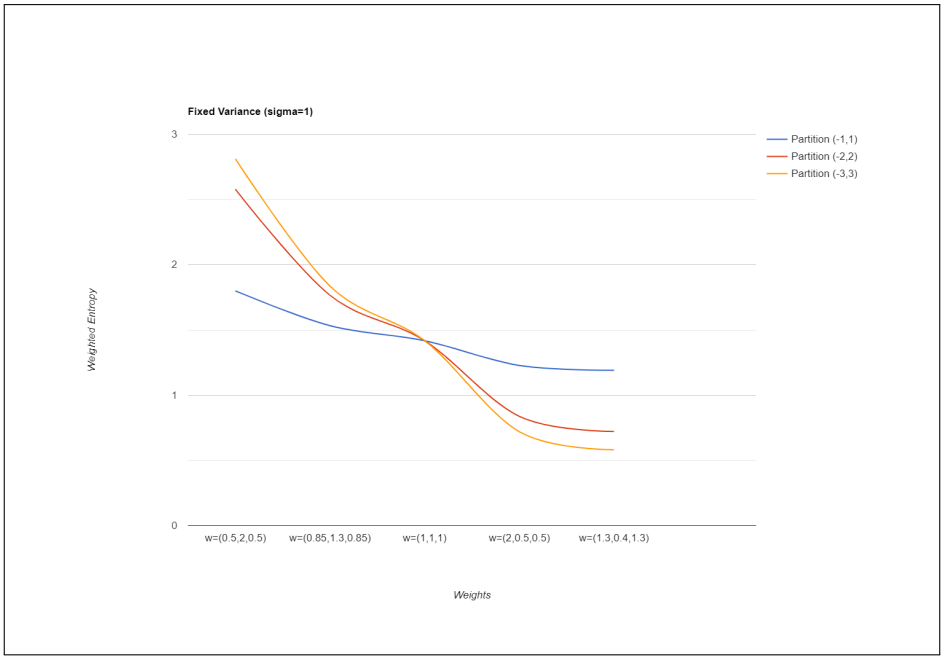
Now, let us assume that the stochastic source X is described by a Normal distribution with parameters $\mu = 0$ and $\sigma^2 = 9$. In the table below we present the weighted entropy of this source for various weights $\underset{\sim}{w}$ applied in different parts of the support.

$\underset{\sim}{w}$	$\mathbf{S_1} = (-\infty, \mathbf{-9}), \mathbf{S_2} = (\mathbf{-9}, \mathbf{9}), \mathbf{S_3} = (\mathbf{9}, \infty)$	$\mathbf{S_1} = (-\infty, \mathbf{-3}), \mathbf{S_2} = (\mathbf{-3}, \mathbf{3}), \mathbf{S_3} = (\mathbf{3}, \infty)$
$\underset{\sim}{w} = (1, 1, 1)^T$	2.517	2.517
$\underset{\sim}{w} = (1.3, 0.4, 1.3)^T$	1.025	1.943
$\underset{\sim}{w} = (0.85, 1.3, 0.85)^T$	3.263	2.804
$\underset{\sim}{w} = (2, 0.5, 0.5)^T$	1.273	2.039
$\underset{\sim}{w} = (0.5, 2, 0.5)^T$	5	3.473

In the following graph we visualize the change of the weighted entropy of various random variables described by a Normal distribution with $\mu = 0$ and $\sigma^2 =$ Variance with a fixed partition of the support $S_1 = (-\infty, -3), S_2 = (-3, 3), S_3 = (3, \infty)$.



Following, we visualize the change of the weighted entropy, with various support partitions, of a random variable described by a Normal distribution with $\mu = 0$ *and* $\sigma^2 = 1$.



Example 13. *(Student's t-distribution)*

Let a stochastic source X *described by a Student's t-distribution with one degree of freedom* $(df = 1)$. *In the table below we present the weighted entropy of this source for various weights* $\underset{\sim}{w}$ *applied in different parts of the support.*

$\underset{\sim}{w}$	$\mathbf{S_1} = (-\infty, -\mathbf{3}), \mathbf{S_2} = (-\mathbf{3}, \mathbf{3}), \mathbf{S_3} = (\mathbf{3}, \infty)$	$\mathbf{S_1} = (-\infty, -\mathbf{1}), \mathbf{S_2} = (-\mathbf{1}, \mathbf{1}), \mathbf{S_3} = (\mathbf{1}, \infty)$
$\underset{\sim}{w} = (1, 1, 1)^T$	2.531	2.531
$\underset{\sim}{w} = (1.3, 0.4, 1.3)^T$	2.012	2.676
$\underset{\sim}{w} = (0.85, 1.3, 0.85)^T$	2.791	2.088
$\underset{\sim}{w} = (2, 0.5, 0.5)^T$	2.098	2.652
$\underset{\sim}{w} = (0.5, 2, 0.5)^T$	3.395	2.289

In this point it seems reasonable to refer that a Student's t-distribution with $df \to \infty$ *is an excellent alternative of Standard Normal distribution so all the values of Example* 8 *are almost the same as in the Example* 7.

Example 14. *(Cauchy-Lorentz distribution)*

Let a stochastic source X *described by a Cauchy distribution with location parameter* $\chi_0 = 0$ *and scale parameter* $\gamma = 2$. *In the table below we present the weighted entropy of this source for various weights* $\underset{\sim}{w}$ *applied in different parts of the support.*

$\underset{\sim}{w}$	$\mathbf{S_1} = (-\infty, -\mathbf{3}), \mathbf{S_2} = (-\mathbf{3}, \mathbf{3}), \mathbf{S_3} = (\mathbf{3}, \infty)$	$\mathbf{S_1} = (-\infty, -\mathbf{1}), \mathbf{S_2} = (-\mathbf{1}, \mathbf{1}), \mathbf{S_3} = (\mathbf{1}, \infty)$
$\underset{\sim}{w} = (1, 1, 1)^T$	3.224	3.224
$\underset{\sim}{w} = (1.3, 0.4, 1.3)^T$	2.953	3.683
$\underset{\sim}{w} = (0.85, 1.3, 0.85)^T$	3.359	2.994
$\underset{\sim}{w} = (2, 0.5, 0.5)^T$	2.998	3.612
$\underset{\sim}{w} = (0.5, 2, 0.5)^T$	3.675	2.458

Example 15. *(Negatively Skewed Normal distribution)*

Let a stochastic source X described by a Negatively Skewed Normal distribution with parameters $\xi = 0, \omega = 1$ and $\alpha = -10$. In the table below we present the weighted entropy of this source for various weights $\underset{\sim}{w}$ applied in different parts of the support.

$\underset{\sim}{w}$	$\mathbf{S_1} = (-\infty, \mathbf{-3}), \mathbf{S_2} = (\mathbf{-3}, \mathbf{3}), \mathbf{S_3} = (\mathbf{3}, \infty)$	$\mathbf{S_1} = (-\infty, \mathbf{-1}), \mathbf{S_2} = (\mathbf{-1}, \mathbf{1}), \mathbf{S_3} = (\mathbf{1}, \infty)$
$\underset{\sim}{w} = (1, 1, 1)^T$	0.797	0.797
$\underset{\sim}{w} = (1.3, 0.4, 1.3)^T$	0.909	0.744
$\underset{\sim}{w} = (0.85, 1.3, 0.85)^T$	0.741	0.824
$\underset{\sim}{w} = (2, 0.5, 0.5)^T$	0.398	1.107
$\underset{\sim}{w} = (0.5, 0.5, 2)^T$	1.383	0.398
$\underset{\sim}{w} = (0.5, 2, 0.5)^T$	0.611	0.886

Because the Skewed Normal Distribution is not a symmetric distribution then the weighted entropy with higher significance in the left part ($\underset{\sim}{w} = (2, 0.5, 0.5)^T$) is not equal with the weighted entropy with higher significance in the right part ($\underset{\sim}{w} = (0.5, 0.5, 2)^T$).

6.2.2 Weighted Corrected Kullback-Leibler Divergence

In the last part of this Chapter we will present examples of the corrected weighted Kullback-Leibler (CWKL) divergence between two continuous distributions. There will be a variety in weights and probability distributions involved in.

Example 16. *(Normal distributions)*

Let P be a Normal distribution with parameters $\mu = 0, \sigma^2 = 1$ and Q an other Normal distribution with parameters $\mu = 0, \sigma^2 = 4$. The CWKL divergence between P and Q for various weights $\underset{\sim}{w}$ applied in different parts of the support is given in the following tables.

$\underset{\sim}{w}$	$\mathbf{S_1} = (-\infty, \mathbf{-3}), \mathbf{S_2} = (\mathbf{-3}, \mathbf{3}), \mathbf{S_3} = (\mathbf{3}, \infty)$	$\mathbf{S_1} = (-\infty, \mathbf{-1.36}), \mathbf{S_2} = (\mathbf{-1.36}, \mathbf{1.36}), \mathbf{S_3} = (\mathbf{1.36}, \infty)$
$\underset{\sim}{w} = (1, 1, 1)^T$	0.318	0.318
$\underset{\sim}{w} = (1.3, 0.4, 1.3)^T$	0.236	0.322
$\underset{\sim}{w} = (0.85, 1.3, 0.85)^T$	0.358	0.316
$\underset{\sim}{w} = (2, 0.5, 0.5)^T$	0.251	0.321
$\underset{\sim}{w} = (0.5, 2, 0.5)^T$	0.453	0.311

$\underset{\sim}{w}$	$\mathbf{S_1} = (-\infty, \mathbf{-6}), \mathbf{S_2} = (\mathbf{-6}, \mathbf{6}), \mathbf{S_3} = (\mathbf{6}, \infty)$
$\underset{\sim}{w} = (1, 1, 1)^T$	0.318
$\underset{\sim}{w} = (1.3, 0.4, 1.3)^T$	0.129
$\underset{\sim}{w} = (0.85, 1.3, 0.85)^T$	0.412
$\underset{\sim}{w} = (2, 0.5, 0.5)^T$	0.161
$\underset{\sim}{w} = (0.5, 2, 0.5)^T$	0.632

Example 17. *(Normal and Student's t-distributions)*

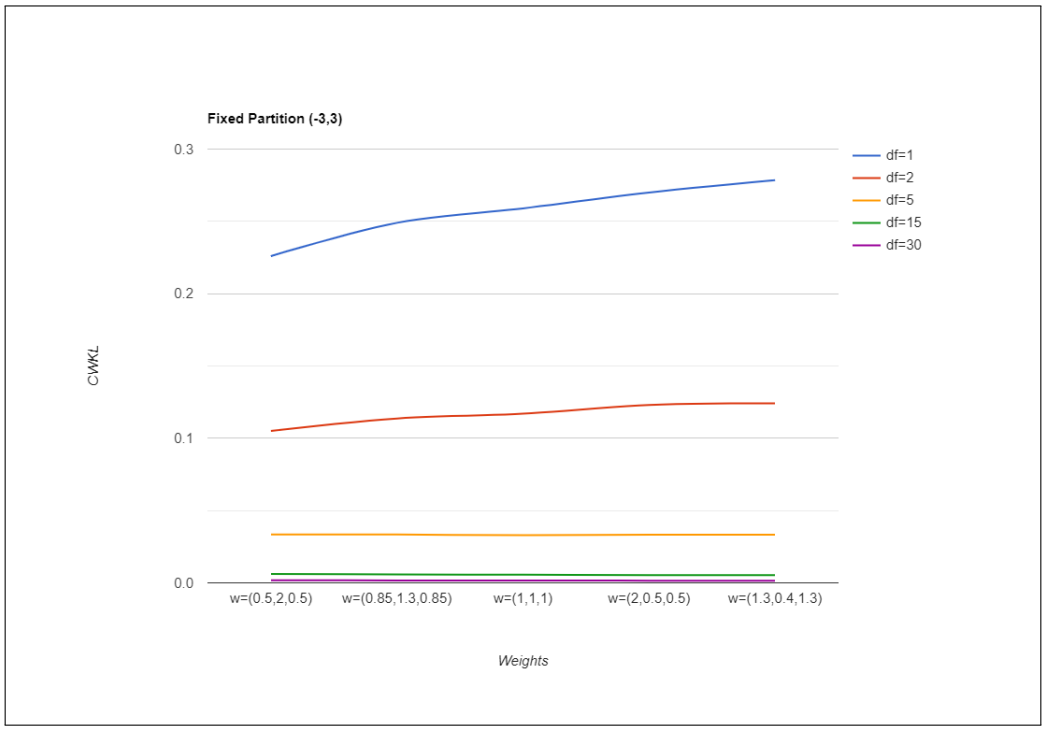
*Let P be a Normal distribution with parameters $\mu = 0, \sigma^2 = 1$ and Q a Student's t-distribution distribution with one degree of freedom (**df=1**). The CWKL divergence between P and Q for various weights $\underset{\sim}{w}$ applied in different parts of the support is given in the following table.*

$\underset{\sim}{w}$	$\mathbf{S_1} = (-\infty, \mathbf{-3}), \mathbf{S_2} = (\mathbf{-3}, \mathbf{3}), \mathbf{S_3} = (\mathbf{3}, \infty)$	$\mathbf{S_1} = (-\infty, \mathbf{-1.85}), \mathbf{S_2} = (\mathbf{-1.85}, \mathbf{1.85}), \mathbf{S_3} = (\mathbf{1.85}, \infty)$
$\underset{\sim}{w} = (1, 1, 1)^T$	0.259	0.259
$\underset{\sim}{w} = (1.3, 0.4, 1.3)^T$	0.278	0.297
$\underset{\sim}{w} = (0.85, 1.3, 0.85)^T$	0.249	0.241
$\underset{\sim}{w} = (2, 0.5, 0.5)^T$	0.275	0.291
$\underset{\sim}{w} = (0.5, 2, 0.5)^T$	0.226	0.196

*Now, let P be a Normal distribution with parameters $\mu = 0, \sigma^2 = 1$ and Q a Student's t-distribution distribution with thirty degrees of freedom (**df=30**). The CWKL divergence between P and Q for various weights $\underset{\sim}{w}$ applied in different parts of the support is given in the following table.*

$\underset{\sim}{w}$	$\mathbf{S_1} = (-\infty, \mathbf{-3}), \mathbf{S_2} = (\mathbf{-3}, \mathbf{3}), \mathbf{S_3} = (\mathbf{3}, \infty)$	$\mathbf{S_1} = (-\infty, \mathbf{-1.85}), \mathbf{S_2} = (\mathbf{-1.85}, \mathbf{1.85}), \mathbf{S_3} = (\mathbf{1.85}, \infty)$
$\underset{\sim}{w} = (1, 1, 1)^T$	0.00162	0.00162
$\underset{\sim}{w} = (1.3, 0.4, 1.3)^T$	0.0015	0.002
$\underset{\sim}{w} = (0.85, 1.3, 0.85)^T$	0.00169	0.00141
$\underset{\sim}{w} = (2, 0.5, 0.5)^T$	0.00152	0.00197
$\underset{\sim}{w} = (0.5, 2, 0.5)^T$	0.00183	0.0009

In the following graph we visualize the variations of the CWKL between a Normal distribution with $\mu = 0$ and $\sigma^2 = 1$ and a Student's t-distribution with various degrees of freedom (df). A plethora of weights are applied on a fixed partition of the support $S_1 = (-\infty, -3), S_2 = (-3, 3), S_3 = (3, \infty)$.



Example 18. *(Normal and Cauchy distributions)*

Let P be a Normal distribution with parameters $\mu = 0, \sigma^2 = 1$ and Q a Cauchy distribution with location parameter $\chi_0 = 0$ and scale parameter $\gamma = 0.8$. The CWKL divergence between P and Q for various weights $\underset{\sim}{w}$ applied in different parts of the support is given in the following table.

$\underset{\sim}{w}$	$\mathbf{S_1} = (-\infty, \mathbf{-3}), \mathbf{S_2} = (\mathbf{-3}, \mathbf{3}), \mathbf{S_3} = (\mathbf{3}, \infty)$	$\mathbf{S_1} = (-\infty, \mathbf{-2}), \mathbf{S_2} = (\mathbf{-2}, \mathbf{2}), \mathbf{S_3} = (\mathbf{2}, \infty)$
$\underset{\sim}{w} = (1, 1, 1)^T$	0.205	0.205
$\underset{\sim}{w} = (1.3, 0.4, 1.3)^T$	0.222	0.234
$\underset{\sim}{w} = (0.85, 1.3, 0.85)^T$	0.196	0.191
$\underset{\sim}{w} = (2, 0.5, 0.5)^T$	0.219	0.229
$\underset{\sim}{w} = (0.5, 2, 0.5)^T$	0.176	0.156

Example 19. *(Cauchy and Student's t-distributions)*

Let P be a Cauchy distribution with location parameter $\chi_0 = 0$ and scale parameter $\gamma = 1.05$ and Q a Student's t-distribution distribution with one degree of freedom (df=1). The CWKL divergence between P and Q for various weights $\underset{\sim}{w}$ applied in different parts of the support is given in the following tables.

w	$\mathbf{S_1} = (-\infty, -\mathbf{1}), \mathbf{S_2} = (-\mathbf{1}, \mathbf{1}), \mathbf{S_3} = (\mathbf{1}, \infty)$	$\mathbf{S_1} = (-\infty, -\mathbf{3}), \mathbf{S_2} = (-\mathbf{3}, \mathbf{3}), \mathbf{S_3} = (\mathbf{3}, \infty)$
$w = (1, 1, 1)^T$	59×10^{-5}	59×10^{-5}
$w = (1.3, 0.4, 1.3)^T$	51×10^{-5}	43×10^{-5}
$w = (0.85, 1.3, 0.85)^T$	63×10^{-5}	67×10^{-5}
$w = (2, 0.5, 0.5)^T$	52×10^{-5}	46×10^{-5}
$w = (0.5, 2, 0.5)^T$	74×10^{-5}	83×10^{-5}

w	$\mathbf{S_1} = (-\infty, -\mathbf{3}), \mathbf{S_2} = (-\mathbf{3}, -\mathbf{1}), \mathbf{S_3} = (-\mathbf{1}, \mathbf{1}), \mathbf{S_4} = (\mathbf{1}, \mathbf{3}), \mathbf{S_5} = (\mathbf{3}, \infty)$
$w = (1, 1, 1, 1, 1)^T$	59×10^{-5}
$w = (1.5, 0.75, 0.5, 0.75, 1.5)^T$	42×10^{-5}
$w = (0.75, 1, 1.5, 1, 0.75)^T$	77×10^{-5}
$w = (0.5, 1, 2, 1, 0.5)^T$	94×10^{-5}
$w = (0.5, 1.5, 1, 1.5, 0.5)^T$	58×10^{-5}
$w = (1, 1.25, 0.5, 1.25, 1)^T$	41×10^{-5}

Chapter 7

Conclusion

As it is clear from our study, the results between the classical and the weighted formulas are completely different when we focus on specific parts of the distributions. This is exactly what we want to showcase through this thesis.

In an experiment with two possible outcomes like Example 1 the Shannon Entropy is symmetric around its maximum value which is the equiprobable condition but the Weighted Shannon Entropy is not symmetric and takes larger values in the conditions with larger weights (associated with bigger significance). On the other hand, in an experiment with many possible outcomes like Example 8 with the assets, it is clear that the Weighted Shannon Entropy is significantly better to detect riskiness than the classical Shannon Entropy. We observe that if we apply larger weights on the riskier states of the assets we get bigger Weighted Shannon Entropy which is directly related to higher risk.

As in the case of Weighted Shannon Entropy and the classical Shannon Entropy the case of CWKL divergence between two assets is completely different from the classical Corrected Kullback-Leibler divergence. It is clear that when the probability distributions of the assets are quite similar then the CWKL with ideal weights performs better than the classical Corrected Kullback-Leibler divergence for the detection of dissimilarities. Even though when we apply bigger weights in the more 'similar' (from the probabilistic point of view) states then the CWKL is smaller than the classical Corrected Kullback-Leibler divergence. This seems logical because we seek for dissimilarities in the parts where they do not exist.

To summarize, the CWKL divergence is larger than the Corrected Kullback-Leibler divergence if we apply bigger weights in the parts with greater dissimilarities. From the other hand, the CWKL divergence is less than the Corrected Kullback-Leibler divergence when we apply bigger weights in the parts with less dissimilarities. Lastly, the CWKL divergence coincides to Corrected Kullback-Leibler divergence if we apply equal weights in all parts.

As it is clear from all the above, the appropriate choice of weights will result in better discrimination between similar distributions. In addition, they provide us with a framework for concentrating on the 'important' parts of the distribution.

These weighted measures could be a useful tool for the construction of 'directed' statistical tests on the parts of the distribution we wish to emphasize. Such tests could include goodness of fit tests or model selection criteria which will concentrate on specific parts of the distribution by assigning appropriate weights. Thus, there is much room for improvement and research on this promising concept.

Bibliography

[Akaike, 1973] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle,[w:] proceedings of the 2nd international symposium on information, bn petrow, f. *Czaki, Akademiai Kiado, Budapest*.

[Ali and Silvey, 1966] Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142.

[Avlogiaris et al., 2016] Avlogiaris, G., Micheas, A., and Zografos, K. (2016). On local divergences between two probability measures. *Metrika*, 79(3):303–333.

[Balakrishnan and Sanghvi, 1968] Balakrishnan, V. and Sanghvi, L. (1968). Distance between populations on the basis of attribute data. *Biometrics*, pages 859–865.

[Barbu et al., 2018] Barbu, V. S., Karagrigoriou, A., and Preda, V. (2018). Entropy and divergence rates for markov chains: 2. the weighted case. *Proceedings of the Romanian Academy, Series A*, 18(4):293–301.

[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.

[Basu et al., 2011] Basu, A., Shioya, H., and Park, C. (2011). *Statistical inference: the minimum distance approach*. Chapman and Hall/CRC.

[Belis and Guiasu, 1968] Belis, M. and Guiasu, S. (1968). A quantitative-qualitative measure of information in cybernetic systems (corresp.). *IEEE Transactions on Information Theory*, 14(4):593–594.

[Bhattacharya, 1943] Bhattacharya, K. (1943). A note on two-fold triple systems. *Sankhyā: The Indian Journal of Statistics*, pages 313–314.

[Birch, 1964] Birch, M. (1964). A new proof of the pearson-fisher theorem. *The Annals of Mathematical Statistics*, pages 817–824.

[Choi and Bulgren, 1968] Choi, K. and Bulgren, W. (1968). An estimation procedure for mixtures of distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(3):444–460.

[Cressie and Pardo, 2002] Cressie, N. and Pardo, L. (2002). Model checking in loglinear models using φ-divergences and mles. *Journal of Statistical Planning and Inference*, 103(1-2):437–453.

[Cressie and Read, 1984] Cressie, N. and Read, T. R. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(3):440–464.

[Di Crescenzo and Longobardi, 2007] Di Crescenzo, A. and Longobardi, M. (2007). On weighted residual and past entropies. *arXiv preprint math/0703489*.

[Fisher, 1925] Fisher, R. A. (1925). Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 22, pages 700–725. Cambridge University Press.

[Gokhale and Kullback, 1978] Gokhale, D. V. and Kullback, S. (1978). *The information in contingency tables*, volume 23. M. Dekker.

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

[Guiaşu, 1986] Guiaşu, S. (1986). Grouping data by using the weighted entropy. *Journal of Statistical Planning and Inference*, 15:63–69.

[Halmos, 2013] Halmos, P. R. (2013). *Measure theory*, volume 18. Springer.

[Hausdorff, 1914] Hausdorff, F. (1914). Bemerkung über den inhalt von punktmengen. *Mathematische Annalen*, 75(3):428–433.

[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*.

[Jones et al., 2001] Jones, M., Hjort, N. L., Harris, I. R., and Basu, A. (2001). A comparison of related density-based minimum divergence estimators. *Biometrika*, 88(3):865–873.

[Kagan, 1963] Kagan, A. (1963). On the theory of fisher's amount of information. In *Sov. Math. Dokl*, volume 4, pages 991–993.

[Kapur, 1994] Kapur, J. N. (1994). *Measures of information and their applications*. Wiley-Interscience.

[Kolmogorov, 1933] Kolmogorov, A. (1933). Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4:83–91.

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

[Kupperman et al., 1958] Kupperman, M. et al. (1958). Probabilities of hypotheses and information-statistics in sampling from exponential-class populations. *The Annals of Mathematical Statistics*, 29(2):571–575.

[Le Cam, 1990] Le Cam, L. (1990). Maximum likelihood: an introduction. *International Statistical Review/Revue Internationale de Statistique*, pages 153–171.

[Lin, 1991] Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

[Macdonald, 1971] Macdonald, P. (1971). Comments and queries comment on "an estimation procedure for mixtures of distributions" by choi and bulgren. *Journal of the Royal Statistical Society: Series B (Methodological)*, 33(2):326–329.

[Mahalanobis, 1936] Mahalanobis, P. C. (1936). On the generalized distance in statistics. volume 2, pages 49–55. 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.

[Mathai and Rathie, 1975] Mathai, A. M. and Rathie, P. (1975). *Basic concepts information theory statistics: axiomatic foundations and aplication*. Number Q 360. M37.

[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.

[Matusita, 1964] Matusita, K. (1964). Distance and decision rules. *Annals of the Institute of Statistical Mathematics*, 16(1):305–315.

[Menéndez et al., 1997] Menéndez, M., Salicrú, M., Morales, D., and Pardo, L. (1997). Divergence measures between populations: applications in the exponential family. *Communications in Statistics-Theory and Methods*, 26(5):1099–1117.

[Morales et al., 1995] Morales, D., Pardo, L., and Vajda, I. (1995). Asymptotic divergence of estimates of discrete distributions. *Journal of Statistical Planning and Inference*, 48(3):347–369.

[Morimoto, 1963] Morimoto, T. (1963). Markov processes and the h-theorem. *Journal of the Physical Society of Japan*, 18(3):328–331.

[Nawrocki and Harding, 1986] Nawrocki, D. N. and Harding, W. H. (1986). State-value weighted entropy as a measure of investment risk. *Applied Economics*, 18(4):411–419.

[Nayak, 1983] Nayak, T. K. (1983). *Applications of entropy functions in measurement and analysis of diversity*. PhD thesis, University of Pittsburgh.

[Pak and Basu, 1998] Pak, R. J. and Basu, A. (1998). Minimum disparity estimation in linear regression models: Distribution and efficiency. *Annals of the Institute of Statistical Mathematics*, 50(3):503–521.

[Pardo, 2018] Pardo, L. (2018). *Statistical inference based on divergence measures*. Chapman and Hall/CRC.

[Pearson, 1900] Pearson, K. (1900). X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

[Pinsker, 1960] Pinsker, M. S. (1960). Dynamical systems with completely positive or zero entropy. In *Doklady Akademii Nauk*, volume 133, pages 1025–1026. Russian Academy of Sciences.

[Polyanskiy and Wu, 2014] Polyanskiy, Y. and Wu, Y. (2014). Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC)*, 6(2012-2016):7.

[Quandt and Ramsey, 1978] Quandt, R. E. and Ramsey, J. B. (1978). Estimating mixtures of normal distributions and switching regressions. *Journal of the American statistical Association*, 73(364):730–738.

[Rathie and Kannappan, 1972] Rathie, P. N. and Kannappan, P. (1972). A directed-divergence function of type β. *Information and Control*, 20(1):38–45.

[Rukhin, 1994] Rukhin, A. (1994). Optimal estimator for the mixture parameter by the method of moments and information affinity, in trans. In *12the Prague Conference on Information Theory*, pages 214–219.

[Saks, 1937] Saks, S. (1937). Theory of the integral.

[Salicrú et al., 1994] Salicrú, M., Morales, D., Menéndez, M., and Pardo, L. (1994). On the applications of divergence type measures in testing statistical hypotheses. *Journal of Multivariate Analysis*, 51(2):372–391.

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

[Simon, 1973] Simon, G. (1973). Additivity of information in exponential family probability laws. *Journal of the American Statistical Association*, 68(342):478–482.

[Smirnov, 1948] Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, 19(2):279–281.

[Suhov and Zohren, 2014] Suhov, Y. and Zohren, S. (2014). Quantum weighted entropy and its properties. *arXiv preprint arXiv:1411.0892*.

[Taneja, 1998] Taneja, I. (1998). On generalized weighted divergence measures.

[Toma et al., 2019] Toma, A., Karagrigoriou, A., and Trentou, P. (2019). Robust criteria for model selection using pseudodistances. *(submited)*.

[Vajda, 1973] Vajda, I. (1973). $\chi\alpha$-divergence and generalized fisher information. In *Transactions of the Sixth Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, page 223. Academia.

[Vajda, 1989] Vajda, I. (1989). *Theory of statistical inference and information*, volume 11. Kluwer Academic Pub.

[Verdu, 1998] Verdu, S. (1998). Fifty years of shannon theory. *IEEE Transactions on information theory*, 44(6):2057–2078.

[Wolfowitz et al., 1957] Wolfowitz, J. et al. (1957). The minimum distance method. *The Annals of Mathematical Statistics*, 28(1):75–88.

[Zografos et al., 1990] Zografos, K., Ferentinos, K., and Papaioannou, T. (1990). Divergence statistics: sampling properties and multinomial goodness of fit and divergence tests. *Communications in Statistics-Theory and Methods*, 19(5):1785–1802.

Appendices

Appendix A

A.1

Measure Theory material §3.1

Definition 21. *(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. $\emptyset, \ X \in \mathcal{T}$
2. *if $\{U_a \in \mathcal{T} : a \in I\}$ is an arbitrary collection of open sets, 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 22. *(σ-Algebra)*

A σ-algebra (or field) on a set X is a collection $\mathcal{A}$ 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 23. *(Generated* σ*-Algebra)*

If $\mathcal{F}$ *is any collection of subsets of a set* X*, then the* σ*-algebra generated by* $\mathcal{F}$ *is*

$$\sigma(\mathcal{F}) = \bigcap \{\mathcal{A} \subset \mathcal{P}(X) : \mathcal{F} \subseteq \mathcal{A} \ \textit{ and } \ \mathcal{A} \ \textit{ is a } \ \sigma - \textit{algebra}\}.$$

This intersection is nonempty, since $\mathcal{P}(X)$ *is a* σ*-algebra that contains* $\mathcal{F}$*, and an intersection of* σ*-algebras is a* σ*-algebra. An immediate consequence of the definition is the following result, which we will use repeatedly.*

Definition 24. *(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 25. *(Measurable Space)*

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

Definition 26. *(Measure)*

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

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

such that

1. $\mu(\emptyset) = 0$

2. 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 27. *(Measure Space)*

A measure space $(X, \mathcal{A}, \mu)$ *consists of a set* X*, a* σ*-algebra* $\mathcal{A}$ *on* X*, and a measure* μ *defined on* $\mathcal{A}$*.*

From now on we will concentrate on $\mathbb{R}^n$ with the Euclidean norm. Whenever we are referring to the standard topology of $\mathbb{R}^n$ we will be referring to the topology induced by the Euclidean metric which by itself is induced by the Euclidean norm. So $\mathcal{T}(\mathbb{R}^n) \subsetneq \mathcal{P}(\mathbb{R}^n)$ we denote from now on the standard topology on $\mathbb{R}^n$. That is, $G \subseteq \mathbb{R}^n$ belongs to $\mathcal{T}(\mathbb{R}^n)$ if for every $x \in G$ there exists $r > 0$ such that $B_r(x) \subseteq G$, where

$$B_r(x) = \{\psi \in \mathbb{R}^n : ||x - \psi|| < r\}$$

is the open ball of radius r centered at $x \in \mathbb{R}_n$ and $|| \cdot ||$ denotes the Euclidean norm.

Definition 28. *(Borel Set)*

The Borel σ*-algebra* $\mathcal{B}(\mathbb{R}^n)$ *on* $\mathbb{R}^n$ *is the* σ*-algebra generated by the open sets of* $\mathbb{R}^n$*, i.e.* $\mathcal{B}(\mathbb{R}^n) = \sigma(\mathcal{T}(\mathbb{R}^n))$*. A set that belongs to the Borel* σ*-algebra is called a Borel set.*

At this point our goal should be to construct a notion of the volume of rather general subsets of $\mathbb{R}^n$ that reduces to the usual volume of elementary geometrical sets such as cubes or rectangles. These generalized notion of volume is the Lebesgue measure.

If $\mathcal{L}(\mathbb{R}^n)$ denotes the collection of Lebesgue measurable sets and

$$\mu : \mathcal{L}(\mathbb{R}^n) \to [0, \infty]$$

denotes Lebesgue measure, then we want $\mathcal{L}(\mathbb{R}^n)$ to contain all n-dimensional rectangles and $\mu(R)$ should be the usual volume of a rectangle R. Moreover, we want μ to be countably additive. That is, if

$$\{A_i \in \mathcal{L}(\mathbb{R}^n) : i \in \mathbb{N}\}$$

is a countable collection of disjoint measurable sets, then their union should be measurable and

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

The reason for requiring countable additivity is that finite additivity is too weak to allow the justification of any limiting processes, while uncountable additivity is too

strong; for example, it would imply that if the measure of a set consisting of a single point is zero, then the measure of every subset of $\mathbb{R}^n$ would be zero.

It is not possible to define the Lebesgue measure of all subsets of $\mathbb{R}^n$ in a geometrically reasonable way. Hausdorff (1914) showed that for any dimension $n \geq 1$, there is no countably additive measure defined on all subsets of R^n that is invariant under isometries (translations and rotations) and assigns measure one to the unit cube. We will skip any additional theoretical background needed to strictly define Lebesgue outer measures and we will just provide two simple cases which will be needed later on.

Proposition 7. *Every rectangle is Lebesgue measurable and its measure is the volume.*

Before continuing we briefly consider a generalization of one-dimensional Lebesgue measure, called Lebesgue-Stieltjes measures on $\mathbb{R}$. These measures are obtained from an increasing, right-continuous function $F : \mathbb{R} \to \mathbb{R}$, and assign to a half-open interval $(a, b]$ the measure

$$\mu_F((a, b]) = F(b) - F(a).$$

Theorem 10. *Suppose that $F : \mathbb{R} \to \mathbb{R}$ is an increasing, right-continuous function. Then there is a unique Borel measure $\mu_F : \mathcal{B}(\mathbb{R}) \to [0, \infty]$ such that*

$$\mu_F((a, b]) = F(b) - F(a)$$

for every $a < b$.

Definition 29. *(Measurable Function)*

Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces. A function $f : X \to Y$ is measurable if $f^{-1}(B) \in \mathcal{A}$ for every $B \in \mathcal{B}$.

In the following part we will briefly present Lebesgue integrability. We will not dive in the full depth of it, since we are more concerned with Riemann integrability.

Definition 30. *(Characteristic Function)*

The characteristic function (or indicator function) of a subset $E \subset X$ is the function $\chi_E : X \to \mathbb{R}$ defined by

$$\chi_E(x) = \begin{cases} 1, & x \in E \\ 0, & x \notin E \end{cases}$$

The function χ_E is measurable if and only if E is a measurable set.

Definition 31. *(Simple Function)*

A simple function $\phi : X \to \mathbb{R}$ on a measurable space $(X, \mathcal{A})$ is a function of the form

$$\phi(x) = \sum_{n=1}^{N} c_n \chi_{E_n}(x)$$

where $c_1, ..., c_N \in \mathbb{R}$ and $E_1, ..., E_N \in \mathcal{A}$.

Note that, according to this definition, a simple function is measurable. The representation of ϕ is not unique, we call it a standard representation if the constants c_n are distinct and the sets E_n are disjoint.

Definition 32. *(Positive Simple Function)*

If $\phi : X \to [0, \infty)$ is a positive simple function, given by

$$\phi = \sum_{i=1}^{N} c_i \chi_{E_i}$$

where $c_i \geq 0$ and $E_i \in \mathcal{A}$, then the integral of ϕ with respect to μ is

$$\int \phi d\mu = \sum_{i=1}^{N} c_i \mu(E_i).$$

Definition 33. *(Lebesgue Integral)*

If $f : X \to [0, \infty]$ is a positive, measurable, extended (it can take infinity as a value) real-valued function on a measure space X, then:

$$\int f d\mu = \sup \left\{ \int \phi d\mu : 0 \leq \phi \leq f, \ \phi \ simple \right\}$$

A positive function $f : X \to [0, \infty]$ is integrable if it is measurable and

$$\int f d\mu < \infty$$

Proposition 8. *If $A \subset X$ is a measurable set and $f : X \to [0, \infty]$ is measurable, we define*

$$\int_A f d\mu = \int f \chi_A d\mu.$$

Proposition 9. *Suppose that $X = [a, b] \subsetneq \mathbb{R}$ is a compact interval and $\mu : \mathcal{L}([a, b]) \to [0, \infty]$ is Lesbegue measure on $[a, b]$. We note that any Riemann integrable function $f : [a, b] \to \mathbb{R}$ is integrable with respect to Lebesgue measure μ, and its Riemann integral is equal to the Lebesgue integral,*

$$\int_a^b f(x)dx = \int_{[a,b]} f d\mu$$

Thus, all of the usual integrals from elementary calculus remain valid for the Lebesgue integral on $\mathbb{R}$. We will write an integral with respect to Lebesgue measure on $\mathbb{R}$, or $\mathbb{R}^n$, as

$$\int f dx$$

Note :*Even though the class of Lebesgue integrable functions on an interval is wider than the class of Riemann integrable functions, some improper Riemann integrals may exist even though the Lebesgue integral does not.*

A.2

Probability Spaces material §3.2

Definition 34. *(Probability Measure)*

If $(\Omega, \mathcal{F})$ is a measurable space and μ is a measure on $(\Omega, \mathcal{F})$, then with the additional condition of $\mu(\Omega) = 1$ we have a probability measure and we often label it by P (it is then straight forward that $P(A) \leq 1$ for all $A \in \mathcal{F}$).

Definition 35. *(Probability Space)*

A measure space is a triplet $(\Omega, \mathcal{F}, \mu)$, with μ a measure on the measurable space $(\Omega, \mathcal{F})$. A measure space $(\Omega, \mathcal{F}, P)$ with P a probability measure is called a probability space.

Definition 36. *(Random Variable)*

A function $X : \Omega \to \mathbb{S}$ between two measurable spaces $(\Omega, \mathcal{F})$ and $(\mathbb{S}, S)$ is called an $(\mathbb{S}, S)$-valued Random Variable if

$$X^{-1}(B) := \{\omega : X(\omega) \in B\} \in \mathcal{F} \quad \forall B \in S$$

Hence, a random variable is a measurable function. A measurable function $X : \Omega \to \mathbb{S}$ is also called a random variable in $\mathbb{S}$. It has the interpretation of a quantity, or state, determined by chance. Where no space $\mathbb{S}$ is mentioned, it is assumed that X takes values in $\mathbb{R}$. Even though the first measurable space $(\Omega, \mathcal{F})$ doesn't need to be a probability space but we prefer to define random variables in probability spaces $(\Omega, \mathcal{F}, P)$. Hence, the following definition.

Definition 37. *(Random Variable)*

A random variable on a probability space $(\Omega, \mathcal{F}, P)$ is a function $X : \Omega \to \mathbb{R}$ that is measurable with respect to the Borel sets.

Definition 38. *(Distribution of a Random Variable)*

The measure $\mu_X = P \circ X^{-1}$ is called the law or distribution of X. For real-valued random variables, μ_X is uniquely determined by its values on the intervals $(-\infty, x]$, $x \in \mathbb{R}$, given by

$$F_X(x) = \mu_X((-\infty, x]) = P(X \leq x)$$

The function F_X is called the distribution function of X. Generalizing the above for any Borel set we get the following definition.

Definition 39. *(Distribution)*

Every random variable X induces a probability measure μ_X on $\mathbb{R}$ (called its distribution) by

$$\mu_X(A) = P(X^{-1}(A))$$

for all $A \in \mathcal{B}(\mathbb{R})$

To check that μ_X is a probability measure, note that since X is a function, if $A_1, A_2, ... \in \mathcal{B}(\mathbb{R})$ are disjoint, then so are $X \in A_1, X \in A_2, ... \in F$, hence

$$\mu_X\left(\bigcup_i A_i\right) = P\left(\{X \in \bigcup_i A_i\}\right) = P\left(\bigcup_i \{X \in A_i\}\right) = \sum_i P\left(\{X \in A_i\}\right) = \sum_i \mu_X(A_i)$$

Theorem 11. *If F is the distribution function of a random variable X, then*
i)F is nondecreasing.
ii)F is right-continuous (i.e. $\lim_{x \to a^+} F(x) = F(a)$ for all $a \in \mathbb{R}$).
iii) $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.
iv)If $F(x^-) = \lim_{y \to x^-} F(y)$ then $F(x^-) = P(X < x)$.
v)$P(X = x) = F(x) - F(x^-)$

Proof. For *i*), note that if $x \leq y$, then $\{X \leq x\} \subseteq \{X \leq y\}$, so $F(x) = P(X \leq x) \leq P(X \leq y) = F(y)$ by monotonicity of the probability measure.
For *ii*), observe that if $x \searrow a$, then $\{X \leq x\} \searrow \{X \leq a\}$, and apply continuity of the probability measure we have desired result.
For *iii*), we have $\{X \leq x\} \searrow \emptyset$ as $x \searrow -\infty$ and $\{X \leq x\} \nearrow \mathbb{R}$ as $x \nearrow \infty$.
For *iv*), $\{X \leq y\} \nearrow \{X < x\}$ as $y \nearrow x$. (Note that the limit exists since F is monotone.)
For *v*), $\{X = x\} = \{X \leq x\} \setminus \{X < x\}$. □

In fact, the first three properties in Theorem 11 are sufficient to characterize a distribution function.

Theorem 12. *If $F : \mathbb{R} \to \mathbb{R}$ satisfies properties i), ii), and iii) from Theorem 11, then it is the distribution function of some random variable.*

Theorem 12 shows that any function satisfying properties *i*) - *iii*) gives rise to a random variable X, and thus to a probability measure μ, the distribution of X. The following result shows that the measure is uniquely determined.

Theorem 13. *If F is a function satisfying $i)-ii)$ in Theorem 10, then there is a unique probability measure μ on $(\mathbb{R},\mathcal{B})$ with $\mu((-\infty,x])=F(x)$ for all $x\in\mathbb{R}$ (as mentioned in definitions 38,39).*

To summarize, every random variable induces a probability measure on $(R,\mathcal{B})$, every probability measure defines a function satisfying properties $i)-iii)$ in Theorem 11, and every such function uniquely determines a probability measure.

Consequently, it is equivalent to give the distribution or the distribution function of a random variable. However, one should be aware that distributions/distribution functions do not determine random variables, even neglecting differences on null sets. For example, if X is uniform on $[-1,1]$ (so that $\mu_X=\frac{1}{2}\mu|_{[-1,1]}$) , then $-X$ also has distribution μ_X, but $-X\neq X$ almost surely. When two random variables X and Y have the same distribution function, we say that they are equal in distribution and write $X\overset{d}{=}Y$. Note that random variables can be equal in distribution even if they are defined on different probability spaces.

Definition 40. *(Probability Density Function of a Random Variable)*

We say that a random variable $X(\omega)$ has a probability density function f_X if and only if its distribution function F_X can be expressed as

$$F_X(a)=\int_{-\infty}^{a}f_X(x)dx,\ \ \forall a\in\mathbb{R}$$

Remark 5. *To make Definition precise we temporarily assume that probability density functions f_X are Riemann integrable and interpret the integral in this sense. We construct Lebesgue's integral and extend the scope of Definition 40 to Lebesgue integrable density functions $f_X\geq 0$ (in particular, accommodating Borel functions f_X).*

Remark 6. *Conversely, if g is a nonnegative measurable function with $\int_{\mathbb{R}}g(x)dx=1$, then $G(x)=\int_{-\infty}^{x}g(t)dt$, satisfies $i)-iii)$ in Theorem 11, so Theorem 12 gives a random variable with density g. In undergraduate probability, such a X is called continuous. Of course this cannot be used as a strict definition. Hence the Definition is provided bellow.*

Definition 41. *If the distribution of X has a density, then we say that X is absolutely continuous.*

A.3

Classification of Random Variables material §3.3

Definition 42. *(Absolute Continuity of Measures)*

If μ and v are measures on measure space $(X, \mathcal{A})$, then we say that v is absolutely continuous with respect to μ (and write $v \ll \mu$) if $v(A) = 0$ for all $A \in \mathcal{A}$ with $\mu(A) = 0$.

Definition 43. *(Mutual Singularity of Measures)*

If μ and v are measures on $(X, \mathcal{A})$, then we say that μ and v are mutually singular (and write $\mu \perp v$) if there exist $E, F \in \mathcal{A}$ such that
i) $E \cap F = \emptyset$
ii) $E \cup F = X$
iii) $\mu(F) = 0 = v(E)$

A fundamental result in measure theory is the Lebesgue-Radon-Nikodym Theorem (which we state only for positive measures).

Theorem 14. *(Lebesgue-Radon-Nikodym)*

If μ and v are σ-finite measures on $(X, \mathcal{A})$, then there exist unique σ-finite measures λ, ρ on $(X, \mathcal{A})$ such that

1. $\lambda \perp \mu$
2. $\rho \ll \mu$
3. $v = \lambda + \rho$

Moreover, there is a measurable function $f : X \to [0, \infty)$ such that $\rho(A) = \int_A f d\mu$ for all $A \in \mathcal{A}$.

Remark 7. *With σ-finite measures on $(X, \mathcal{A})$ we mean that $\mu_X < \infty$ and $v_X < \infty$.*

Note that :

- The function f from Theorem 14 is called the Radon-Nikodym derivative of ρ with respect to μ, and one writes $f = \frac{d\rho}{d\mu}$ (or $d\rho = f d\mu$).
- If v is a finite measure, then λ and ρ are finite, so f is μ-integrable.

- If a random variable X has distribution μ which is absolutely continuous with respect to Lebesgue measure, then we say that (the distribution of) X has density function $f = \frac{d\mu}{dm}$, where m is the Lebesgue measure.

Remark 8. *Based on the above and Definitions 38, 39 we have that for all* $A \in \mathcal{B}_r$, $P(X \in A) = \mu(A) = \int_A f(x)dx$.

Remark 9. *For* f *integrable and* $A \in \mathcal{F}$, *we define the integral of* f *over* A *as* $\int_A f d\mu = \int f I_A d\mu$. *When we do not wish to emphasize the dependence on the argument, we write* $\int f d\mu$, *or sometimes* $\int f(x)\mu(dx)$.

Proposition 10. *For any* $a, b \in \mathbb{R}$ *and any integrable functions* f, g, $\int (af + bg)d\mu = a\int f d\mu + b\int g d\mu$. *If* $f \leq g$ *a.e., then* $\int f d\mu \leq \int g d\mu$.

Definition 44. *(Discrete Measure)*

A measure μ *is said to be discrete if there is a countable set* $\mathbb{S}$ *with* $\mu(\mathbb{S}^C) = 0$. *A random variable, is called discrete if its distribution is.*

Note that :

- if X is discrete, then $\mu \perp m$.
- The set $\mathbb{S}$ in the above definition plays the role of support

Definition 45. *(Continuous Measure)*

A measure μ *is called continuous if* $\mu(\{x\}) = 0$ *for all* $x \in \mathbb{R}$.

Remark 10. *By countable additivity, a discrete probability measure is not continuous and vice versa. Absolutely continuous distributions are continuous, but it is possible for a continuous distribution to be singular with respect to Lebesgue measure.*

Definition 46. *(Singular Continuity of a Random Variable)*

A random variable X with continuous distribution $\mu \perp m$ is called singular continuous.

Analogous to the singular/absolutely continuous decomposition in the Theorem 14, we have the following result for finite Borel measures on $\mathbb{R}$.

Remark 11. *An example is given by the 'uniform distribution on the Cantor set' formed by taking* $[0, 1]$ *and successively removing the open middle third of all remaining intervals.*

Theorem 15. *Any finite Borel measure can be uniquely written as*

$$\mu = \mu_d + \mu_c$$

where μ_d is discrete and μ_c is continuous.

Proof. Let $E = \{x \in \mathbb{R} : \mu(\{x\}) > 0\}$.

For any countable $F \subseteq E$, $\sum_{x \in F} \mu(\{x\}) = \mu(F) < \infty$ by countable additivity and finiteness.

It follows that $E_k = \{x \in \mathbb{R} : \mu(\{x\}) > k^{-1}\}$ is finite for all $k \in \mathbb{N}$.

Consequently, $E = \bigcup_{k=1}^{\infty} E_k$ is a countable union of finite sets and thus is countable.

The result follows by defining $\mu_d(A) = \mu(A \cap E), \;\; \mu_c(A) = \mu(A \cap E^c)$. □

Remark 12. *Thus if μ is a probability distribution, then it follows from the Radon-Nikodym Theorem that $\mu = \mu_{ac} + \mu_s$ where $\mu_{ac} << m$ and $\mu_s \perp m$. By Theorem 14, $\mu_s = \mu_d + \mu_{sc}$ where μ_d is discrete and μ_{sc} is singular continuous. Since μ is a probability measure, each of μ_{ac}, μ_d, μ_{sc} is finite and thus is identically zero or a multiple of a probability measure.*

Appendix B

B.1

Proof. Proposition 3, §4.2.1 *(Vajda, 1995)*

Using the non-negativity of the function $\psi(x) \equiv \phi(x) - \phi^{'}(1)(x-1)$, we have $D_{\psi}(\theta_1, \theta_2) \geq 0$, but we know that $D_{\phi}(\theta_1, \theta_2) = D_{\psi}(\theta_1, \theta_2)$, then $D_{\phi}(\theta_1, \theta_2) \geq 0$.

It is known that for every convex function ϕ the following inequality holds

$$\phi(t) \leq \phi(0) + t \lim_{r \to \infty} \frac{\phi(r)}{r}, \quad (t \geq 0) \tag{B.1}$$

If ϕ is strictly convex at some $t_0 \in (0, \infty)$ then the inequality in $(B.1)$ is strict for all $t > 0$. Using $(B.1)$ we have

$$D_{\phi}(\theta_1, \theta_2) \leq \int_X f_{\theta_2}(x) \left(\phi(0) + \frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} \lim_{r \to \infty} \frac{\phi(r)}{r} \right) d\mu(x) = \phi(0) + t \lim_{r \to \infty} \frac{\phi(r)}{r}.$$

It is clear that $P_{\theta_1} = P_{\theta_2}$ implies $D_{\phi}(\theta_1, \theta_2) = 0$.

If $S_1 \cap S_2 = \emptyset$, we have

$$D_{\phi}(\theta_1, \theta_2) = \int_X f_{\theta_2}(x) \phi \left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} \right) d\mu(x) =$$

$$= \int_{S_1^c \cap S_2} f_{\theta_2}(x) \phi \left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} \right) d\mu(x) + \int_{S_1 \cap S_2^c} f_{\theta_2}(x) \phi \left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} \right) d\mu(x) =$$

$$= \phi(0) + t \lim_{r \to \infty} \frac{\phi(r)}{r}.$$

Now we are going to establish that if ϕ is strictly convex at $t = 1$, then $D_{\phi}(\theta_1, \theta_2) = 0$ implies $P_{\theta_1} = P_{\theta_2}$.

In fact, if ϕ is strictly convex at $t = 1$ then

$$\psi\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right) > 0$$

for $\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} > 1$ and for $\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} < 1$. If $D_\psi(\theta_1, \theta_2) = 0$ then $\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} \leq 1$ or $\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} \geq 1$. First we suppose that $\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} \leq 1$. We know that

$$D_\phi(\theta_1, \theta_2) = D_\psi(\theta_1, \theta_2) = 0$$

and

$$0 = D_\psi(\theta_1, \theta_2) = \int_X f_{\theta_2}(x)\psi\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right) d\mu(x) =$$

$$= \int_X f_{\theta_2}(x)\left(\phi\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right) - \phi^{'}(1)\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} - 1\right)\right) d\mu(x) =$$

$$= D_\phi(\theta_1, \theta_2) - \phi^{'}(1)\int_X f_{\theta_2}(x)\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} - 1\right) d\mu(x) =$$

$$= 0 - \phi^{'}(1)\int_X f_{\theta_2}(x)\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} - 1\right) d\mu(x) =$$

$$= -\phi^{'}(1)\int_X f_{\theta_2}(x)\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} - 1\right) dP_{\theta_2}.$$

Since ϕ is strictly convex at $t = 1$, it must be $P_{\theta_1} = P_{\theta_2}$. For $\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} \geq 1$, the result can be established in the same way.

The strict convexity of ϕ at $t = 1$ implies the strict inequality in $(B.1)$, i.e.,

$$\phi(t) < \phi(0) + t \lim_{r\to\infty} \frac{\phi(r)}{r}, \quad \forall t > 0.$$

Which implies that

$$l(t) = \phi(0) - \phi(t) + t \lim_{r\to\infty} \frac{\phi(r)}{r} > 0, \quad \forall t > 0.$$

If we take $x \in S_1$, i.e., x such that $f_{\theta_1} > 0$, then $t = \frac{f_{\theta_1}(x)}{f_{\theta_2(x)}} > 0$ and $l\left(\frac{f_{\theta_1}(x)}{f_{\theta_2(x)}}\right) > 0$.

Therefore,

$$D_l(\theta_1,\theta_2) = \int_X f_{\theta_2}(x) l\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right) d\mu(x) =$$

$$= \int_X f_{\theta_2(x)}\left(\phi(0) - \phi\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right) + \frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} \lim_{r\to\infty}\frac{\phi(r)}{r}\right) d\mu(x) =$$

$$= -D_\phi(\theta_1,\theta_2) + \phi(0) + \lim_{r\to\infty}\frac{\phi(r)}{r},$$

but by (4.6) we have

$$D_\phi(\theta_1,\theta_2) = \phi(0) + \lim_{r\to\infty}\frac{\phi(r)}{r}$$

therefore,

$$D_l(\theta_1,\theta_2) = \int_X f_{\theta_2}(x) l\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right) d\mu(x) = 0$$

with

$$l\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right) > 0$$

Then, $f_{\theta_2}(x) = 0$, because $D_l(\theta_1,\theta_2) = 0$ and $l\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right) > 0$, i.e., $x \notin S_2$. This completes the proof. □

B.2

Proof. Proposition 5, §4.2.1 *(Vajda, 1995)*

We assume that μ is the Lebesgue measure. We define

$$\tilde{D}_\phi(\theta_1, \theta_2) = \int_{\mathbb{R}} f_{\theta_1}(x)\phi\left(\frac{f_{\theta_2}(x)}{f_{\theta_1}(x)}\right) dx,$$

and we shall establish

$$\tilde{D}_\phi(\theta_1, \theta_2) \leq \tilde{D}_\phi(\theta_1, \theta_3),\ \phi \in \Phi^* \tag{B.2}$$

If $(B.2)$ holds, then (4.10) also holds, because if we consider the function

$$\phi(t) = t\phi(\frac{1}{t}) \in \Phi^*$$

we have

$$\tilde{D}_\phi(\theta_1, \theta_2) = \int_{\mathbb{R}} f_{\theta_1}(x)\frac{f_{\theta_2}(x)}{f_{\theta_1}(x)}\phi\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right) dx = D_\phi(\theta_1, \theta_2).$$

Since, by hypothesis, the family of distributions $\{P_\theta\}_{\theta\in\Theta\subset\mathbb{R}}$ has monotone non decreasing likelihood ratio, then

$$h_2(x) = \frac{f_{\theta_2}(x)}{f_{\theta_1}(x)} \ \ and \ \ h_3(x) = \frac{f_{\theta_3}(x)}{f_{\theta_1}(x)}$$

are non decreasing functions of x. The same happens with

$$\frac{h_3(x)}{h_2(x)} = \frac{f_{\theta_3}(x)}{f_{\theta_2}(x)} \tag{B.3}$$

From $(B.3)$ we recognize three possibilities:

1. $h_3(x) < h_2(x),\ \forall x$
2. $h_3(x) > h_2(x),\ \forall x$
3. $\exists a$ s.t. $h_3(x) \leq h_2(x)$ for $x < a$ and $h_3(x) \geq h_2(x)$ for $x > a$.

We know that

$$E_{\theta_1}[h_3(X)] = \int_{\mathbb{R}} f_{\theta_1}(x)\frac{f_{\theta_3}(x)}{f_{\theta_1}(x)}dx = E_{\theta_1}[h_2(X)] = 1.$$

If $E_{\theta_1}[h_3(X)] = E_{\theta_1}[h_2(X)] = 1$, the possibilities (1) and (2) are not true, hence it should be true (3). Using the monotonicity of $h_2(x)$ and $h_3(x)$ we have

$$x : h_2(x) \leq b \subset x : h_3(x) \leq b, \;\; if \;\; b < h_2(a).$$

and

$$x : h_2(x) \leq b \supset x : h_3(x) \leq b, \;\; if \;\; b < h_2(a)$$

If we denote

$$F_{h_2(X)}(t) = Pr_{\theta_1}(h_2(X) \leq t) = Pr_{\theta_1}(x \in \mathbb{R} : h_2(x) \leq t)$$

$$F_{h_3(X)}(t) = Pr_{\theta_1}(h_3(X) \leq t) = Pr_{\theta_1}(x \in \mathbb{R} : h_3(x) \leq t)$$

we have for $t < h_2(a)$

$$F_{h_2(X)}(t) = Pr_{\theta_1}(x \in \mathbb{R} : h_2(x) \leq t) \leq Pr_{\theta_1}(x \in \mathbb{R} : h_3(x) \leq t) = F_{h_3(X)}(t),$$

and for $t > h_2(a)$

$$F_{h_2(X)}(t) \geq F_{h_3(X)}(t).$$

Now we shall establish that the statements

1. $E_{\theta_1}[h_3(X)] = E_{\theta_1}[h_2(X)]$
2. $F_{h_2(X)}(t) \leq F_{h_3(X)}(t)$ for $t < h_2(a)$ and $F_{h_2(X)}(t) \geq F_{h_3(X)}(t)$ for $t > h_2(a)$

imply $\forall k$,

$$E_{\theta_1}[|h_2(X) - k|] \leq E_{\theta_1}[|h_3(X) - k|]. \tag{B.4}$$

It is well known that the expectation of a non negative random variable X can be written as

$$E[X] = \int_0^\infty (1 - F_X(x))dx.$$

In our case,

$$E_{\theta_1}[h_3(X)] = \int_0^\infty (1 - F_{h_3(X)}(x))dx = \int_0^\infty (1 - F_{h_2(X)}(x))dx = E_{\theta_1}[h_2(X)].$$

Denoting

$$I_1 \equiv \int_0^{h_2(a)} (1 - F_{h_3(X)}(x)) - (1 - F_{h_2(X)}(x))dx$$

and

$$I_2 \equiv \int_{h_2(a)}^\infty (1 - F_{h_3(X)}(x)) - (1 - F_{h_2(X)}(x))dx$$

we have

$$I_1 = \int_0^{h_2(a)} (F_{h_2(X)}(x) - F_{h_3(X)})dx,\ I_2 = \int_{h_2(a)}^\infty (F_{h_2(X)}(x) - F_{h_3(X)})dx$$

Therefore,

$$E_{\theta_1}[h_3(X)] - E_{\theta_1}[h_2(X)] = \int_0^{h_2(a)} (F_{h_2(X)}(x) - F_{h_3(X)})dx + \int_{h_2(a)}^\infty (F_{h_2(X)}(x) - F_{h_3(X)})dx.$$

Finally, we have

$$\int_0^{h_2(a)} (F_{h_2(X)}(x) - F_{h_3(X)})dx = \int_{h_2(a)}^\infty (F_{h_3(X)}(x) - F_{h_2(X)})dx. \tag{B.5}$$

Now we prove $(B.4)$. It is easy to check that

$$E_{\theta_1}[|h_i(X) - k|] = \int_0^k F_{h_i(X)}(x)dx + \int_k^\infty (1 - F_{h_i(X)}(x))dx.$$

Assuming that $k \geq h_2(a)$, an analogous proof can be done if $k < h_2(a)$; we have

$$E_{\theta_1}[|h_i(X) - k|] = \int_0^{h_2(a)} F_{h_i(X)}(x)dx + \int_{h_2(a)}^k F_{h_i(X)}(x)dx + \int_k^\infty (1 - F_{h_i(X)}(x))dx,$$

for $i = 2, 3$. Let us define

$$s = E_{\theta_1}[|h_3(X) - k|] - E_{\theta_1}[|h_2(X) - k|],$$

so that

$$s = \int_0^{h_2(a)} (F_{h_3(X)}(x) - F_{h_2(X)}(x))dx - \int_{h_2(a)}^k (F_{h_2(X)}(x) - F_{h_3(X)}(x))dx + \int_k^\infty (F_{h_2(X)}(x) - F_{h_3(X)}(x))dx.$$

By $(B.5)$ we have

$$\int_0^{h_2(a)} (F_{h_3(X)}(x) - F_{h_2(X)}(x))dx = \int_{h_2(a)}^\infty (F_{h_2(X)}(x) - F_{h_3(X)}(x))dx \geq \int_{h_2(a)}^k (F_{h_2(X)}(x) - F_{h_3(X)}(x))dx$$

Then we get that

$$s \geq \int_k^\infty (F_{h_2(X)}(x) - F_{h_3(X)}(x))dx$$

Thus,

$$E_{\theta_1}[|h_3(X) - k|] \geq E_{\theta_1}[|h_2(X) - k|]. \tag{B.6}$$

Finally we prove $(B.2)$ or equivalently that

$$E_{\theta_1}[\phi(h_3(X))] \geq E_{\theta_1}[\phi(h_2(X))]$$

Since ϕ is continuous and convex we have

$$\phi(z) - \phi(0) = \int_0^z b(k)dk,$$

where b is non decreasing and bounded in $[0, z]$. Integrating by parts it yields,

$$\phi(z) - \phi(0) = zb(z) - \int_0^z kdb(k) = \int_0^z (z-k)db(k) + zb(0).$$

Now we consider the function

$$b^*(k) = \begin{cases} b(k), & if \;\; k \in [0, z] \\ c, & if \;\; k > z \end{cases}.$$

Then we have

$$\phi(z) - \phi(0) = \int_0^z (z-k)db^*(k) + zb^*(0) + \int_z^\infty (z-k)db^*(k) = \int_0^\infty (z-k)db^*(k) + zb^*(0),$$

where we have taken into account that $\int_z^\infty (z-k)db^*(k) = 0$ and $\int_z^\infty (z-k)db^*(k) = 0$.

Therefore

$$E[\phi(Z)] = E\left[\int_z^\infty (Z-k)db^*(k) + Zb^*(0) + \phi(0)\right] =$$

$$= \int_0^\infty \int_0^\infty (z-k)db^*(k)dF_Z(z) + E[Z]b^*(0) + \phi(0) =$$

$$= \int_0^\infty E[Z-k]db^*(k) + E[Z]b^*(0) + \phi(0).$$

But,

$$\int_0^\infty E[|Z-k|]db^*(k) = \int_0^\infty \left(\int_0^\infty |z-k|dF_Z(z)\right) db^*(k) =$$

$$= \int_0^\infty \left(\int_0^z (z-k)db^*(k) + \int_0^z -(z-k)db^*(k)\right) dF_Z(z).$$

Then,

$$\int_0^\infty E[|Z-k|]db^*(k) = \int_0^\infty E[Z-k]db^*(k),$$

and thus

$$E[\phi(Z)] = \frac{1}{2}\int_0^\infty E[(Z-k)+|Z-k|]db^*(k) + E[Z]b^*(0) + \phi(0).$$

If we consider $Z \equiv h_2(X)$, we have

$$E_{\theta_1}[\phi(h_2(X))] = \frac{1}{2}\int_0^\infty 1-k+E[|h_2(X)-k|]db^*(k) + b^*(0) + \phi(0)$$

because

$$E_{\theta_1}[h_i(X)] = 1,\ i = 2,3.$$

In the same way

$$E_{\theta_1}[\phi(h_3(X))] = \frac{1}{2}\int_0^\infty 1-k+E[|h_3(X)-k|]db^*(k) + b^*(0) + \phi(0).$$

Applying $(B.6)$ we have the desired result. □