Applied Mathematics & Information Sciences **5**(2) (2011), 276-107
– An International Journal

Jensen's difference without probability vectors and actuarial applications

Athanasios Sachlas[1] and Takis Papaioannou[2]

[1]Department of Statistics & Insurance Science, University of Piraeus, 185 34, Piraeus, Greece

Email Address: asachlas@unipi.gr

[2]Department of Statistics & Insurance Science, University of Piraeus, 185 34, Piraeus, Greece

Email Address: takpap@unipi.gr

In mathematics and statistics there exist many divergences. One of them, which has a special appeal since it originates from Shannon's entropy (a well known index of diversity) and its concavity property, is *Jensen's difference* as it was called by Burbea and Rao [9]. Continuing our research on the properties and the use of divergence and information measures in the actuarial field, in the present paper, we investigate the properties of the Jensen difference in the case of non-probability vectors. This appears in actuarial graduation. Jensen's difference without probability vectors is an appropriate divergence if the vectors have equal element totals. We also investigate the use of Jensen's difference in the problem of determining a client's disability distribution [6].

Keywords: Jensen difference, Jensen-Shannon divergence, non-probability vectors, divergence measures, limiting properties, graduation, lagrangian duality, disability distribution.

1 Introduction

The bibliography provides a lot of measures of information that have been proposed and studied in the literature (see for example [21], [29]). These are mainly categorized in two groups, namely *entropy type measures* and *measures of divergence*.

A useful notion in Information Theory is *Shannon's entropy* given by

$$H(X) = -\sum_x p(x)\ln p(x) \;\text{ or }\; H(X) = -\int f(x)\ln f(x)\,dx$$

depending on whether the random variable X is discrete or continuous, with probability distribution $p(x)$ or $f(x)$, respectively. In the latter case, $H(X)$ is also called *differential entropy*. This measure quantifies the expected uncertainty related with the result of an experiment, which means that it provides information for the predictability of the outcome

of a random variable X. The larger the entropy the less concentrated the distribution of X and thus an observation of X provides a little information.

A bivariate function $D(f,g)$ of two functions or vectors f, g is a measure of divergence if $D(f,g) \geq 0$ with equality if and only if $f = g$ (c.f. [1]). It expresses the "distance" between the two functions or vectors. The main representative of this group of measures of information is the Kullback-Leibler or relative entropy. Other well known members of the group are the Cressie-Read power divergence [29] and the more general Csiszar divergence or ϕ - divergence [11], which for finite probability vectors $\mathbf{p}^\star = (p_1^\star, \ldots, p_n^\star)^T$ and $\mathbf{q}^\star = (q_1^\star, \ldots, q_n^\star)^T$ is defined by

$$I^C(\mathbf{p}^\star, \mathbf{q}^\star) = \sum_{i=1}^{n} q_i^\star \phi\left(\frac{p_i^\star}{q_i^\star}\right).$$

Function ϕ is convex in $[0,\infty)$ such that $0\phi(0/0) = 0$, $lim_{u\to 0}\phi(u) = \phi(0)$ and $0\phi(u/0) = u\phi_\infty$, where $\phi_\infty = lim_{u\to\infty}[\phi(u)/u]$, $u > 0$, $\phi(1) = 0$ and $\phi(u)$ strictly convex at $u = 1$. Special choices of ϕ lead to known measures of divergnence including the Cressie-Read power divergence. In probability and statistics these divergencies are almost exclusively used with probability distributions. As we shall see below there are cases where non-probability distributions are involved. A good reference book on measures of divergence is that of Pardo [27]. Notation with $\star$ will indicate a probability vector, while without $\star$ a non-probability vector.

A measure of divergence with a special appeal since it originates from Shannon's entropy and its concavity property is *Jensen's difference* as it was called by Burbea and Rao [9]. It is also known as *information radius* [34]. The Jensen difference between probability vectors is given by

$$J(\mathbf{p}^\star, \mathbf{q}^\star) \quad \equiv \quad H\left(\tfrac{1}{2}(\mathbf{p}^\star + \mathbf{q}^\star)\right) - \tfrac{1}{2}\left[H(\mathbf{p}^\star) + H(\mathbf{q}^\star)\right],$$

where $H(\mathbf{p}^\star) = -\sum_i p_i^\star \ln p_i^\star$ is the Shannon entropy between the finite probability vectors $\mathbf{p}^\star$ and $\mathbf{q}^\star$.

The aim of the present paper is on one hand to study the basic properties of statistical information theory for Jensen's difference with and without probability vectors and on the other to explore their use in actuarial science. It is a sequel of a recent paper by the authors [32] where similar problems have been studied with the Kullback-Leibler (KL) and the Cressie-Read (CR) power divergences. Special attention is paid to the Lagrangian duality for the Jensen difference in connection with the graduation problem.

In Section 2 of the paper we present two actuarial problems involving divergences. The first one is the determination of a client's disability distribution and the second is the graduation of mortality rates. Both of them have been presented and solved in the seminal paper of Brockett [6] via the Kullback-Leibler divergence. In this paper the emphasis is on the Jensen difference which we study in detail in Section 3. A special feature of our

approach is the use of non-probability vectors which appear in the graduation problem but may appear in other situations as well. In Section 4 we give a numerical example concerning the two problems while in Section 5 we give concluding results.

2 Actuarial problems

Information theory is related to actuarial science through the use of information measures for the treatment and solution of actuarial problems. In general terms we can categorize the use of information theory as follows: through entropy, through the Kullback-Leibler divergence or relative entropy and through other measures.

A well known method of estimating probability models is the maximum entropy principle (MEP). In this method, starting with some moments, which provide the only available information for the model, the model which maximizes the entropy is selected. This method is widely used in several fields such as economics, accounting, biology, medicine, ecology etc. [15]. Use of MEP in actuarial science can be found, among others, in references [3], [8], [12], [15], [20] and [24] dealing with topics such as loss distributions, credit risk, insurance problems, non-life insurance pricing, risk management, portfolio optimization, etc. The Kullback-Leibler directed divergence was first introduced in actuarial problems as an information theoretic method for actuarial graduation in [7] and [39]. Brockett [6] gives a very good description of the use of information theory in actuarial science. Other uses of the Kullback-Leibler directed divergence in the actuarial field can be found in [25] and [38].

Two actuarial problems that can be solved via information theoretic methods are the determination of a client's disability distribution and the graduation of mortality rates [6]. The latter appears to be more interesting since it involves non-probability vectors.

2.1 Determination of a client's disability distribution

Most insurance companies adopt a reference or standard distribution for losses. However this distribution might not be immediately applicable to a particular client's situation. So it is more common to make adjustments in order to reflect the known characteristics of the client. Particularly, for the determination of the distribution of the duration of a disability for a client with expected duration μ different from that of the standard table, which is the less distinguishable from the distribution of the table, we can minimize any divergence measure

$$D(\mathbf{p}^\star, \mathbf{q}^\star) \text{ subject to } \sum_{i=1}^{n} p_i^\star = 1 \text{ and } \sum_{i=1}^{n} x_i p_i^\star = \mu,$$

where $\mathbf{p}^\star = (p_i^\star, \ldots, p_n^\star)^T$ and $\mathbf{q}^\star = (q_i^\star, \ldots, q_n^\star)^T$. The $q_i^\star$ is the known probability of the disability having a duration of x_i days, obtained from a reference table, $\sum_{i=1}^{n} p_i^\star = 1$, $p_i^\star$ is

the unknown probability of a duration x_i days to be developed for the particular client and $x_1, \ldots, x_n$ are n discrete times of interest given in the standard table. The first constraint assures that the $p_i^\star$'s form a probability distribution.

Brockett [6] describes the minimization of the Kullback - Leibler divergence subject to the two above mentioned constraints. We note that Brockett solves the above minimization problem via its unconstrained dual convex programming problem.

2.2 Graduation through divergences

A common matter for an actuary is the description of the actual but unknown mortality pattern of a population. In order to achieve this the actuary calculates from raw data crude mortality rates, death probabilities or forces of mortality. Since these entities form an irregular series, the actuary revises the initial estimates with the aim of producing smoother estimates, with a procedure called *graduation*. There are several methods of graduation classified into parametric curve fitting and non-parametric smoothing methods. A very good reference book for graduation is that of London [23].

Brockett and Zhang [7] were the first to propose the use of information theoretic ideas in graduation. Zhang and Brockett [39] tried to construct a smooth series of n annual death probabilities $\{v_x\}$, $x = 1, 2, \ldots, n$ which is as close as possible to the observed series $\{u_x\}$ and in addition they assumed that the true but unknown underlying mortality pattern is (i) smooth, (ii) increasing with age x, i.e. monotone, (iii) more steeply increasing in higher ages, i.e. convex. They also assumed that (iv) the total number of deaths in the graduated data equals the total number of deaths in the observed data, and (v) the total age of death in the graduated data equals the total age of death in the observed data. By total age of death we mean the sum of the product of the number of deaths at every age by the corresponding age. The last two constraints imply that the average age of death is required to be the same for the observed and graduated mortality data. For the mathematical description of the constraints the interested reader is refered to [32].

In order to obtain the graduated values, Brockett in [6] minimized the Kullback-Leibler divergence between the crude death probabilities $\mathbf{u} = (u_1, \ldots, u_n)^T$ and the new death probabilities $\mathbf{v} = (v_1, \ldots, v_n)^T$,

$$I^{KL}(\mathbf{v}, \mathbf{u}) = \sum_x v_x \ln \frac{v_x}{u_x},$$

subject to the constraints (i) - (v).

It is easily seen that the annual mortality rates (death probabilities) $\mathbf{u}$ and $\mathbf{v}$ are not probability vectors since $\sum_{x=1}^n u_x$ and $\sum_{x=1}^n v_x$ may be larger or smaller than one. To solve this problem, Sachlas and Papaioannou in [32] investigated the properties of the Kullback-Leibler and Cressie-Read divergence measures in the case of non-probability vectors, concluding that under some circumstances these can be used as proper divergence measures

and proposed the use of an extra constraint in the minimization problem, i.e.,

$$\text{(vi)} \sum_{x=1}^{n} v_x = \sum_{x=1}^{n} u_x.$$

Constraint (vi) has the meaning that the overall probability of observing a death in the n year span is the same for both the observed and the graduated values.

A unifying way to obtain the graduated values v_x was proposed and investigated by the authors in [32]. This is to minimize the Cressie-Read divergence between $\mathbf{v}$ and $\mathbf{u}$

$$I^{CR}(\mathbf{v}, \mathbf{u}) = \frac{1}{\lambda(\lambda+1)} \sum_{x} v_x \left[\left(\frac{v_x}{u_x} \right)^{\lambda} - 1 \right], \ \lambda \in R - \{0, -1\},$$

for given λ subject to constraints (i) - (v) and/or (vi). In this paper, as stated above, we investigate the role of the Jensen difference.

3 The Jensen difference

The Jensen difference $J(\mathbf{p}^\star, \mathbf{q}^\star)$ is a special case of the Jensen-Shannon divergence (JSD) defined in [22] as

$$JS(\mathbf{p}^\star, \mathbf{q}^\star) = H(a\mathbf{p}^\star + (1-a)\mathbf{q}^\star) - aH(\mathbf{p}^\star) - (1-a)H(\mathbf{q}^\star)$$

for $a = 1/2$ [26]. Since then there is a confusion in the use of names Jensen-Shannon divergence and Jensen difference in the bibliography. In the present paper when will use the name Jensen difference we will refer to the measure $J(\mathbf{p}^\star, \mathbf{q}^\star)$. If we consider the function $\phi(x) = ax \ln x - [ax + (1-a)] \ln(ax + 1 - a)$, JSD can be seen as a particular case of the ϕ-divergence ([26], [36]).

The Jensen difference is a natural measure of divergence between the probability vectors $\mathbf{p}^\star$ and $\mathbf{q}^\star$ as it satisfies the two basic properties of a divergence measure. It is non-negative and vanishes if and only if $\mathbf{p}^\star = \mathbf{q}^\star$. An interesting property of $J(\mathbf{p}^\star, \mathbf{q}^\star)$ is that considered as a function of $(\mathbf{p}^\star, \mathbf{q}^\star)$, it is convex [9].

3.1 The Jensen difference with probability vectors

The properties of Jensen's difference as a measure of divergence have not been fully investigated. The following lemma gives the connection between the Jensen difference and the well known Kullback - Leibler directed divergence.

Lemma 3.1. *The Jensen difference with probability vectors $\mathbf{p}^\star$, $\mathbf{q}^\star$ is connected with the Kullback - Leibler directed divergence through the equation*

$$J(\mathbf{p}^\star, \mathbf{q}^\star) = \frac{1}{2} \left[I^{KL}(\mathbf{p}^\star, (\mathbf{p}^\star + \mathbf{q}^\star)/2) + I^{KL}(\mathbf{q}^\star, (\mathbf{p}^\star + \mathbf{q}^\star)/2) \right].$$

The above equation can be used in order to examine the information theoretic properties of Jensen's difference. In terms of the symmetric Jefrey's J-divergence $Jef(\mathbf{p}^\star, \mathbf{q}^\star) = I^{KL}(\mathbf{p}^\star, \mathbf{q}^\star) + I^{KL}(\mathbf{q}^\star, \mathbf{p}^\star)$, Crooks in [10] gave an upper bound for Jensen's difference,

$$J(\mathbf{p}^\star, \mathbf{q}^\star) \leq \ln \frac{2}{1 + \exp\left\{-\frac{1}{2} Jef(\mathbf{p}^\star, \mathbf{q}^\star)\right\}}.$$

The classical $J(\mathbf{p}^\star, \mathbf{q}^\star)$ exhibits several interesting properties [17]. Among them we mention that it is symmetric and always well defined, it takes values between 0 and 1, and its square root $\sqrt{J(\mathbf{p}^\star, \mathbf{q}^\star)}$ verifies the triangle inequality while $J(\mathbf{p}^\star, \mathbf{q}^\star)$ does not [13]. For an exhaustive enumeration of the JSD properties we refer to [14].

Generalizations of the Jensen difference have been given in [35]. A relationship between the well known Fisher information measure and two different scalar parametric generalizations of Jensen's difference divergence was established in [28]. Relationships with the Cramer-Rao inequality were also established in the same paper.

Measure J has been used for measuring the distance between random graphs, for testing the goodness-of-fit of point estimations, in the analysis of DNA sequences and in the segmentation of textured images. In addition, by making use of its ability to be generalized to an arbitrary number of probability distributions, J has been used to quantify the complex heterogeneity of DNA sequences as well as to detect borders between coding and noncoding DNA [14].

We now turn to study the sampling properties of estimated Jensen differences. For the sake of Lemmas 3.2 to 3.5, which follow below, we change the notation: k will denote the dimension of the two discrete finite probability distributions $\mathbf{p}^\star$ and $\mathbf{q}^\star$ and n or m the size of multinomial samples. If $\mathbf{q}^\star$ is known and an estimate $\hat{\mathbf{p}}^\star$ of $\mathbf{p}^\star$ is available from a multinomial sample $x_1, \ldots, x_k$, $\sum_{i=1}^k x_i = n$ obtained from population $\mathbf{p}^\star$ then $J(\mathbf{p}^\star, \mathbf{q}^\star)$ is estimated by $J(\hat{\mathbf{p}}^\star, \mathbf{q}^\star)$ with $\hat{p}_i^\star = x_i/n$, $i = 1, \ldots, k$. If $\mathbf{q}^\star$ is unknown but it is estimated on the basis of a multinomial sample $(y_1, \ldots, y_k)$, independent of $(x_1, \ldots, x_k)$ and $\sum_{i=1}^k y_i = m$, $\hat{q}_i^\star = y_i/m$, $i = 1, \ldots, k$ then $J(\mathbf{p}^\star, \mathbf{q}^\star)$ is estimated by $J(\hat{\mathbf{p}}^\star, \hat{\mathbf{q}}^\star)$. The means and variances of $J(\hat{\mathbf{p}}^\star, \mathbf{q}^\star)$ and $J(\hat{\mathbf{p}}^\star, \hat{\mathbf{q}}^\star)$ are given in the following lemmas.

Lemma 3.2. *The mean of $J(\hat{\mathbf{p}}^\star, \mathbf{q}^\star)$ is given by*

$$\begin{aligned} E[J(\hat{\mathbf{p}}^\star, \mathbf{q}^\star)] &\approx J(\mathbf{p}^\star, \mathbf{q}^\star) + \frac{1}{4n} \sum_{i=1}^k \frac{q_i^\star (1 - p_i^\star)}{p_i^\star + q_i^\star} \\ &= J(\mathbf{p}^\star, \mathbf{q}^\star) + \frac{1}{4n} \left(k - 1 - \sum_{i=1}^k \frac{p_i^\star (1 - p_i^\star)}{p_i^\star + q_i^\star} \right) \end{aligned}$$

while its variance is given by

$$Var[J(\hat{\mathbf{p}}^\star, \mathbf{q}^\star)] \approx \frac{1}{4n} \left\{ \sum_{i=1}^k p_i^\star \left[\ln \left(\frac{2p_i^\star}{p_i^\star + q_i^\star} \right) \right]^2 - \left[I^{KL}(\mathbf{p}^\star, (\mathbf{p}^\star + \mathbf{q}^\star)/2) \right]^2 \right\}.$$

Lemma 3.3. *The mean of $J(\hat{\mathbf{p}}^\star, \hat{\mathbf{q}}^\star)$ is given by*

$$E[J(\hat{\mathbf{p}}^\star, \hat{\mathbf{q}}^\star)] \approx J(\mathbf{p}^\star, \mathbf{q}^\star) + \frac{1}{4n}\sum_{i=1}^{k}\frac{q_i^\star(1-p_i^\star)}{p_i^\star+q_i^\star} + \frac{1}{4m}\sum_{i=1}^{k}\frac{p_i^\star(1-p_i^\star)}{p_i^\star+q_i^\star}$$

while its variance is given by

$$\begin{aligned} Var[J(\hat{\mathbf{p}}^\star, \hat{\mathbf{q}}^\star)] \approx\ & \frac{1}{4n}\left\{\sum_{i=1}^{k} p_i^\star\left[\ln\left(\frac{2p_i^\star}{p_i^\star+q_i^\star}\right)\right]^2 - \left[I^{KL}(\mathbf{p}^\star, (\mathbf{p}^\star+\mathbf{q}^\star)/2)\right]^2\right\} \\ & -\frac{1}{4m}\left\{\sum_{i=1}^{k} q_i^\star\left[\ln\left(\frac{2p_i^\star}{p_i^\star+q_i^\star}\right)\right]^2 + \left[I^{KL}(\mathbf{q}^\star, (\mathbf{p}^\star+\mathbf{q}^\star)/2)\right]^2\right\}. \end{aligned}$$

Proof. The results for both lemmas follow after some algebra and using known results on the multinomial distribution, the mean and the variance of the estimated ϕ-divergence [41] and the fact that the Jensen difference is a particular case of the ϕ-divergence. □

The following lemmas give the asymptotic distributions of $J(\hat{\mathbf{p}}^\star, \mathbf{q}^\star)$ and $J(\hat{\mathbf{p}}^\star, \hat{\mathbf{q}}^\star)$.

Lemma 3.4. *If $(x_1, \ldots, x_k)$ is multinomial $M(n, q_1^\star, \ldots, q_k^\star)$, then the quantity*

$$\begin{aligned} 8nJ(\hat{\mathbf{p}}^\star, \mathbf{q}^\star) =\ & 4\left[n\sum_{i=1}^{k} q_i^\star \ln q_i^\star + \sum_{i=1}^{k} x_i \ln x_i \right. \\ & \left. -n\ln n - \sum_{i=1}^{k}(x_i + nq_i^\star)\ln\left(\frac{x_i+nq_i^\star}{2n}\right)\right] \xrightarrow{L} \chi^2_{k-1}. \end{aligned}$$

Proof. The lemma is an application of [41, Theorem 3.2] and thus we omit the proof. □

Lemma 3.5. *Let $(x_1, \ldots, x_k)$, $(y_1, \ldots, y_k)$ two independent random samples from multinomials $M(n, p_1^\star, \ldots, p_k^\star)$ and $M(m, q_1^\star, \ldots, q_k^\star)$, respectively. Then if $\mathbf{p} = \mathbf{q}$ the quantity*

$$\begin{aligned} \frac{8mn}{n+m}J(\hat{\mathbf{p}}^\star, \hat{\mathbf{q}}^\star) =\ & \frac{8}{n+m}\left\{m\sum_{i=1}^{k} x_i \ln x_i - mn\ln n + n\sum_{i=1}^{k} y_i \ln y_i - mn\ln m \right. \\ & \left. -\sum_{i=1}^{k}(mx_i+ny_i)\ln\left(\frac{mx_i+ny_i}{mn}\right)\right\} \xrightarrow{L} \chi^2_{k-1}. \end{aligned}$$

Proof. The lemma is an application of [18, Corollary 2] and thus the proof is ommited. □

Lemmas 3.2 and 3.3 indicate that $J(\hat{\mathbf{p}}^\star, \mathbf{q}^\star)$ and $J(\hat{\mathbf{p}}^\star, \hat{\mathbf{q}}^\star)$ are asymptoticaly unbiased estimates of their corresponding counterparts and give their asymptotic variances. Lemmas 3.4 and 3.5 provide the means to construct tests of goodness of fit and tests of equality of divergences, etc based on one or more samples from multinomial populations. For details see [41] and [18].

3.2 The Jensen difference without probability vectors

In this subsection we will explore the properties of the Jensen difference when we have non-probability vectors. This supplements our previous research on the properties of divergence measures without probability vectors [32]. In the sequel with $\mathbf{p} = (p_1, \ldots, p_n)^T$ we will denote the non-probability vector with real nonnegative components, while with $\mathbf{p}^\star = (p_1^\star, \ldots, p_n^\star)^T$, $p_i^\star = p_i / \sum_i p_i$, $i = 1, \ldots, n$ the corresponding probability vector. Similarly for $\mathbf{q}$ and $\mathbf{q}^\star$.

Definition 3.1. We define by

$$J(\mathbf{p}, \mathbf{q}) \equiv H\left(\tfrac{1}{2}(\mathbf{p}+\mathbf{q})\right) - \tfrac{1}{2}\left[H(\mathbf{p}) + H(\mathbf{q})\right], \tag{3.1}$$

the Jensen difference between the non-probability vectors $\mathbf{p} = (p_1, \ldots, p_n)^T$ and $\mathbf{q} = (q_1, \ldots, q_n)^T$, where $\sum_i p_i \neq 1$, $\sum_i q_i \neq 1$ and $H(\mathbf{p}) = -\sum_i p_i \ln p_i$ is the "Shannon entropy" of $\mathbf{p}$.

Lemma 3.6. *For the Jensen's difference without probability vectors* $\mathbf{p}, \mathbf{q}$ *the following equation holds*

$$\begin{aligned} J(\mathbf{p}, \mathbf{q}) &= -\sum_{i=1}^{n} \frac{1}{2}\left(p_i^\star \sum_{i=1}^{n} p_i + q_i^\star \sum_{i=1}^{n} q_i\right) \ln\left[\frac{1}{2}\left(p_i^\star \sum_{i=1}^{n} p_i + q_i^\star \sum_{i=1}^{n} q_i\right)\right] \\ &\quad -\frac{1}{2}\left\{\sum_{i=1}^{n} p_i\left[H(\mathbf{p}^\star) - \ln \sum_{i=1}^{n} p_i\right] + \sum_{i=1}^{n} q_i\left[H(\mathbf{q}^\star) - \ln \sum_{i=1}^{n} q_i\right]\right\}. \end{aligned}$$

Proof. For the Shannon's entropy without probability vectors we have that

$$\begin{aligned} H(\mathbf{p}) &= -\sum_{i=1}^{n} p_i \ln p_i = -\sum_{i=1}^{n}\left(p_i^\star \sum_{i=1}^{n} p_i\right) \ln\left(p_i^\star \sum_{i=1}^{n} p_i\right) \\ &= -\sum_{i=1}^{n} p_i\left[\sum_{i=1}^{n} p_i^\star \ln p_i^\star + \ln \sum_{i=1}^{n} p_i\right] = \sum_{i=1}^{n} p_i\left[H(\mathbf{p}^\star) - \ln \sum_{i=1}^{n} p_i\right], \end{aligned}$$

where $H(\mathbf{p}^\star) = -\sum_{i=1}^{n} p_i^\star \ln p_i^\star$ is Shannon's entropy related to the probability vector $\mathbf{p}^\star$. Similarly we have that

$$H(\mathbf{q}) = \sum_{i=1}^{n} q_i\left[H(\mathbf{q}^\star) - \ln \sum_{i=1}^{n} q_i\right]$$

and

$$H(\tfrac{1}{2}(\mathbf{p}+\mathbf{q})) = -\sum_{i=1}^{n} \frac{1}{2}\left(p_i^\star \sum_{i=1}^{n} p_i + q_i^\star \sum_{i=1}^{n} q_i\right) \ln\left[\frac{1}{2}\left(p_i^\star \sum_{i=1}^{n} p_i + q_i^\star \sum_{i=1}^{n} q_i\right)\right].$$

Thus replacing the above expressions to Equation 3.1 the desirable result is obtained. □

We observe that entropy $H(\frac{1}{2}(\mathbf{p}+\mathbf{q}))$ cannot be written in terms of $H(\frac{1}{2}(\mathbf{p}^\star+\mathbf{q}^\star))$. This makes it difficult to find an easy and general expression connecting Jensen's difference without probability vectors with Jensen's difference with probability vectors and then study its properties.

In the sequel we will assume that $\sum_{i=1}^n p_i = \sum_{i=1}^n q_i$. This is the minimal requirement for a measure of divergence without probability vectors to be considered as a typical measure of divergence [32]. The relation connecting Jensen's difference without probability vectors and Jensen's difference with probability vectors is given in the following lemma.

Lemma 3.7. *If $\sum_i p_i = \sum_i q_i$, then for the Jensen difference involving non-probability vectors $\mathbf{p}, \mathbf{q}$, it holds that*

$$J(\mathbf{p},\mathbf{q}) = \left(\sum_i p_i\right) J(\mathbf{p}^\star,\mathbf{q}^\star),$$

where $J(\mathbf{p}^\star,\mathbf{q}^\star)$ is the Jensen difference between the two probability vectors $\mathbf{p}^\star,\mathbf{q}^\star$.

Proof. The desired result is easily obtained, substituting into Equation 3.1, $H(\mathbf{p})$, $H(\mathbf{q})$ and $H(\frac{1}{2}(\mathbf{p}+\mathbf{q}))$ given in the proof of Lemma 3.6. □

Now we have to see if this measure has information theoretic and divergence properties.

Proposition 3.1. *Let $\sum_i p_i = \sum_i q_i > 0$. Then $J(\mathbf{p},\mathbf{q}) \geq 0$ with equality if and only if $\mathbf{p} = \mathbf{q}$, where $\mathbf{p}$ and $\mathbf{q}$ are non-probability vectors. Moreover $J(\mathbf{p},\mathbf{q}) \leq \sum_i p_i$.*

Proof. The proof is obvious since $J(\mathbf{p}^\star,\mathbf{q}^\star) \geq 0$ if and only if $\mathbf{p}^\star = \mathbf{q}^\star$ and $J(\mathbf{p}^\star,\mathbf{q}^\star) \leq 1$. □

Note also that in view of the properties of $J(\mathbf{p}^\star,\mathbf{q}^\star)$, $\sqrt{J(\mathbf{p},\mathbf{q})}$ is a metric for non-probability vectors.

Definition 3.2. (Bivariate Shannon entropy) Let $p(x,y)$ be a bivariate non-probability function associated with two discrete variables X, Y in R^2 for which it holds $\sum_x \sum_y p(x,y) \neq 1$. We define the Shannon entropy involving a non-probability function p as

$$H_{X,Y}(p) = -\sum_x \sum_y p(x,y) \ln p(x,y).$$

Definition 3.3. (Conditional Shannon entropy) For the discrete variables X, Y and the bivariate non-probability function $p(x,y)$, as given above let $f(x) = \sum_y p(x,y)$, $h(y|x) = \frac{p(x,y)}{f(x)}$, $g(y) = \sum_x p(x,y)$, and $r(x|y) = \frac{p(x,y)}{g(y)}$. We set

$$H_{Y|X=x}(h) = \sum_y h(y|x) \ln h(y|x), H_{X|Y=y}(r) = \sum_x r(x|y) \ln r(x|y)$$

and define

$$H_{Y|X}(h) = E_X\left[H_{Y|X=x}(h)\right] = \sum_x f(x) \sum_y h(y|x) \ln h(y|x),$$

$$H_{X|Y}(r) = E_Y\left[H_{X|Y=y}(r)\right] = \sum_y g(y) \sum_x r(x|y) \ln r(x|y).$$

Definition 3.4. (Bivariate Jensen difference) Let $p_i(x,y)$, $i = 1,2$, be two bivariate non-probability functions associated with two discrete variables X, Y in R^2 for which it holds $\sum_x \sum_y p_i(x,y) \neq 1$. We define the Jensen difference between two bivariate non-probability functions p_1, p_2 as

$$\begin{aligned} J_{X,Y}(p_1,p_2) &= H\left(\tfrac{1}{2}(p_1+p_2)\right) - \tfrac{1}{2}\left[H(p_1)+H(p_2)\right] \\ &= -\sum_x \sum_y \tfrac{1}{2}(p_1(x,y)+p_2(x,y)) \ln(\tfrac{1}{2}(p_1(x,y)+p_2(x,y))) \\ &\quad -\tfrac{1}{2}\left[-\sum_x \sum_y p_1(x,y)\ln p_1(x,y) - \sum_x \sum_y p_2(x,y)\ln p_2(x,y)\right]. \end{aligned}$$

Definition 3.5. (Conditional Jensen difference) For the discrete variables X, Y and the bivariate non-probability functions $p_i(x,y)$, $i = 1,2$, as given above let $f_i(x) = \sum_y p_i(x,y)$, $h_i(y|x) = \frac{p_i(x,y)}{f_i(x)}$, $g_i(y) = \sum_x p_i(x,y)$, and $r_i(x|y) = \frac{p_i(x,y)}{g_i(y)}$, $i = 1,2$. We set

$$\begin{aligned} J_{Y|X=x}(h_1,h_2) &= H\left(\tfrac{1}{2}(h_1+h_2)\right) - \tfrac{1}{2}\left[H(h_1)+H(h_2)\right] \\ &= -\sum_x \sum_y \tfrac{1}{2}(h_1(y|x)+h_2(y|x)) \ln(\tfrac{1}{2}(h_1(y|x)+h_2(y|x))) \\ &\quad -\tfrac{1}{2}\left[-\sum_x \sum_y h_1(y|x)\ln h_1(y|x) - \sum_x \sum_y h_2(y|x)\ln h_2(y|x)\right], \end{aligned}$$

and define

$$\begin{aligned} J_{Y|X}(h_1,h_2) &= E_X\left[J_{Y|X=x}(h_1,h_2)\right] \\ &= -\sum_x f_1(x) \sum_y \tfrac{1}{2}(h_1(y|x)+h_2(y|x)) \ln(\tfrac{1}{2}(h_1(y|x)+h_2(y|x))) \\ &\quad -\tfrac{1}{2}\left[-\sum_x f_1(x)\sum_y h_1(y|x)\ln h_1(y|x) - \sum_x f_1(x)\sum_y h_2(y|x)\ln h_2(y|x)\right]. \end{aligned}$$

The conditional Jensen's difference $J_{X|Y}(r_1,r_2)$ is defined analogously.

Proposition 3.2. *(Strong Additivity) Let p_1, p_2 be two bivariate non-probability functions associated with two discrete variables X, Y in R^2 as in Definition 3.5. Then*

$$J_{X,Y}(p_1,p_2) = J_X(f_1,f_2) + J_{Y|X}(h_1,h_2) = J_Y(g_1,g_2) + J_{X|Y}(r_1,r_2),$$

where the functions f_i, h_i, g_i, r_i, $i = 1,2$ are as in Definition 3.5.

Proof. It is known [33] that

$$H_{X^\star,Y^\star}(p_i^\star) = H_{X^\star}(f_i^\star) + H_{Y^\star|X^\star}(h_i^\star) = H_{Y^\star}(g_i^\star) + H_{X^\star|Y^\star}(r_i^\star),\ i = 1, 2.$$

Thus, we have that

$$\begin{aligned} J_{X^\star,Y^\star}(p_1^\star, p_2^\star) &= H_{X^\star,Y^\star}(\tfrac{1}{2}(p_1^\star + p_2^\star)) - \tfrac{1}{2}\left[H_{X^\star,Y^\star}(p_1^\star) + H_{X^\star,Y^\star}(p_2^\star)\right] \\ &= H_{X^\star}(\tfrac{1}{2}(f_1^\star + f_2^\star)) + H_{Y^\star|X^\star}(\tfrac{1}{2}(h_1^\star + h_2^\star)) \\ &\quad -\tfrac{1}{2}\left[H_{X^\star}(f_1^\star) + H_{Y^\star|X^\star}(h_1^\star) + H_{X^\star}(f_2^\star) + H_{Y^\star|X^\star}(h_2^\star)\right] \\ &= J_{X^\star}(f_1^\star, f_2^\star) + J_{Y^\star|X^\star}(h_1^\star, h_2^\star), \end{aligned}$$

which means that the strong additivity property holds for the Jensen difference. Similarly it holds that

$$J_{X^\star,Y^\star}(p_1^\star, p_2^\star) = J_{Y^\star}(g_1^\star, g_2^\star) + J_{X^\star|Y^\star}(r_1^\star, r_2^\star).$$

For the variables X, Y we have that

$$\begin{aligned} J_{X,Y}(p_1, p_2) &= \left(\sum_x \sum_y p_1(x, y)\right) J_{x^\star,y^\star}(p_1^\star, p_2^\star) \\ &= \left(\sum_x \sum_y p_1(x, y)\right) \left[J_{X^\star}(f_1^\star, f_2^\star) + J_{Y^\star|X^\star}(h_1^\star, h_2^\star)\right] \\ &= J_X(f_1, f_2) + J_{Y|X}(h_1, h_2), \end{aligned}$$

since $\sum_x \sum_y p_1(x, y) = \sum_x f_1(x) = \sum_y g_1(y)$. In a similar way, we prove that

$$J_{X,Y}(p_1, p_2) = J_Y(g_1, g_2) + J_{X|Y}(r_1, r_2).$$

□

For weak additivity we have the following proposition.

Proposition 3.3. *(Weak additivity) If $h_i(y|x) = g_i(y)$ and thus $p_i(x, y) = f_i(x)g_i(y)$, $i = 1, 2$, we have that the random variables $X^\star, Y^\star$, which are the "standardized" values of X, Y, are independent, then*

$$J_{X,Y}(p_1, p_2) = J_X(f_1, f_2) + J_Y(g_1, g_2).$$

Proof. It is known [33] that

$$H_{X^\star,Y^\star}(p_i^\star) = H_{X^\star}(f_i^\star) + H_{Y^\star}(g_i^\star),\ i = 1, 2.$$

Thus, we have that

$$\begin{aligned} J_{X^\star,Y^\star}(p_1^\star, p_2^\star) &= H_{X^\star,Y^\star}(\tfrac{1}{2}(p_1^\star + p_2^\star)) - \tfrac{1}{2}\left[H_{X^\star,Y^\star}(p_1^\star) + H_{X^\star,Y^\star}(p_2^\star)\right] \\ &= H_{X^\star}(\tfrac{1}{2}(f_1^\star + f_2^\star)) + H_{Y^\star}(\tfrac{1}{2}(g_1^\star + g_2^\star)) \\ &\quad -\tfrac{1}{2}\left[H_{X^\star}(f_1^\star) + H_{Y^\star}(g_1^\star) + H_{X^\star}(f_2^\star) + H_{Y^\star}(g_2^\star)\right] \\ &= J_{X^\star}(f_1^\star, f_2^\star) + J_{Y^\star}(g_1^\star, g_2^\star), \end{aligned}$$

which means that the weak additivity property holds for the Jensen difference.

Then for the variables X, Y we have that

$$\begin{aligned} J_{X,Y}(p_1, p_2) &= \left(\sum_x \sum_y p_1(x,y)\right) J_{X^\star,Y^\star}(p_1^\star, p_2^\star) \\ &= \left(\sum_x \sum_y p_1(x,y)\right) \left[J_{X^\star}(f_1^\star, f_2^\star) + I_{Y^\star}(g_1^\star, g_2^\star)\right]. \end{aligned}$$

Since it holds that $\sum_x \sum_y p_i(x,y) = \sum_x f_i(x) = \sum_y g_i(y)$, $i = 1, 2$ we finally have that

$$J_{X,Y}(p_1, p_2) = J_X(f_1, f_2) + J_Y(g_1, g_2).$$

□

Proposition 3.4. *(Maximal information and sufficiency) Let $Y = T(X)$ be a measurable transformation of X and $p_i = p_i(x)$, $g_i = g_i(y)$, $i = 1, 2$. Then*

$$J_X(p_1, p_2) \geq J_Y(g_1, g_2),$$

with equality if and only if $Y^\star$ is sufficient with respect to the pair of distributions $p_1^\star$ and $p_2^\star$, $Y^\star$ and $X^\star$ being the normalized versions of Y and X, respectively.

Proof. Let $g_i(y)$ be the measure associated with Y. Then $g_i(y) = \sum_{x:T(x)=y} p_i(x)$. The following inequalities are equivalent

$$\begin{aligned} J_X(p_1, p_2) &\geq J_Y(g_1, g_2) \Leftrightarrow \\ \left(\sum_x p_1(x)\right) J_{X^\star}(p_1^\star, p_2^\star) &\geq \left(\sum_y g_1(y)\right) J_{Y^\star}(g_1^\star, g_2^\star). \end{aligned}$$

Since $\sum_x p_i(x) = \sum_y g_i(y)$, $i = 1, 2$, the last inequality is equivalent to

$$J_{X^\star}(p_1^\star, p_2^\star) \geq J_{Y^\star}(g_1^\star, g_2^\star),$$

which always holds. Equality holds if and only if the statistic $Y^\star = T(X^\star)$ is sufficient.

□

One basic property of measures of information and divergence is the limiting property. This property asserts that a series $\{X_n\}$ of random variables converges to a random variable X in distribution when $n \to \infty$ if and only if $I_{X_n} \to I_X$, where I denotes the information measure. Under some conditions the limiting property holds for the Kullback-Leibler divergence (see [16] and [40]).

The limiting property holds for the Csiszar's measure of divergence (ϕ-divergence) [40]. So the limiting property holds for the Jensen difference with probability vectors as it is a member of the ϕ-divergence family for proper $\phi(x)$. In the next proposition we investigate whether the limiting property holds in case we do not have probability vectors.

Proposition 3.5. *(The limiting property) Let $\{\mathbf{p}_n\}$ be a bounded from above sequence of non-probability vectors. Then $\boldsymbol{p}_n \to \boldsymbol{p}$ if and only if $J(\boldsymbol{p}_n, \boldsymbol{p}) \to 0$.*

Proof. Let $\mathbf{p}_n \to \mathbf{p}$. Using Lemma 3.7 we have

$$\lim_{n\to\infty} J(\mathbf{p}_n, \mathbf{p}) \quad = \quad \left(\lim_{n\to\infty} \sum_i p_n(i)\right) \lim_{n\to\infty} J(\mathbf{p}_n^\star, \mathbf{p}^\star) = 0,$$

because $\lim_{n\to\infty} J(\mathbf{p}_n^\star, \mathbf{p}^\star) = 0$.

On the other hand, let $J(\mathbf{p}_n, \mathbf{p}) \to 0$. Then

$$\lim_{n\to\infty} \sum_i p(i)\phi\left(\frac{p_n(i)}{p(i)}\right) = 0,$$

where $\phi(x) = \frac{1}{2}\left[x \ln x - (x+1)\ln\left(\frac{x+1}{2}\right)\right]$, $x > 0$ is a continuous function with $\phi(1) = 0$.

Suppose that $\mathbf{p}_n \to \mathbf{p}$ does not hold. So there is a subsequence $n_1 < n_2 < \ldots < n_s < \ldots$ of integers and a vector $\mathbf{q}$ such that

$$\lim_{s\to\infty} \mathbf{p}_{n_s} = \mathbf{q} \text{ and } \mathbf{p} \neq \mathbf{q}. \tag{3.2}$$

Because ϕ is continuous we have that

$$\lim_{s\to\infty} \sum_i p(i)\phi\left(\frac{p_{n_s}(i)}{p(i)}\right) = \sum_i p(i)\phi\left(\frac{q(i)}{p(i)}\right).$$

However $\left\{\sum_i p(i)\phi\left(\frac{p_{n_s}(i)}{p(i)}\right)\right\}$ is a subsequence of $\left\{\sum_i p(i)\phi\left(\frac{p_n(i)}{p(i)}\right)\right\}$, which converges to $\phi(1) = 0$. Thus

$$\sum_i p(i)\phi\left(\frac{q(i)}{p(i)}\right) = \phi(1) = 0,$$

which is possible only if $p(i) = q(i)$, which contradicts Equation 3.2. Thus we have that $\mathbf{p}_n \to \mathbf{p}$, so the limiting property holds for the Jensen difference. □

Summarizing the above results, we have that the Jensen difference $J(\mathbf{p},\mathbf{q})$ for non-probability vectors, under some conditions is nonnegative, additive, invariant under sufficient transformations, it shares the property of maximal information and the limiting one. Thus, we can regard $J(\mathbf{p},\mathbf{q})$ as a measure of divergence, provided that $\sum_i p_i = \sum_i q_i$.

The study of the properties of Jensen's difference without probability vectors, along with the results of our previous research on KL and CR measures [32], lead us to say that the equality $\sum_i p_i = \sum_i q_i$ constitutes the minimal requirement for a bivariate function $D(\mathbf{p},\mathbf{q})$ to be a measure of divergence along with the statement that $D(\mathbf{p},\mathbf{q}) \geq 0$ with equality if and only if $\mathbf{p} = \mathbf{q}$.

Since the Jensen difference can be considered as a divergence measure we can use it in order to graduate actuarial entities, in the way we describe in Section 2.2. This involves convex minimization with constraints and can be done using standard routines. However, it is of interest to examine its Lagrangian dual.

3.3 Lagrangian duality for the Jensen difference

The quadratically constrained Jensen difference problem is defined as finding $\mathbf{x} = (x_1,\ldots,x_n)^T \in R^n$ which solves the primal problem

$$(P) \qquad \min -\sum_{j=1}^{n} \tfrac{1}{2}(x_j+d_j)\ln\left(\tfrac{1}{2}(x_j+d_j)\right) + \tfrac{1}{2}\left[\sum_{j=1}^{n} x_j \ln x_j + \sum_{j=1}^{n} d_j \ln d_j\right]$$

subject to

$g_i(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{D}_i\mathbf{x} + \mathbf{b}_i^T\mathbf{x} + c_i \leq 0$, $i = 1,2,\ldots,m$, $m = 2(n+1)$, $\mathbf{x} \geq \mathbf{0}$,

where $\mathbf{d} = (d_1,\ldots,d_n)^T$ is a given vector with strictly positive components, $\mathbf{D}_i$ is a given positive semi-definite matrix for each i, $\mathbf{b}_i \in R^n$ and c_i are given constants not both equal to zero. Constraints (i) - (v) of the actuarial graduation problem of Subsection 2.2 can be written in the previous form of $g_i(\mathbf{x}) \leq 0$. For details see [32].

In the sequel, we will try to derive a dual representation of the primal problem (P) by means of Lagrangian duality by using a simple decomposition argument to convert problem (P) into an equivalent convex program with linear and quadratic constraints. Because $\mathbf{D}_i$ is a semipositive definite $n \times n$ matrix, we can express it as $\mathbf{D}_i = \mathbf{A}_i^T\mathbf{A}_i$, where $\mathbf{A}_i$ is an $n_i \times n$ matrix and n_i is the rank of $\mathbf{D}_i$, $i = 1,2,\ldots,m$. In this case the constraints can be written as $g_i(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}_i^T\mathbf{A}_i\mathbf{x} + \mathbf{b}_i^T\mathbf{x} + c_i$. Defining the new variables $\mathbf{u}_i = \mathbf{A}_i\mathbf{x}$, $\mathbf{u}_i \in R^{n_i}$, $i = 1,2,\ldots,m$, the problem (P) is equivalent to the following convex program with linear equality and quadratic inequality constraints:

$$(P^\star) \qquad \min_{\mathbf{x},\mathbf{u}_i} -\sum_{j=1}^{n} \tfrac{1}{2}(x_j+d_j)\ln\left(\tfrac{1}{2}(x_j+d_j)\right) + \tfrac{1}{2}\left[\sum_{j=1}^{n} x_j \ln x_j + \sum_{j=1}^{n} d_j \ln d_j\right]$$

subject to

$\frac{1}{2}\mathbf{u}_i^T\mathbf{u}_i + \mathbf{b}_i^T\mathbf{x} + c_i \leq 0$, $\mathbf{A}_i\mathbf{x} = \mathbf{u}_i$, $\mathbf{u}_i \in R^{n_i}$, $i = 1,2,\ldots,m$, $\mathbf{x} \geq \mathbf{0}$.

Let $\mathbf{u} = (u_1, \ldots, u_m)^T \in R^{n_i} \times \ldots \times R^{n_m}$ and $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_m)^T \in R^{n_i} \times \ldots \times R^{n_m}$. We now have

Theorem 3.1. *The Lagrangian dual problem of* (P) *is given by*

$$(D) \qquad \sup_{\boldsymbol{\lambda} \in R_+^m, \mathbf{y}_i \in R^{n_i}} \left\{ -\sum_{j=1}^{n} \frac{d_j}{2e^{2s_j} - 1} \left[e^{2s_j} \ln \left(\frac{d_j e^{2s_j}}{2e^{2s_j} - 1} \right) - \frac{1}{2} \ln \left(\frac{d_j}{2e^{2s_j} - 1} \right) - s_j \right] \right.$$
$$\left. -\frac{1}{2} \sum_{i=1}^{m} \frac{\|\mathbf{y}_i\|^2}{\lambda_i} + \boldsymbol{\lambda}^T \mathbf{c} + \mathbf{d}^T \mathbf{z} \right\},$$

where $\mathbf{z}^T = (\ln d_1, \ldots, \ln d_n)$.

Proof. The Lagrangian function for problem (P) is

$$\begin{aligned} L(\mathbf{x}, \mathbf{u}; \boldsymbol{\lambda}, \mathbf{y}) &= -\sum_{j=1}^{n} \tfrac{1}{2}(x_j + d_j) \ln \left(\tfrac{1}{2}(x_j + d_j) \right) + \frac{1}{2} \left[\sum_{j=1}^{n} x_j \ln x_j + \sum_{j=1}^{n} d_j \ln d_j \right] \\ &\quad + \sum_{i=1}^{m} \lambda_i \left(\frac{1}{2} \mathbf{u}_i^T \mathbf{u}_i + \mathbf{b}_i^T \mathbf{x} + c_i \right) + \sum_{i=1}^{m} \mathbf{y}_i^T \left(\mathbf{A}_i \mathbf{x} - \mathbf{u}_i \right) \\ &= -\sum_{j=1}^{n} \tfrac{1}{2}(x_j + d_j) \ln \left(\tfrac{1}{2}(x_j + d_j) \right) + \frac{1}{2} \left[\sum_{j=1}^{n} x_j \ln x_j + \sum_{j=1}^{n} d_j \ln d_j \right] \\ &\quad + \sum_{i=1}^{m} (\lambda_i \mathbf{b}_i^T + \mathbf{y}_i^T \mathbf{A}_i) \mathbf{x} + \boldsymbol{\lambda}^T \mathbf{c} + \sum_{i=1}^{m} \left(\frac{1}{2} \lambda_i \mathbf{u}_i^T \mathbf{u}_i - \mathbf{y}_i^T \mathbf{u}_i \right), \end{aligned}$$

where $\mathbf{y}_i \in R^{n_i}$, $i = 1, 2, \ldots, m$ are vector Lagrange multipliers while the Lagrangian dual objective function of problem (P) is given by

$$h(\boldsymbol{\lambda}, \mathbf{y}) = \inf_{\mathbf{x} \geq \mathbf{0}, \mathbf{u}_i \in R^{n_i}} L(\mathbf{x}, \mathbf{u}; \boldsymbol{\lambda}, \mathbf{y}).$$

The dual problem associated with (P) is defined as

$$(D) \qquad \sup_{\boldsymbol{\lambda} \in R_+^m, \mathbf{y}_i \in R^{n_i}} h(\boldsymbol{\lambda}, \mathbf{y}).$$

With $\boldsymbol{\lambda}$ we denote the vector of Lagrange multipliers of the primal problem associated with constraints $\frac{1}{2} \mathbf{u}_i^T \mathbf{u}_i + \mathbf{b}_i^T \mathbf{x} + c_i \leq 0$ while with $\mathbf{y}$ the vector of vector Lagrange multipliers associated with constraints $\mathbf{A}_i \mathbf{x} = \mathbf{u}_i$, $\mathbf{u}_i \in R^{n_i}$, $i = 1, 2, \ldots, m$. Using the fact that the Lagrangian function is separable in the two decision variables, $\mathbf{x}$ and $\mathbf{u}$ [5], we derive an

explicit form for the dual objective function $h(\boldsymbol{\lambda}, \mathbf{y})$ [37] as follows:

$$\begin{aligned} h(\boldsymbol{\lambda}, \mathbf{y}) &= -\sum_{j=1}^{n} \inf_{x_j \geq 0} \left\{ \tfrac{1}{2}(x_j + d_j) \ln\left(\tfrac{1}{2}(x_j + d_j)\right) + \frac{1}{2} x_j \ln x_j \right. \\ &\qquad \left. + \sum_{i=1}^{m} (\lambda_i \mathbf{b}_i^T + \mathbf{y}_i^T \mathbf{A}_i)_j x_j \right\} \\ &\quad + \sum_{i=1}^{m} \inf_{\mathbf{u}_i \in R^{n_i}} \left\{ \frac{1}{2} \lambda_i \mathbf{u}_i^T \mathbf{u}_i - \mathbf{y}_i^T \mathbf{u}_i \right\} + \mathbf{c}^T \boldsymbol{\lambda} + \mathbf{d}^T \mathbf{z}. \end{aligned} \tag{3.3}$$

Let us now denote the terms involving x's of the right hand side of $L(\mathbf{x}, \mathbf{u}; \boldsymbol{\lambda}, \mathbf{y})$ by

$$\begin{aligned} f(\mathbf{x}; \boldsymbol{\lambda}, \mathbf{y}) &= -\sum_{j=1}^{n} \tfrac{1}{2}(x_j + d_j) \ln\left(\tfrac{1}{2}(x_j + d_j)\right) \\ &\quad + \frac{1}{2} \sum_{j=1}^{n} x_j \ln x_j + \sum_{i=1}^{m} (\lambda_i \mathbf{b}_i^T + \mathbf{y}_i^T \mathbf{A}_i)\mathbf{x}. \end{aligned} \tag{3.4}$$

It is easy to see that

$$\frac{\partial}{\partial x_j} f(\mathbf{x}; \boldsymbol{\lambda}, \mathbf{y}) = -\frac{1}{2} \ln\left(\tfrac{1}{2}(x_j + d_j)\right) + \frac{1}{2} \ln x_j + \sum_{i=1}^{m} (\lambda_i \mathbf{b}_i^T + \mathbf{y}_i^T \mathbf{A}_i)_j.$$

In order to find the optimal point, we set the above equation equal to zero, so we have

$$\begin{aligned} \frac{\partial}{\partial x_j} f(\mathbf{x}; \boldsymbol{\lambda}, \mathbf{y}) &= 0 \Leftrightarrow \\ x_j &= \frac{d_j}{2 \exp\left\{ 2 \sum_{i=1}^{m} (\lambda_i \mathbf{b}_i^T + \mathbf{y}_i^T \mathbf{A}_i)_j \right\} - 1}. \end{aligned} \tag{3.5}$$

Substituting Equation 3.5 to Equation 3.4, and setting $s_j = \sum_{i=1}^{m} (\lambda_i \mathbf{b}_i^T + \mathbf{y}_i^T \mathbf{A}_i)_j$, $j = 1, \ldots, n$ we have that

$$\begin{aligned} & -\sum_{j=1}^{n} \frac{1}{2} \left(\frac{d_j}{2e^{2s_j} - 1} + d_j \right) \ln\left(\frac{1}{2} \left(\frac{d_j}{2e^{2s_j} - 1} + d_j \right) \right) \\ & + \frac{1}{2} \sum_{j=1}^{n} \frac{d_j}{2e^{2s_j} - 1} \ln\left(\frac{d_j}{2e^{2s_j} - 1} \right) + \sum_{j=1}^{n} s_j \frac{d_j}{2e^{2s_j} - 1} \\ = & -\sum_{j=1}^{n} \frac{d_j e^{2s_j}}{2e^{2s_j} - 1} \ln\left(\frac{d_j e^{2s_j}}{2e^{2s_j} - 1} \right) + \frac{1}{2} \sum_{j=1}^{n} \frac{d_j}{2e^{2s_j} - 1} \ln\left(\frac{d_j}{2e^{2s_j} - 1} \right) + \sum_{j=1}^{n} \frac{d_j s_j}{2e^{2s_j} - 1} \\ = & -\sum_{j=1}^{n} \frac{d_j}{2e^{2s_j} - 1} \left\{ e^{2s_j} \ln\left(\frac{d_j e^{2s_j}}{2e^{2s_j} - 1} \right) - \frac{1}{2} \ln\left(\frac{d_j}{2e^{2s_j} - 1} \right) - s_j \right\}, \end{aligned}$$

which is the minimum value of the first infimum in Equation 3.3.

Setting $g(\mathbf{u};\boldsymbol{\lambda},\mathbf{y})$ for the last term of the Lagrangian function $L(\mathbf{x},\mathbf{u};\boldsymbol{\lambda},\mathbf{y})$, i.e.

$$g(\mathbf{u};\boldsymbol{\lambda},\mathbf{y}) = \sum_{i=1}^{m}\left(\frac{1}{2}\lambda_i\mathbf{u}_i^T\mathbf{u}_i - \mathbf{y}_i^T\mathbf{u}_i\right) \tag{3.6}$$

we have that

$$\frac{\partial}{\partial u_i}g(\mathbf{u};\boldsymbol{\lambda},\mathbf{y}) = \lambda_i\mathbf{u}_i - \mathbf{y}_i.$$

In order to find the optimal point, we set the above equation equal to zero, so we have

$$\begin{aligned}\frac{\partial}{\partial u_i}g(\mathbf{u};\boldsymbol{\lambda},\mathbf{y}) &= \mathbf{0} \Leftrightarrow\\ \lambda_i\mathbf{u}_i &= \mathbf{y}_i,\end{aligned}$$

which means that

$$\mathbf{u}_i = \frac{1}{\lambda_i}\mathbf{y}_i. \tag{3.7}$$

By substitution of Equation 3.7 to Equation 3.6 we have that

$$\sum_{i=1}^{m}\left(\frac{1}{2}\lambda_i\left(\frac{1}{\lambda_i}\mathbf{y}_i\right)^T\left(\frac{1}{\lambda_i}\mathbf{y}_i\right) - \mathbf{y}_i^T\left(\frac{1}{\lambda_i}\mathbf{y}_i\right)\right) = -\frac{1}{2}\sum_{i=1}^{m}\frac{\|\mathbf{y}_i\|^2}{\lambda_i},$$

which is the minimum value of the second infimum in Equation 3.3. □

Theorem 3.2. *(a) If* (P) *is feasible then* $\inf(P)$ *is attained and* $\min(P) = \sup(D)$. *Moreover, if there exists an* $\mathbf{x} \in R^n$ *satisfying* $\mathbf{x} > 0$, $g_i(\mathbf{x}) < 0$, $i = 1,\ldots,m$, *then* $\sup(D)$ *is attained and* $\min(P) = \max(D)$.

(b) If $\mathbf{x}^\star$ *solves the primal problem* (P) *and* $\mathbf{y}_i^\star \in R^{n_i}$, $\boldsymbol{\lambda}^\star \in R_+^m$ *solve the dual problem* (D), *then*

$$x_j^\star = \frac{d_j}{2\exp\left\{2\sum\limits_{i=1}^{m}(\lambda_i^\star\mathbf{b}_i^T + \mathbf{y}_i^{\star T}\mathbf{A}_i)_j\right\} - 1}.$$

Proof. The proof of the theorem can be obtained via standard duality results (see for example [19], [30] or [5]).

(a) Because of the nonnegativity of the constraints, i.e. $\boldsymbol{\lambda} \in R_+^m$, of the dual problem (D), this satisfies the strongest constraint qualification which implies lack of duality gap and attainment of the primal infimum. Thus the first part follows immediately. The second part follows from the definition of duality.

(b) The optimality condition for $\mathbf{x} = \mathbf{x}^\star$ to solve the minimization of $h(\boldsymbol{\lambda},\mathbf{y})$ given in Equation (3.3) is the optimal solution $x_j^\star$, given above, and thus the desired result follows.

By part (a) we have that a saddle point $(\mathbf{x}^\star,\mathbf{y}_i^\star,\boldsymbol{\lambda}^\star)$ exists and so $\min_{\mathbf{x}\geq 0} L(\mathbf{x},\mathbf{y}_i^\star,\boldsymbol{\lambda}^\star) = L(\mathbf{x}^\star,\mathbf{y}_i^\star,\boldsymbol{\lambda}^\star)$ [2]. □

4 Numerical Illustration

4.1 Determination of a client's disability distribution with Jensen's difference

In this subsection we use the Jensen difference to determine the disability distribution that meets the special characteristics of a client than the reference table that the insurance company uses. The data that we use comes from [4, Table 13.2]. It is a standard table of probabilities q_i with mean duration $\mu_{st} = 31.35$ days given in the second column of Table 4.1. It is easy to notice that $\sum_{i=1}^{n} q_i^\star = 1$. Suppose that we have a client with expected disability duration of $\mu = 21$ days and we want to construct a duration table for this particular client which is the least distinguishable from the standard one. This problem was also solved by Brockett [6] by minimizing the Kullback-Leibler divergence between the unknown probabilities for the client and the corresponding probabilities of the standard table subject to the constraints $\sum_{i=1}^{n} p_i^\star = 1$ and $\sum_{i=1}^{n} x_i p_i^\star = 21$. His results are shown in the third column of Table 4.1. Our approach is the minimization of an alternative divergence - the Jensen difference - subject to the same constraints.

The results are shown in the fourth column of Table 4.1. We followed the same procedure in order to derive the duration table for two clients with $\mu = 26.8$ and $\mu = 38$, respectively. Comparing the results via the smoothness measure $S = \sum_{i=1}^{n-3} \left(\Delta^3 p_i^\star\right)^2$, where Δ is the difference operator, and the mean square error $MSE = \frac{1}{n}\sum_{i=1}^{n}(q_i^\star - p_i^\star)^2$ (given in Table 4.2), the best disability distribution for $\mu = 21$ and $\mu = 26.8$ is obtained through the minimization of the Jensen difference while for $\mu = 38$ the best disability distribution is obtained via the minimization of the Kullback - Leibler divergence. A further numerical illustration allows us to propose the use of the Jensen difference when $\mu < \mu_{st}$ and the use of the Kullback - Leibler divergence when $\mu > \mu_{st}$, where μ_{st} is the mean duration of the standard table.

4.2 Actuarial graduation

For the illustration, we will use a data set of death probabilities coming from [23, p. 20]. It consists of 15 death probabilities belonging to ages 70 to 84 (computed from a total of 2073 observations). These data set was graduated by London in [23] by graphic means and a linear transformation of the graduated values and by Brockett in [6] via the minimization of the Kullback-Leibler divergence subject to constraints (i) - (v).

We graduated the crude values via the minimization of the Jensen difference. The minimization was conducted subject to constraints (i) - (v), proposed in [6], the additional constraint (vi) that Sachlas and Papaioannou proposed in [32] and finally subject to constraints (i) - (iii) and (vi). The sixth constraint we propose has no any particular actuarial interpretation. However it is necessary in the light of the information theoretic properties. The relevant results are presented along with the raw data in Table 4.3(a). Graphically, the

		$\mu = 21$		$\mu = 26.8$		$\mu = 38$	
x	Standard	K-L	Jensen	K-L	Jensen	K-L	Jensen
1	0.03500	0.05081	0.05298	0.04101	0.04130	0.02777	0.02810
2	0.03474	0.04968	0.05151	0.04048	0.04073	0.02775	0.02806
3	0.03349	0.04717	0.04865	0.03880	0.03901	0.02694	0.02721
4	0.03318	0.04604	0.04724	0.03822	0.03841	0.02687	0.02712
5	0.03195	0.04367	0.04459	0.03660	0.03675	0.02606	0.02627
6	0.03160	0.04254	0.04324	0.03599	0.03612	0.02595	0.02614
7	0.03040	0.04031	0.04079	0.03443	0.03453	0.02514	0.02530
8	0.03002	0.03921	0.03951	0.03381	0.03388	0.02499	0.02513
9	0.02885	0.03712	0.03725	0.03231	0.03236	0.02419	0.02430
10	0.02701	0.03423	0.03423	0.03008	0.03011	0.02280	0.02289
11	0.02530	0.03159	0.03147	0.02801	0.02803	0.02150	0.02157
12	0.02370	0.02915	0.02894	0.02609	0.02609	0.02028	0.02033
13	0.02222	0.02692	0.02664	0.02433	0.02431	0.01915	0.01917
14	0.02083	0.02485	0.02452	0.02268	0.02265	0.01808	0.01808
15	0.01953	0.02295	0.02258	0.02114	0.02111	0.01706	0.01706
16	0.01831	0.02120	0.02080	0.01971	0.01967	0.01611	0.01609
17	0.01772	0.02021	0.01978	0.01897	0.01892	0.01570	0.01567
18	0.01662	0.01867	0.01823	0.01769	0.01764	0.01483	0.01479
19	0.01611	0.01783	0.01737	0.01705	0.01699	0.01447	0.01442
20	0.01510	0.01646	0.01600	0.01589	0.01583	0.01366	0.01361
21	0.01465	0.01573	0.01526	0.01533	0.01527	0.01334	0.01328
22	0.01374	0.01453	0.01408	0.01430	0.01423	0.01260	0.01254
23	0.01334	0.01390	0.01344	0.01380	0.01374	0.01232	0.01225
24	0.01295	0.01329	0.01283	0.01332	0.01325	0.01204	0.01196
25	0.01214	0.01227	0.01183	0.01242	0.01235	0.01136	0.01129
26	0.01180	0.01175	0.01132	0.01200	0.01193	0.01112	0.01104
27	0.01106	0.01085	0.01044	0.01119	0.01112	0.01050	0.01041
28	0.01076	0.01039	0.00999	0.01082	0.01075	0.01028	0.01020
31	0.06361	0.05873	0.05634	0.06290	0.06247	0.06206	0.06145
38	0.04832	0.04014	0.03846	0.04592	0.04557	0.04947	0.04886
45	0.03753	0.02805	0.02698	0.03428	0.03402	0.04032	0.03976
52	0.02980	0.02004	0.01943	0.02616	0.02598	0.03360	0.03312
59	0.02399	0.01452	0.01424	0.02024	0.02013	0.02839	0.02800
66	0.01939	0.01056	0.01051	0.01573	0.01567	0.02408	0.02380
73	0.01586	0.00777	0.00787	0.01236	0.01235	0.02067	0.02051
80	0.01300	0.00573	0.00592	0.00974	0.00976	0.01778	0.01773
87	0.01077	0.00427	0.00451	0.00776	0.00780	0.01546	0.01552
91	0.12561	0.04690	0.05025	0.08844	0.08918	0.18531	0.18697

Table 4.1: Disability distribution determination through Kullback-Leibler divergence and Jensen's difference

	$\mu = 21$		$\mu = 26.8$		$\mu = 38$	
	KL	$Jensen$	KL	$Jensen$	KL	$Jensen$
S	0.02308	**0.0217**	0.0298	**0.0296**	**0.5027**	0.0504
MSE	0.000219	**0.000214**	0.000045	**0.000044**	**0.000108**	0.000112

Table 4.2: Comparison of the disability distributions

results in a logarithmic scale are presented in Figure 4.1.

The results appear nearly equivalent to those presented by London and Brockett. The differences are small. The value of the smoothness measure $S = \sum_{x=1}^{n-3} \left(\Delta^3 v_x\right)^2$ and the goodness of fit measures, i.e. $F = \sum_{x=1}^{n} w_x (u_x - v_x)^2$, where $w_x = l_x/(u_x(1-u_x))$ are weights with l_x being the number of people at risk at age x, log-likelihood $\log L(v) = \sum_{i=1}^{n} [d_x \log v_x + (l_x - d_x)\log(1-v_x)]$, deviance $D(v) = 2\log L(u) - 2\log L(v)$ and $\chi^2 = \sum_{i=1}^{n} \frac{(d_x - l_x v_x)^2}{l_x v_x (1-v_x)}$, are given in Table 4.3(b). The numerical investigation in [32], with same data set, compared the graduations made in [6], [23], and through the use of Cressie-Read power divergence. The overall winner is the graduation through the minimization of the Jensen difference subject to constraints (i) - (v), as judged by smoothness and fidelity. However we believe that constraint (vi) is necessary as this is the minimal requirement for the Jensen difference (and other measures, such as the Kullback - Leibler divergence and the Cressie and Read power divergence) with non-probability vectors to be a measure of divergence.

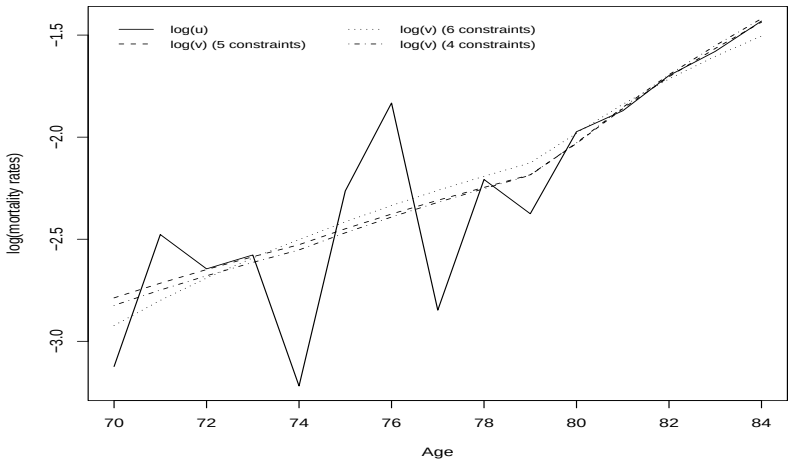

Figure 4.1: Several graduations through the Jensen difference

(a) Graduated values

x	u_x	v_x (5 constraints)	v_x (6 constraints)	v_x (4 constraints)
70	0.044	0.062	0.054	0.059
71	0.084	0.066	0.061	0.064
72	0.071	0.071	0.068	0.069
73	0.076	0.075	0.075	0.073
74	0.040	0.080	0.082	0.078
75	0.104	0.086	0.089	0.085
76	0.160	0.093	0.097	0.092
77	0.058	0.099	0.104	0.098
78	0.110	0.106	0.112	0.105
79	0.093	0.113	0.119	0.112
80	0.139	0.131	0.138	0.132
81	0.154	0.156	0.159	0.157
82	0.183	0.182	0.180	0.184
83	0.206	0.209	0.201	0.212
84	0.239	0.238	0.222	0.242

(b) Smoothness and goodness of fit values

	5 constraints	6 constraints	4 constraints
S	0.000199	0.0002	0.0002
F	16.62	16.70	16.93
Deviance	16.40	16.89	16.48
log-likelihood	-713.12	-713.37	-713.16
χ^2	16.59	16.68	16.93

Table 4.3: Several graduations through Jensen's difference

5 Conclusions

In this paper we studied the use of Jensen's difference in actuarial science as an alternative measure of divergence in problems where the Kullback - Leibler divergence is mainly used. Specifically, we investigated its use in two actuarial problems - the determination of a client's disability distribution and the graduation of mortality rates. Because in the latter case, mortality rates do not form probability vectors, and in order to use $J(\mathbf{p}, \mathbf{q})$ for this purpose, we investigated the properties of the Jensen difference in the case of non-probability vectors. We showed that, under some conditions it is nonnegative, additive and invariant under sufficient transformations. It also shares the property of maximal information and the limiting one. So, we can regard $J(\mathbf{p}, \mathbf{q})$ as a measure of divergence, provided that $\sum_i p_i = \sum_i q_i$, and use it for graduation. Combining with results from our previous research on KL and CR measures [32], this condition should be considered as the minimal requirement for a bivariate function $D(\mathbf{p}, \mathbf{q})$ to be a measure of divergence along with the statement that $D(\mathbf{p}, \mathbf{q}) \geq 0$ with equality if and only if $\mathbf{p} = \mathbf{q}$ when we have non-probability vectors.

We also provided Lagrangian duality results for the problem of minimizing the Jensen difference subject to quadratic and linear inequality constraints. Especially, we derived the Lagrangian dual problem, which proved to be unconstrained, and its solution. These results are important in actuarial science, especially in the problem of graduation.

In the case of the determination of a client's disability distribution, the Jensen difference is a comparable alternative to the Kullback - Leibler divergence. The numerical illustration allows us to propose the following empirical rule: use the Jensen difference when $\mu < \mu_{st}$ while use the Kullback - Leibler divergence when $\mu > \mu_{st}$, where μ_{st} is the mean duration of the standard table. The numerical investigation concerning the graduation of mortality rates indicated that the minimization of the Jensen difference between the crude and graduated rates seems to be the best "divergence" method. Targeting on smoothness and goodness of fit, its results are comparable with those obtained using the Kullback - Leibler directed divergence and the Cressie and Read power divergences.

Acknowledgments. The authors would like to thank the Editor and the Referee for their valuable comments and suggestions.

References

[1] A. Basu, I.R. Harris, N.L. Hjort and M.C. Jones, Robust and Efficient Estimation by Minimising a Density Power Divergence, *Biometrika*, **85**, 3, 1998, 549 - 559.

[2] A. Ben-Tal and A. Charnes, A dual optimization framework for some problems in information theory and statistics, *Problems of Control and Information Theory*, **8**, 1979, 387 - 401.

[3] B. Berliner and B. Lev, On the use of maximum entropy concepts in insurance, *Proc. Ist. Int. Cong. Actuaries*, 1978, 78 - 81.

[4] N.L.Jr. Bowers, H.U. Gerber, J.C. Hickman, D.A. Jones and C.J. Nesbitt, *Actuarial Mathematics*, 2nd ed., Schaumburg, Ill.: Society of Actuaries, 1997.

[5] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, United Kingdom, 2006.

[6] P.L. Brockett, Information Theoretic Approach to Actuarial Science: A Unification and Extension of Relevant Theory and Applications, *Transactions of the Society of Actuaries*, **43**, 1991, 73 - 114.

[7] P.L. Brockett and J. Zhang, Information Theoretical Mortality Graduation, *Scandinavian Actuarial Journal*, 1986, 131 - 140.

[8] V. Brunel, Minimal models for credit risk: an information theory approach, *working paper*, 2004, http://vivienbrunel.free.fr/WorkingPapers/Entropy.pdf.

[9] J. Burbea and C.R. Rao, On the Convexity of Some Divergence Measures Based on Entropy Functions, *IEEE Transactions on Information Theory*, vol **28**, no 3, 1982, 489 - 495.

[10] G.E. Crooks, Inequalities between the Jenson-Shannon and Jeffreys divergences, Technical Report 004, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA, 2008.

[11] I. Csiszar, Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Bewis der Ergodizitat on Markhoffschen Ketten, *Publ. of the Math. Inst. of the Hungarian Academy of Sc.*, **8**, 1963, 84 - 108.

[12] A.H. Darooneh, Non-life insurance pricing: multiagent model, *Insurance: Mathematics and Economics*, **24**, 2004, 23 - 29.

[13] D.M. Endres and J.E. Schindelin, A New Metric for Probability Distributions, *IEEE Transactions on Information Theory*, vol **49**, no 7, 2003, 1858 - 1860.

[14] I. Grosse, P. Bernaola-Galvan, P. Carpena, R. Roman-Roldan, J. Oliver and H.E. Stanley, Analysis of symbolic sequences using the Jensen-Shannon divergence, *PHYSICAL REVIEW*, E, VOLUME **65**, 2002, 1 - 16.

[15] J.N. Kapur, *Maximum-Entropy Models in Science and Engineering*. Wiley, New York, 1989.

[16] S. Kullback, *Information Theory and Statistics*, John Wiley & Sons, New York, 1959.

[17] P.W. Lamberti, A.P. Majtei, A. Borras, M. Casas and A. Plastino, Metric character of the quantum Jensen-Shannon divergence, *PHYSICAL REVIEW A* **77**, 2008, 052311.

[18] E. Landabaru and L. Pardo, Asymptotic behaviour and statistical applications of weighted (h, ϕ) - divergences, *Kybernetes*, Vol. **33**, No 9/10, 2004, 1518 - 1534.

[19] P.J. Laurent, *Approximation et Optimisation*, Hermann, Paris, 1972.

[20] A. Levin and A. Tchernitser, Multifactor stochastic variance models in risk management: Maximum entropy approach and Levy processes, in: *Handbook of Heavy Tailed Distributions in Finance*, (S.T. Rachev, ed), 2003, 443 - 480, Elsevier Science B.V., Amsterdam.

[21] F. Liese and I. Vajda, *Convex Statistical Distances*, B. G. Teubner, Leipzig, 1987.

[22] J. Lin, Divergence measures based on the Shannon entropy, *IEEE Trans. Information Theory*, Vol. **37**, No. 1, 1991, 145 - 151.

[23] D. London, *Graduation: The Revision of Estimates*, ACTEX Publications, Winsted, Connecticut, 1985.

[24] H.J. Luthi and J. Doege, Convex risk measures for portfolio optimization and concepts of flexibility, *Mathematical Programming Series*, B, **104**, 2005, 541 - 559.

[25] D.L. McLeish and R.M. Reesor, Risk, Entropy, and the Transformation of Distributions, *North American Actuarial Journal*, **7** (2), 2003, 128 - 144.

[26] M.L. Menendez, J.A. Pardo, L. Pardo and M.C. Pardo, The Jensen-Shannon Divergence, *J. Franklin Inst.*, Vol. **334**B, No. 2, 1997, 307 - 318.

[27] L. Pardo, *Statistical Inference Based on Divergence Measures*, Chapman and Hall, London, 2006.

[28] L. Pardo, D. Morales and I.J. Taneja, Generalized Jensen difference divergence measures and Fisher measure of information, *Kybernetes*, Vol. **24**, No. 2, 1995, 15 - 28.

[29] T.R.C. Read and N.A.C. Cressie, *Goodness - of - Fit Statistics for Discrete Multivariate Data*, Springer - Verlag, New York, 1998.

[30] J.H. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[31] A. Sachlas and T. Papaioannou, On a minimization problem involving divergences and its applications, in: *Advances in Data Analysis: Theory and Applications to Reliability and Inference, Data Mining, Bioinformatics, Lifetime Data and Neural Networks*, (C. H. Skiadas, Ed.), 2010, 81 - 94, Birkhauser, Boston.

[32] A. Sachlas and T. Papaioannou, Divergences without probability vectors and their applications, *Applied Stochastic Models in Business and Industry*, **26**, 2010, 448 - 472.

[33] C.E. Shannon, A mathematical theory of communication, *Bell System Technical Journal*, **27**, 1948, 379 - 432.

[34] R. Sibson, Information radius, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, Vol. **14**, 1969, 149 - 160.

[35] I.J. Taneja, On generalized information measures and their applications, *Advances in Electronics and Electron Physics*, Vol. **76**, 1989, 327 - 413.

[36] I.J. Taneja, On a Difference of Jensen Inequality and its Applications to Mean Divergence Measures, *RGMIA Research Report Collection*, http://rgmia.vu.edu.au, **7** (4), Art. 16, 2004.

[37] M. Teboulle, A Simple Duality Proof for Quadratically Constrained Entropy Functionals and Extension to Convex Constraints, *SIAM J. Appl. Math.*, Vol. **49**, No. 6, 1989, 1845 - 1850.

[38] L. Xu, D.L. Bricker and K.O. Kortanek, Bounds for stop-loss premium under restrictions on I-divergence, *Insurance: Mathematics and Economics*, **23**, 1987, 119 - 139.

[39] J. Zhang and P.L. Brockett, Quadratically Constrained Information Theoretic Analysis, *SIAM Journal of Applied Mathematics*, Vol. **47**, No. 4, 1987, 871 - 885.

[40] K. Zografos, K. Ferentinos and T. Papaioannou, Limiting Properties of Some Measures of Information, *Annals of the Institute of Statistical Mathematics*, B, Vol. **41**, No. 3, 1989, 451 - 460.

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

Takis Papaioannou is Professor of Statistics, Emeritus, of the Universities of Piraeus and Ioannina, Greece. He received his Ph.D. from Iowa State University and has taught at the Universities of Georgia and McGill. He returned to Greece in 1976 as Professor (Chair) of Probability and Statistics, Department of Mathematics at the University of Ioannina, where he served and developed the statistics unit and laboratory for twenty three years. Then he moved to the University of Piraeus as Professor of Statistics and Chairman of the Department of Statistics and Insurance Science, from which, after serving for five years, he retired. His research interests are statistical information theory, categorical data analysis, biostatistics, statistical inference, and applied statistics. He is author of more than forty papers published in international ISI journals. He has taught at the Universities of Arizona and Cyprus as Visiting Professor. He is a member of the American Statistical Association, the Institute of Mathematical Statistics, and elected member of the International Statistical Institute (ISI). He is also a member and former President of the Greek Statistical Institute.

Athanasios Sachlas received his Ph.D. degree in Statistics from University of Piraeus, Greece. His research focuses on applications of Statistical Information Theory. His research interests also include Biostatistics, Demography, Actuarial Science and Statistical Process Control. Occasionally, he has worked as a consultant to various seminars on statistics. He is currently working as a contract lecturer in the Department of Nursing, University Peloponnese teaching Biostatistics. He is a member of the Greek Statistical Society and of the development and management team of the educational portal www.learn-biostatistics.gr.