Dispersion indices based on Kerridge inaccuracy measure and Kullback-Leibler divergence

Dispersion indices based on Kerridge inaccuracy measure and Kullback-Leibler divergence

Narayanaswamy Balakrishnan[1], Francesco Buono[2], Camilla Calì[3] and Maria Longobardi[4]

[1] McMaster University, Hamilton, Ontario, Canada

bala@mcmaster.ca

[2] Dipartimento di Matematica e Applicazioni "Renato Caccioppoli"

Università degli Studi di Napoli Federico II, Naples, Italy

francesco.buono3@unina.it

[3] Dipartimento di Biologia

Università degli Studi di Napoli Federico II, Naples, Italy

camilla.cali@unina.it

[4] Dipartimento di Biologia

Università degli Studi di Napoli Federico II, Naples, Italy

malongob@unina.it

Keywords: Kerridge inaccuracy measure, Kullback-Leibler divergence, Varentropy.

AMS Subject Classification: 62N05, 60E15, 94A17.

Abstract

Recently, a new dispersion index, as a measures of information, has been introduced and called varentropy. In this paper, we introduce new measures of variability based on two measures of uncertainty, namely, the Kerridge inaccuracy measure and the Kullback-Leibler divergence. Their generating functions are considered and their infinite series representations are given. These new measures and associated properties, bounds and illustrative examples are all presented in detail. Finally, an application of Kullback-Leibler divergence and its dispersion index is illustrated by using the mean-variance rule.

1 Introduction

Let X be a non-negative and absolutely continuous random variable with cumulative distribution function (cdf) F and probability density function (pdf) f. Shannon (1948) introduced a measure of uncertainty as the average level of information associated with the random variable X, known as Shannon entropy or differential entropy. It is defined as

$$H(X) = \mathbb{E}_f[-\log f(X)] = -\int_0^{+\infty} f(x) \log f(x) dx,$$

where log is the natural logarithm. Since then, several properties of Shannon entropy have been studied and different generalizations of this measure have been introduced. One can observe that the Shannon entropy is position-free, in the sense that X and $X + b$, with $b \in \mathbb{R}$, have the same entropy.

The differential entropy has been extended to study the discrepancy between two distributions. In the context of measures discussed in this paper, f and g are two pdf's associated with a single random variable X in problems in which f is the pdf of the "true" distribution of X while g is the pdf suggested by the results of an experiment (Kerridge, 1961). In another viewpoint, f could be suitable to be selected as a model since it is closest to a reference pdf g (Kullback and Leibler, 1951). To be precise, let us consider two absolutely continuous non-negative random variables X and Y with cdf's F, G and pdf's f, g, respectively. If F is the distribution corresponding to the observations and G is the distribution assigned by the experimenter, then the inaccuracy measure of X and Y (also named cross entropy of Y on X or relative distance between X and Y) has been given by Kerridge (1961) as

$$I(f; g) \equiv H_f(g) := \mathbb{E}_f[-\log g(X)] = -\int_0^{+\infty} f(x) \log g(x) dx. \quad (1)$$

As in the previous definition, the inaccuracy is an extension of the entropy $H(X)$. This measure of uncertainty has been widely studied in the literature and has been adapted to different contexts as well (see, for instance, Ghosh and Kundu (2018), Khorashadizadeh (2018) and Kundu et al. (2016)). Also, Taneja and Tuteja (1986) have discussed weighted version of this inaccuracy measure which is a shift dependent measure of uncertainty, while the study of residual and past lifetime distributions through the inaccuracy measure has been carried out by Taneja et al. (2009) and Kumar et al. (2011), respectively. Recently, Balakrishnan et al. (2022) have presented a unified formulation of entropy.

As an information distance between two random variables X and Y, Kullback and Leibler (1951) have introduced a directed divergence defined as

$$K(f:g) = \mathbb{E}_f\left[\log \frac{f(X)}{g(X)}\right] = \int_0^{+\infty} f(x) \log \frac{f(x)}{g(x)} dx. \tag{2}$$

It is also known as information divergence, information gain, relative entropy or discrimination measure. The Kullback-Leibler divergence is a measure of similarity (closeness) between two distributions and plays an important role in information theory, reliability and many other applied fields. Several extensions of this measure have been proposed in the literature; for instance, see Park et al. (2012) and Sunoj et al. (2017). It is important to observe that Kullback-Leibler divergence is non-negative and is equal to 0 if and only if X and Y are identically distributed. This characteristic allows one to use the estimated Kullback-Leibler information as a goodness-of-fit test statistic; see Arizono and Ohta (1989) and Balakrishnan et al. (2007) for pertinent details in this regard. Finally, the Kullback-Leibler divergence and the inaccuracy measure are related as follows:

$$K(f:g) = I(f;g) - H(X). \tag{3}$$

Recently, the study of variability of measures of information has received considerable attention in the literature. In fact, a dispersion index would be useful in understanding about the reliability of the measure. Song (2001) studied the concept of varentropy as an efficient alternative measure for comparing heavy-tailed distributions instead of using the traditional kurtosis measure. In this respect, Fradelizi et al. (2016) have discussed the notion of varentropy defined as

$$VarH(X) \coloneqq Var_f[-\log f(X)] = \int_0^{+\infty} f(x) \log^2 f(x) dx - [H(X)]^2, \tag{4}$$

for which Goodarzi et al. (2017) have presented a useful bound. It is clear that the notation $VarH(X)$ is only a way to denote the varentropy, though it is not the variance of the entropy.

Further, Liu (2007) studied some properties of varentropy under the notion of information volatility which has been followed up in the recent work of Buono et al. (2022).

In this paper, we study the variability of the measures of uncertainty described above. In fact, we point out that these measures can be defined as expectations and their generating functions are then introduced. With these, we can evaluate their dispersion through the variance, in the sense that a measure with a lower level of variance can be considered to be more reliable. The rest of the paper is organized as follows. In Section 2, the definition of the generating function for Shannon entropy and its infinite series representation are given. In Sections 3 and 4, we introduce a dispersion index of Kerridge inaccuracy measure and Kullback-Leibler divergence, respectively. For these new measures, we present some properties, bounds and illustrative examples. In Section 5, we use the mean-variance rule in order to apply the dispersion index of Kullback-Leibler divergence for some illustrative examples.

2 A generating function

A generating function for the Shannon entropy $H(X)$ can be introduced as follows:

$$\begin{aligned} G_{H(X)}(t) &= \mathbb{E}_f\left[e^{t(-\log f(X))}\right] \\ &= \mathbb{E}_f\left[e^{\log\left(\frac{1}{f(X)}\right)^t}\right] = \mathbb{E}_f\left[\frac{1}{(f(X))^t}\right] \\ &= \int_0^{+\infty} \frac{1}{(f(x))^t} f(x)dx. \end{aligned}$$

We can then present an infinite series representation as

$$G_{H(X)}(t) = \mathbb{E}_f\left[1 + \frac{t}{1!}(-\log f(X)) + \frac{t^2}{2!}(-\log f(X))^2 + \frac{t^3}{3!}(-\log f(X))^3 + \dots\right],$$

from which we readily find

$$\left.\frac{d^k}{dt^k} G_{H(X)}(t)\right|_{t=0} = \mathbb{E}_f\left[(-\log f(X))^k\right].$$

In particular, we get $H(X)$ when $k=1$ and $VarH(X)+(H(X))^2$ when $k=2$. Just as $VarH(X)$ is a dispersion index for the Shannon entropy $H(X)$, we can also use the higher-order derivatives of the generating function $G_{H(X)}(t)$ to introduce some other analogous measures associated with $H(X)$. For example,

$$SkewH(X)=\mathbb{E}\left[(-\log f(X))^3\right]-3H(X)\mathbb{E}\left[(\log f(X))^2\right]+2(H(X))^3$$

and

$$\begin{aligned} KurtH(X) &= \mathbb{E}\left[(\log f(X))^4\right]-4H(X)\mathbb{E}\left[(-\log f(X))^3\right] \\ &\quad +6(H(X))^2E\left[(\log f(X))^2\right]-3(H(X))^4 \end{aligned}$$

could be used as suitable measures to explain the skewness (tail tendency) and the kurtosis (tail heaviness) associated with the Shannon entropy $H(X)$.

We now illustrate these measures by considering some examples.

Example 1. Let us take $X \sim Exp(\lambda)$ with pdf $f(x)=\lambda e^{-\lambda x}$, $x>0$, $\lambda>0$. Then,

$$\begin{aligned} G_{H(X)}(t) &= \int_0^{+\infty} \frac{1}{(\lambda e^{-\lambda x})^t}\lambda e^{-\lambda x}dx \\ &= \lambda^{1-t}\int_0^{+\infty} e^{-(1-t)\lambda x}dx \\ &= \frac{\lambda^{1-t}}{(1-t)\lambda}=\frac{\lambda^{-t}}{1-t}\ , \end{aligned}$$

provided $t<1$. Then,

$$\begin{aligned} \left.\frac{d}{dt}G_{H(X)}(t)\right|_{t=0} &= \left.\frac{(1-t)(-\lambda^{-t}\log\lambda)+\lambda^{-t}}{(1-t)^2}\right|_{t=0} \\ &= -\log\lambda+1=H(X). \end{aligned}$$

We have

$$G_{H(X)}(t)=\frac{\lambda^{-t}}{1-t}\ ,$$

(hence, $G_{H(X)}(t)|_{t=0}=1$) and so

$$\log G_{H(X)}(t)=-t\log\lambda-\log(1-t).$$

Consequently, we find

$$\frac{G'_{H(X)}(t)}{G_{H(X)}(t)} = -\log\lambda + \frac{1}{1-t}$$

so that

$$H(X) = G'_{H(X)}(t)|_{t=0} = (-\log\lambda + 1)G_{H(X)}(t)|_{t=0} = -\log\lambda + 1.$$

Next,

$$\frac{G_{H(X)}(t)G''_{H(X)}(t) - \left(G'_{H(X)}(t)\right)^2}{\left(G_{H(X)}(t)\right)^2} = \frac{1}{(1-t)^2}$$

so that

$$G''_{H(X)}(t)|_{t=0} - \left(G'_{H(X)}(t)\right)^2|_{t=0} = 1,$$

which yields

$$G''_{H(X)}(t)|_{t=0} = 1 + (1-\log\lambda)^2.$$

Thence, we find

$$VarH(X) + (H(X))^2 = 1 + (H(X))^2$$

so that

$$VarH(X) = 1.$$

One can similarly derive expressions for $SkewH(X)$ and $KurtH(X)$ measures as well.

Example 2. Next, let us take $X \sim Power(\alpha)$ with pdf $f(x) = \alpha x^{\alpha-1}$, $0 < x < 1$, $\alpha > 0$. Then,

$$\begin{aligned} G_{H(X)}(t) &= \int_0^1 \frac{1}{\alpha^t x^{t(\alpha-1)}} \alpha x^{\alpha-1} dx \\ &= \alpha^{1-t} \int_0^1 x^{(\alpha-1)(1-t)} dx \\ &= \alpha^{1-t} \left. \frac{x^{\alpha-1-t(\alpha-1)+1}}{\alpha-1-t(\alpha-1)+1} \right|_0^1 \\ &= \alpha^{1-t} \frac{1}{\alpha - t(\alpha-1)}, \end{aligned}$$

provided $\alpha - t(\alpha-1) > 0$. So,

$$\begin{aligned}
\frac{d}{dt} G_{H(X)}(t)\Big|_{t=0} &= \frac{(\alpha - t(\alpha-1))\alpha^{1-t}(-\log\alpha) + \alpha^{1-t}(\alpha-1)}{(\alpha - t(\alpha-1))^2}\Big|_{t=0} \\
&= \frac{-\alpha^2\log\alpha + \alpha(\alpha-1)}{\alpha^2} = \frac{\alpha - 1 - \alpha\log\alpha}{\alpha} = H(X).
\end{aligned}$$

In this case, we have

$$G_{H(X)}(t) = \frac{\alpha^{1-t}}{\alpha - (\alpha-1)t} \ ,$$

(hence, $G_{H(X)}(t)|_{t=0} = 1$) and so

$$\log G_{H(X)}(t) = -(1-t)\log\alpha - \log(\alpha - (\alpha-1)t).$$

Consequently, we find

$$\frac{G'_{H(X)}(t)}{G_{H(X)}(t)} = -\log\alpha + \frac{\alpha-1}{\alpha - (\alpha-1)t}$$

so that

$$H(X) = G'_{H(X)}(t)|_{t=0} = -\log\alpha + \frac{\alpha-1}{\alpha},$$

which is monotone decreasing for $\alpha > 1$ and monotone increasing for $\alpha < 1$. Next,

$$\frac{G_{H(X)}(t)G''_{H(X)}(t) - \left(G'_{H(X)}(t)\right)^2}{\left(G_{H(X)}(t)\right)^2} = \frac{(\alpha-1)^2}{(\alpha-(\alpha-1)t)^2}$$

so that

$$G''_{H(X)}(t)|_{t=0} - \left(G'_{H(X)}(t)\right)^2|_{t=0} = \frac{(\alpha-1)^2}{\alpha^2}$$

which yields

$$G''_{H(X)}(t)|_{t=0} = \frac{(\alpha-1)^2}{\alpha^2} + (H(X))^2.$$

Thence, we find

$$VarH(X) + (H(X))^2 = \frac{(\alpha-1)^2}{\alpha^2} + (H(X))^2$$

so that

$$VarH(X) = \left(1 - \frac{1}{\alpha}\right)^2,$$

which is easily seen to be monotone decreasing in α. One can similarly derive expressions for $SkewH(X)$ and $KurtH(X)$ measures.

3 Varinaccuracy

The Kerridge inaccuracy measure can be expressed in terms of expectation of $-\log g(X)$ (see (1)) and for this reason, it is useful to study the variance of this random variable. In the following definition, we introduce the varinaccuray as a dispersion index, also known as cross entropy.

Definition 1. Let X and Y be two non-negative random variables with pdf's f and g, respectively. Then, the varinaccuracy of X and Y is defined as

$$\begin{aligned} VarI(f;g) &:= Var_f[-\log g(X)] \\ &= \int_0^{+\infty} f(x) \log^2 g(x) dx - [I(f;g)]^2 . \end{aligned} \tag{5}$$

Here again, $VarI(f;g)$ does not represent the variance of $I(f;g)$, but is rather a notation. Of course, if X and Y are identically distributed, then, as the inaccuracy measure reduces to the Shannon entropy, the varinaccuracy reduces to the well-known varentropy in (4).

Remark 1. In should be mentioned that Definition 1 could be given in a more general context omitting the non-negativity assumption. In this case, all integrals have to be understood as extended to the common support of X and Y, but its use and interpretation may have to be looked into very carefully.

As done in Section 2, we can introduce a generating function for Kerridge inaccuracy as

$$\begin{aligned} G_{I(f;g)}(t) &= \mathbb{E}_f\left[e^{t(-\log g(X))}\right] \\ &= \mathbb{E}_f\left[e^{\log\left(\frac{1}{g(X)}\right)^t}\right] = \mathbb{E}_f\left[\frac{1}{(g(X))^t}\right] \\ &= \int_0^{+\infty} \frac{1}{(g(x))^t} f(x) dx. \end{aligned}$$

An infinite series representation can then be presented as

$$G_{I(f;g)}(t) = \mathbb{E}_f\left[1 + \frac{t}{1!}(-\log g(X)) + \frac{t^2}{2!}(-\log g(X))^2 + \frac{t^3}{3!}(-\log g(X))^3 + \dots\right],$$

from which we readily find

$$\frac{d^k}{dt^k} G_{I(f;g)}(t)\Big|_{t=0} = \mathbb{E}_f\left[(-\log g(X))^k\right].$$

In particular, we get $I(f;g)$ when $k=1$ and $VarI(f;g)+(I(f;g))^2$ when $k=2$. Using the higher-order derivatives of $G_{I(f:g)}(t)$ we can similarly define the following measures:

$$SkewI(f:g) = \mathbb{E}_f\left[(-\log g(X))^3\right] - 3I(f;g)\mathbb{E}_f\left[(\log g(X))^2\right] + 2(I(f;g))^3,$$

$$\begin{aligned} KurtI(f:g) &= \mathbb{E}_f\left[(\log g(X))^4\right] - 4I(f;g)\mathbb{E}_f\left[(-\log g(X))^3\right] \\ &\quad +6(I(f;g))^2\mathbb{E}_f\left[(\log g(X))^2\right] - 3(I(f;g))^4. \end{aligned}$$

Now, we give some examples to demonstrate the evaluation of varinaccuracy for different distributions.

Example 3. Let $X \sim Exp(\lambda)$ and $Y \sim Exp(\eta)$. Then, by (1), the inaccuracy measure of X and Y is given by

$$I(f;g) = -\int_0^{+\infty} \lambda e^{-\lambda x} \log\left(\eta e^{-\eta x}\right) dx = \frac{\eta}{\lambda} - \log\eta.$$

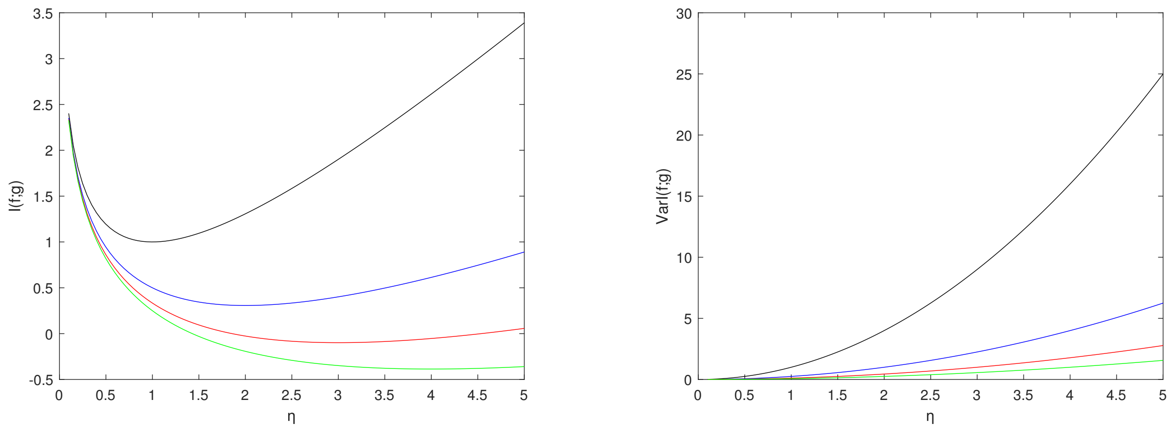
Hence, the varinaccuracy is obtained from (5) as

$$\begin{aligned} VarI(f;g) &= \frac{d^2}{dt^2} G_{I(f;g)}(t)\Big|_{t=0} - (I(f;g))^2 \\ &= \int_0^{+\infty} \lambda e^{-\lambda x} \log^2\left(\eta e^{-\eta x}\right) dx - \left(\frac{\eta}{\lambda} - \log\eta\right)^2 \\ &= \log^2\eta - 2\frac{\eta}{\lambda}\log\eta + 2\frac{\eta^2}{\lambda^2} - \frac{\eta^2}{\lambda^2} - \log^2\eta + 2\frac{\eta}{\lambda}\log\eta = \frac{\eta^2}{\lambda^2}. \end{aligned}$$

In Figure 1, the inaccuracy and the varinaccuracy measures of X and Y are plotted as functions of η, for $\lambda = 1,2,3,4$, with black, blue, red and green lines, respectively. Observe that $I(f;g)$ has its minimum at $\eta = \lambda$ and $VarI(f;g)$ is increasing in η, as one would expect.

Example 4. Let $X \sim U(0,1)$ and Y have power distribution function with parameter $\alpha > 0$ and probability density function g as

$$g(y) = \alpha y^{\alpha-1}, \quad y \in (0,1).$$



(a) $I(f;g)$ (b) $VarI(f;g)$

Figure 1: Plot of I and $VarI$ in Example 3 as a function of η for $\lambda = 1, 2, 3, 4$.

Then, by (1), the inaccuracy measure of X and Y is given by

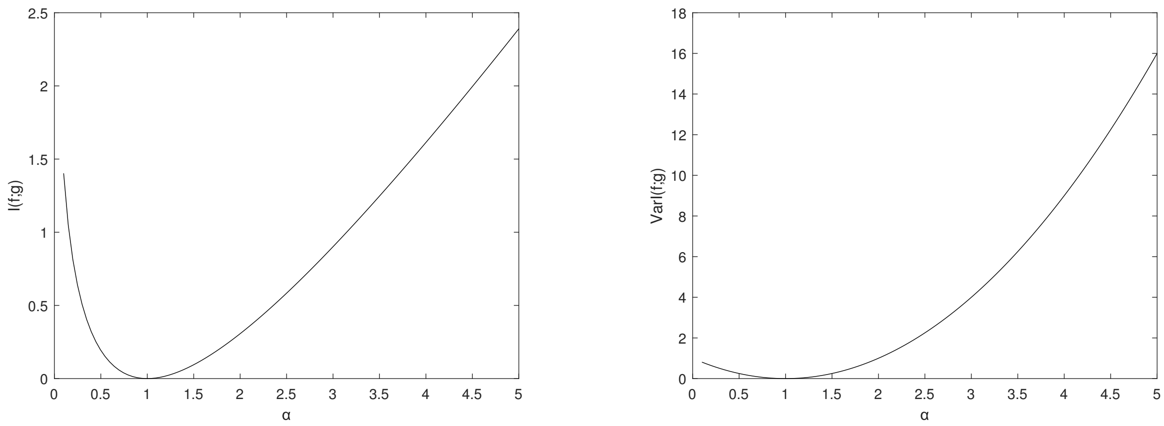
$$I(f;g) = -\int_0^1 \log\left(\alpha x^{\alpha-1}\right) dx = \alpha - 1 - \log\alpha.$$

Hence, the varinaccuracy is obtained from (5) as

$$\begin{aligned} VarI(f;g) &= \frac{d^2}{dt^2} G_{I(f;g)}(t)\bigg|_{t=0} - (I(f;g))^2 \\ &= \int_0^1 \log^2\left(\alpha x^{\alpha-1}\right) dx - (\alpha - 1 - \log\alpha)^2 \\ &= \log^2\alpha - 2(\alpha-1)\log\alpha + 2(\alpha-1)^2 \\ &\quad -(\alpha-1)^2 - \log^2\alpha + 2(\alpha-1)\log\alpha = (\alpha-1)^2. \end{aligned}$$

In Figure 2, the inaccuracy and the varinaccuracy measures of X and Y are plotted as functions of α. In this case, the inaccuracy reaches the minimum at $\alpha = 1$, that is, when Y also has a uniform distribution in $(0, 1)$, and $VarI$ is monotone decreasing for $\alpha < 1$ and monotone increasing for $\alpha > 1$.

In the following proposition, we examine the behaviour of varinaccuracy under affine transformations. The proof of it follows by a simple change of variable technique and is similar to the property of Shannon entropy along with invariance of the variance under translation.



(a) $I(f;g)$ (b) $VarI(f;g)$

Figure 2: Plot of I and $VarI$ in Example 4 as a function of α.

Proposition 3.1. *Let X and Y be two random variables with common support S and pdf's f and g, respectively. Let $a > 0$, $b \geq 0$ and the variables $\tilde{X}$, $\tilde{Y}$ be $\tilde{X} = aX + b$, $\tilde{Y} = aY + b$ with pdf's $\tilde{f}$ and $\tilde{g}$, respectively. Then, we have*

$$VarI(\tilde{f};\tilde{g}) = VarI(f;g).$$

Proposition 3.2. *Let X and Y be two random variables with common support S and pdf's f and g, respectively. Let ϕ be a strictly monotone function and the variables $\tilde{X}$, $\tilde{Y}$ be $\tilde{X} = \phi(X)$, $\tilde{Y} = \phi(Y)$ with pdf's $\tilde{f}$ and $\tilde{g}$, respectively. Then, we have*

$$VarI(\tilde{f};\tilde{g}) = VarI(f;g) + Var_f[\log|\phi'(X)|] - 2cov_f(\log g(X), \log|\phi'(X)|).$$

Proof. Without loss of generality, let us take $S = (0, +\infty)$ with ϕ strictly increasing from $\phi(0)$ to $+\infty$. Then, the common support of $\tilde{X}$ and $\tilde{Y}$ is $(\phi(0), +\infty)$. The relation between the pdf's of $\tilde{X}$, $\tilde{Y}$ and X, Y is given by

$$\tilde{f}(x) = \frac{f(\phi^{-1}(x))}{\phi'(\phi^{-1}(x))}, \quad \tilde{g}(x) = \frac{g(\phi^{-1}(x))}{\phi'(\phi^{-1}(x))}, \quad x \in (\phi(0), +\infty),$$

where $\frac{1}{\phi'(\phi^{-1}(x))} = \frac{d}{dy}\phi^{-1}(y)\Big|_{y=\phi^{-1}(x)}$. Then, the inaccuracy measure of $\tilde{X}$ and $\tilde{Y}$ can be expressed as

$$\begin{aligned} I(\tilde{f};\tilde{g}) &= -\int_{\phi(0)}^{+\infty} \frac{f(\phi^{-1}(x))}{\phi'(\phi^{-1}(x))} \log\left[\frac{g(\phi^{-1}(x))}{\phi'(\phi^{-1}(x))}\right] dx \\ &= I(f;g) + \mathbb{E}_f[\log\phi'(X)]. \end{aligned}$$

Then, the varinaccuracy of $\tilde{X}$ and $\tilde{Y}$ can be obtained as

$$\begin{aligned} VarI(\tilde{f};\tilde{g}) &= \int_{\phi(0)}^{+\infty} \frac{f(\phi^{-1}(x))}{\phi'(\phi^{-1}(x))} \log^2\left[\frac{g(\phi^{-1}(x))}{\phi'(\phi^{-1}(x))}\right] dx \\ &\quad -[I(f;g) + \mathbb{E}_f[\log\phi'(X)]]^2 \\ &= \int_0^{+\infty} f(x)\log^2 g(x)dx + \int_0^{+\infty} f(x)\log^2(\phi'(x))dx \\ &\quad -2\int_0^{+\infty} f(x)\log(\phi'(x))\log(g(x))dx - [I(f;g) + \mathbb{E}_f[\log\phi'(X)]]^2 \\ &= VarI(f;g) + Var_f[\log\phi'(X)] - 2cov_f(\log g(X), \log\phi'(X)), \end{aligned}$$

as required. □

Proposition 3.3. *Let X and Y be two random variables with common support S and pdf's f and g, respectively. Then, $VarI(f;g) = 0$ if and only if Y is uniformly distributed in S.*

Proof. The varinaccuracy is defined as a variance, which vanishes only for degenerate distributions. In particular, $\log g(x)$ needs to be constant for $x \in S$, i.e., g needs to be a constant function in which case Y is uniformly distributed in S. □

In the following proposition, we obtain a lower bound for varinaccuracy based on Chebyshev inequality which, for a random variable W with mean $\mathbb{E}(W)$ and variance $Var(W)$, is given by

$$\mathbb{P}\left(|W - \mathbb{E}(W)| < \varepsilon\right) \geq 1 - \frac{Var(W)}{\varepsilon^2}, \quad \varepsilon > 0. \tag{6}$$

Proposition 3.4. *Let X and Y be two random variables with common support S and pdf's f and g, respectively, and let $\varepsilon > 0$. Then, a lower bound for varinaccuracy is given by*

$$VarI(f;g) \geq \varepsilon^2\left[\mathbb{P}\left(g(X) \leq e^{-\varepsilon - I(f;g)}\right) + \mathbb{P}\left(g(X) \geq e^{\varepsilon - I(f;g)}\right)\right]. \tag{7}$$

Proof. Based on the definitions of inaccuracy and varinaccuracy in (1) and (5), Chebyshev inequality in (6) yields

$$VarI(f;g) \geq \varepsilon^2 \mathbb{P}(|\log g(X) + I(f;g)| \geq \varepsilon). \tag{8}$$

The second factor on the right hand side of the above equation can be written as

$$\begin{aligned}
&\mathbb{P}(|\log g(X) + I(f;g)| \geq \varepsilon) \\
&= \mathbb{P}\left(\log g(X) + I(f;g) \leq -\varepsilon\right) + \mathbb{P}\left(\log g(X) + I(f;g) \geq \varepsilon\right) \\
&= \mathbb{P}\left(g(X) \leq e^{-\varepsilon - I(f;g)}\right) + \mathbb{P}\left(g(X) \geq e^{\varepsilon - I(f;g)}\right),
\end{aligned} \tag{9}$$

and the proof then gets completed by combining (8) and (9). □

In the following corollaries, we specialize the result of Proposition 3.4 to the cases when g is strictly increasing and decreasing.

Corollary 3.1. *Let X and Y be two random variables with common support S, pdf's f and g and cdf's F and G, respectively, and let $\varepsilon > 0$. If g is strictly decreasing in S, then*

$$VarI(f;g) \geq \varepsilon^2 \left[\overline{F}\left(g^{-1}(e^{-\varepsilon - I(f;g)})\right) + F\left(g^{-1}(e^{\varepsilon - I(f;g)})\right)\right], \tag{10}$$

where $\overline{F}(\cdot) = 1 - F(\cdot)$ is the survival function of X.

Corollary 3.2. *Let X and Y be two random variables with common bounded support S, pdf's f and g and cdf's F and G, respectively, and let $\varepsilon > 0$. If g is strictly increasing in S, then*

$$VarI(f;g) \geq \varepsilon^2 \left[F\left(g^{-1}(e^{-\varepsilon - I(f;g)})\right) + \overline{F}\left(g^{-1}(e^{\varepsilon - I(f;g)})\right)\right]. \tag{11}$$

Example 5. Let $X \sim Exp(\lambda)$ and $Y \sim Exp(\eta)$. In Example 3, we have plotted the varinaccuracy measure of X and Y. Here, we use Corollary 3.1 to evaluate a lower bound. In fact, in this case, the pdf g of Y is strictly decreasing and we have

$$g^{-1}(z) = -\frac{1}{\eta} \log \frac{z}{\eta}, \quad z \in (0, \eta).$$

Moreover, the inaccuracy measure of X and Y is given by

$$I(f;g) = -\log\eta + \frac{\eta}{\lambda}.$$

If $\varepsilon\lambda > \eta$, we have

$$e^{\varepsilon - I(f;g)} > \eta,$$

and then $\mathbb{P}(g(X) \geq e^{\varepsilon - I(f;g)}) = 0$. Thus, we can conclude

$$VarI(f;g) \geq \begin{cases} \varepsilon^2 \left(e^{-1-\varepsilon\lambda/\eta} + 1 - e^{-1+\varepsilon\lambda/\eta}\right), & \text{if } \varepsilon\lambda \leq \eta \\ \varepsilon^2 e^{-1-\varepsilon\lambda/\eta}, & \text{if } \varepsilon\lambda > \eta. \end{cases} \tag{12}$$

In Figure 3, we have plotted the varinaccuracy and the bound in the case $\lambda = 4$ as a function of η for different choices of ε.

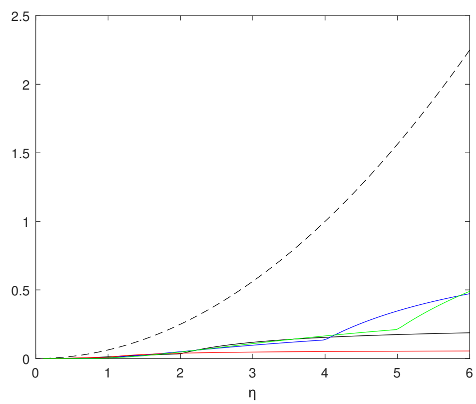


Figure 3: Plot of $VarI(f;g)$ (dashed line) and lower bounds in Example 5 as a function of η for $\lambda = 4$ and $\varepsilon = 0.25, 0.5, 1, 1.25$ (red, black, blue and green lines, respectively).

Example 6. Let $X \sim U(0,1)$ and $Y \sim Power(\alpha)$ with $\alpha > 1$. In Example 4, we have plotted the varinaccuracy measure of X and Y. Here, Corollary 3.2 is used to evaluate the lower bound. In fact, in this case, the pdf g of Y is strictly increasing, and we have

$$g^{-1}(z) = \left(\frac{z}{\alpha}\right)^{\frac{1}{\alpha-1}}, \quad z \in (0,\alpha).$$

Moreover, the inaccuracy measure of X and Y is given by

$$I(f;g) = -\log\alpha + (\alpha - 1).$$

If $1 < \alpha < 1+\varepsilon$, we have

$$e^{\varepsilon - I(f;g)} > \alpha,$$

and then $\mathbb{P}(g(X) \geq e^{\varepsilon - I(f;g)}) = 0$. Thus, we can conclude

$$VarI(f;g) \geq \begin{cases} \varepsilon^2 \left(e^{\frac{1-\varepsilon-\alpha}{\alpha-1}} + 1 - e^{\frac{1+\varepsilon-\alpha}{\alpha-1}} \right), & \text{if } \alpha \geq 1+\varepsilon \\ \varepsilon^2 e^{\frac{1-\varepsilon-\alpha}{\alpha-1}}, & \text{if } 1 < \alpha < 1+\varepsilon. \end{cases} \tag{13}$$

In Figure 4, we have plotted the varinaccuracy and the bound as a function of α for different choices of ε.

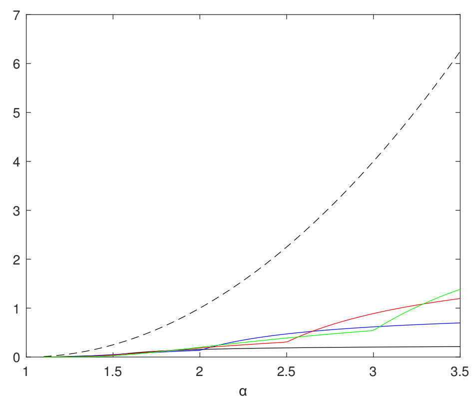


Figure 4: Plot of $VarI(f;g)$ (dashed line) and lower bounds in Example 6 as a function of α for $\varepsilon = 0.5, 1, 1.5, 2$ (black, blue, red and green lines, respectively).

4 A dispersion index for Kullback-Leibler divergence

In the following definition, we introduce a dispersion index for Kullback-Leibler divergence based on (2).

Definition 2. Let X and Y be two non-negative random variables with pdf's f and g, respectively. Then, a dispersion index for Kullback-Leibler divergence of X and Y is defined as

$$
\begin{aligned}
VarK(f:g) &:= Var_f\left[\log\frac{f(X)}{g(X)}\right]\\
&= \mathbb{E}_f\left[\log^2\frac{f(X)}{g(X)}\right]-\left[K(f:g)\right]^2\\
&= \int_0^{+\infty} f(x)\log^2\frac{f(x)}{g(x)}dx-\left[K(f:g)\right]^2. \qquad (14)
\end{aligned}
$$

It is important to keep in mind here again that $VarK(f:g)$ does not represent the variance of Kullback-Leibler divergence, but is rather a short notation.

Remark 2. As mentioned earlier for the varinaccuracy measure, the definition of $VarK$ can also be given for variables with a common support S not necessarily equal to $(0,+\infty)$, but its use and interpretation may need to be examined carefully.

In analogy with Section 2, we can introduce a generating function for $K(f:g)$ as

$$
\begin{aligned}
G_{K(f:g)}(t) &= \mathbb{E}_f\left[e^{t(\log f(X)-\log g(X))}\right]\\
&= \mathbb{E}_f\left[e^{\log\left(\frac{f(X)}{g(X)}\right)^t}\right]=\mathbb{E}_f\left[\left(\frac{f(X)}{g(X)}\right)^t\right].
\end{aligned}
$$

We can then present an infinite series representation for it as

$$
\begin{aligned}
G_{K(f:g)}(t) &= \mathbb{E}_f\Bigg[1+\frac{t}{1!}(\log f(X)-\log g(X))+\frac{t^2}{2!}(\log f(X)-\log g(X))^2\\
&\quad+\frac{t^3}{3!}(\log f(X)-\log g(X))^3+\dots\Bigg],
\end{aligned}
$$

from which we readily find

$$
\left.\frac{d^k}{dt^k}G_{K(f:g)}(t)\right|_{t=0}=\mathbb{E}_f\left[(\log f(X)-\log g(X))^k\right]. \qquad (15)
$$

In particular, when $k=1$, we get $K(f:g)$, and when $k=2$, we get $\mathbb{E}_f[(\log f(X)-\log g(X))^2]$ which is related to $VarK(f:g)$. We can then proceed to define $SkewK(f:g)$ and $KurtK(f:g)$ measures, based on higher-order derivatives of the generating functions $G_{K(f:g)}(t)$, in a manner similar to what was done in the preceding sections.

In the following proposition, in analogy with the relation in (3), we present a connection between varentropy, varinaccuracy and $VarK$ measures.

Proposition 4.1. *Let X and Y be two non-negative random variables with common support S and pdf's f and g, respectively. Then,*

$$VarK(f:g) = VarH(X) + VarI(f;g) - 2cov_f(\log f(X), \log g(X)). \tag{16}$$

Proof. From (14) and by taking into account the expression of the variance of the sum, we obtain

$$\begin{aligned} VarK(f:g) &= Var_f\left[\log \frac{f(X)}{g(X)}\right] = Var_f\left[\log f(X) - \log g(X)\right] \\ &= Var_f[\log f(X)] + Var_f[\log g(X)] - 2cov_f(\log f(X), \log g(X)) \\ &= Var_f[-\log f(X)] + Var_f[-\log g(X)] \\ &\quad -2cov_f(\log f(X), \log g(X)) \end{aligned}$$

and, by recalling (4) and (5), we get the required result. □

Proposition 4.2. *Let X and Y be two random variables with common support S and pdf's f and g, respectively. Then, $VarK(f:g) = 0$ if and only if X and Y are identically distributed.*

Proof. $VarK$ is defined as a variance, and so vanishes only for degenerate distributions. In particular, $\log \frac{f(x)}{g(x)}$ needs to be constant for $x \in S$, i.e.,

$$\frac{f(x)}{g(x)} = a, \quad x \in S,$$

where a is a non-negative constant. In view of the normalization condition, we have $a = 1$ and so X and Y are identically distributed. □

Remark 3. Proposition 4.2 enables one to consider $VarK$ as a measure of divergence since it shares the positive-definiteness property with the Kullback-Leibler divergence. Moreover, as the Kullback-Leibler divergence, it can not be considered as a metric since it is not symmetric and does not satisfy the triangular inequality. The former is quite intuitive from the definition, while the latter is shown by the following counterexample. Let X, Y and Z

follow the power distribution function with parameters 0.5, 3 and 2, and let us denote the corresponding pdf's by f, g, and h, respectively. We then readily find

$$VarK(f:g) = 25, \quad VarK(f:h) = 9, \quad VarK(h:g) = 0.25,$$

so that

$$VarK(f:g) > VarK(f:h) + VarK(h:g),$$

and so the triangular inequality is not satisfied.

For furher developments, it would be of interest to analyze the relationships between this new divergence measure and some well-known measures such as Kullback-Leibler, Rényi, Cressie-Read and Chernoff α divergences (see Bedbur and Kamps (2021) for their definitions).

5 VarK applications in testing the underlying distribution

The Kullback-Leibler divergence is a measure of similarity between two distributions. If we consider X to be distributed as the observed data, then we can choose Y in different ways and then compare the values of $K(f:g)$, where f and g are the pdf's of X and Y, respectively. Of course, a lower value of Kullback-Leibler divergence corresponds to a higher similarity between the distributions of Y and that of the data. There may be situations in which Y_1 and Y_2 are two different random variables with pdf's g_1 and g_2, respectively, such that $K(f:g_1) \simeq K(f:g_2)$. In this case, we can choose the more suitable distribution by considering $VarK$, in the sense that we would prefer a distribution with a lower variance even if it has a slightly higher or similar value of K, between the two chosen models.

In order to obtain a criterion based on Kullback-Leibler divergence and the related dispersion index, we set a threshold r such that if $K(f:g_i)$, $i = 1, 2$, exceeds the value r, we can not accept such a distribution. To be specific, let us suppose $K(f:g_1) \leq K(f:g_2)$.

Moreover, consider the case in which $K(f:g_2) < r$. As $K(f:g_2)$ tends to r, it becomes more difficult to prefer Y_2 to Y_1, but we can tolerate a higher value of Kullback-Leibler divergence if we balance with a lower value of variance. We may then use $VarK$ to standardize the difference between r and K and make comparisons. We would prefer Y_2 to Y_1 if the following inequality is satisfied:

$$\frac{r - K(f:g_1)}{\sqrt{VarK(f:g_1)}} < \frac{r - K(f:g_2)}{\sqrt{VarK(f:g_2)}}. \tag{17}$$

Remark 4. Observe that the criterion in (17) is reasonable since when $K(f:g_1) = K(f:g_2)$, the variable with lower $VarK$ is preferred. Moreover, with the same variance, the variable with lower Kullback-Leibler divergence is still preferred. Finally, if Y_1 has lower K as well as $VarK$, it will be preferred to Y_2.

In order to apply the criterion to concrete situations, we have to choose a value for the threshold r. It may not be possible to fix a numerical value for r, but we can relate this quantity to the Kullback-Leibler divergences. In particular, we choose $r = 2K(f:g_1)$, where $K(f:g_1) \leq K(f:g_2)$. Hence, the criterion in (17) can be reformulated in the following manner: Y_2 is preferred to Y_1 if the following inequality is satisfied:

$$\frac{K(f:g_1)}{\sqrt{VarK(f:g_1)}} < \frac{2K(f:g_1) - K(f:g_2)}{\sqrt{VarK(f:g_2)}},$$

which is equivalent to

$$K(f:g_2) < \left(2 - \sqrt{\frac{VarK(f:g_2)}{VarK(f:g_1)}}\right) K(f:g_1). \tag{18}$$

Remark 5. The same dispersion index given in Definiton 2 can also be introduced in the discrete case. When we have two discrete probability distributions P and Q defined on the same probability space $\mathcal{X}$, the Kullback-Leibler divergence of P and Q is defined as

$$K(P:Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}. \tag{19}$$

The corresponding index of dispersion is then

$$VarK(P:Q) = \sum_{x \in \mathcal{X}} P(x) \log^2 \frac{P(x)}{Q(x)} - [K(P:Q)]^2. \tag{20}$$

In the following, we illustrate three applications of the above method in different scenarios. In the first one, we will have Kullback-Leibler divergences which do not satisfy similarity property. In the second one, we will present the case in which we have two equal Kullback-Leibler divergences. In the final one, we will present the more important situation, i.e., we will find a distribution with lower K, but with higher $VarK$.

Example 7. The data presented in Table 1 were obtained from 200 repetitions of an experiment involving 3 tosses of a coin and then recording the number of heads observed.

Table 1: Data of Example 7.

Number of heads	0	1	2	3
Number of observations	20	63	84	33

If we denote by X the random variable distributed as the data, from Table 1, we get the distribution of X as

$$p_0 = \mathbb{P}(X = 0) = 0.1, \quad p_1 = 0.315, \quad p_2 = 0.42, \quad p_3 = 0.165.$$

Our intention is to establish a suitable distribution for these data, for which we determine Kullback-Leibler divergence and its variance between X and three different variables Y_1, Y_2, Y_3, with probability mass functions P, Q_1, Q_2, Q_3, respectively. In particular, Y_1 follows a binomial distribution $B(3, 0.55)$, where 0.55 is obtained by maximum likelihood estimation, Y_2 follows a beta-binomial distribution with parameters $n = 3$, $\alpha = 12$ and $\beta = 10$, and Y_3 follows a discrete uniform distribution over the four elements. The values of Kullback-Leibler divergence and its variance are presented in Table 2. As the binomial distribution has lower Kullback-Leibler divergence and lower $VarK$, we can conclude that the binomial distribution is more appropriate than the Beta-binomial and the discrete uniform distributions. We also see that the Beta-binomial distribution is preferred to the discrete uniform.

Table 2: $K(P:Q_i)$ and $VarK(P:Q_i)$, for $i=1,2,3$, in Example 7.

Distribution	$K(P:Q)$	$VarK(P:Q)$
Binomial	0.0011	0.0023
Beta-binomial	0.0027	0.0054
Uniform	0.1305	0.2253

Example 8. Consider the real data (see Data Set 4.1 in Murthy et al. (2004)) presenting times to failure of 20 units: $11.24, 1.92, 12.74, 22.48, 9.60, 11.50, 8.86, 7.75, 5.73, 9.37, 30.42, 9.17, 10.20, 5.52, 5.85, 38.14, 2.99, 16.58, 18.92, 13.36$. The data are distributed as the random variable X whose pdf is f. We estimate the density function through a kernel estimator with *MATLAB* function `ksdensity`. In order to establish if the distribution of the data is similar to a Weibull distribution $W2(\alpha,\lambda)$ with pdf

$$g(x)=\lambda\alpha x^{\alpha-1}\exp\left(-\lambda x^{\alpha}\right), \quad x>0,$$

we consider two different Weibull distribution, $Y_1 \sim W2(1.5487, 0.0166)$, with parameters given by the maximum likelihood method, and $Y_2 \sim W2(1.6, 0.0127)$. Note that both distributions are accepted by using Kolmogorov-Smirnov test at a significance level of 5%. In Figure 5, we have plotted the estimated pdf of the data and the pdf's g_1, g_2 of Y_1, Y_2. With these distributions, we obtain

$$K(f:g_1)=K(f:g_2)=0.0990.$$

Hence, to choose the more suitable distribution, we have to compare the values of $VarK$, for which we obtain

$$VarK(f:g_1)=0.3350 \;>\; VarK(f:g_2)=0.2936.$$

So, we choose Y_2 since its Kullback-Leibler divergence has a lower variability.

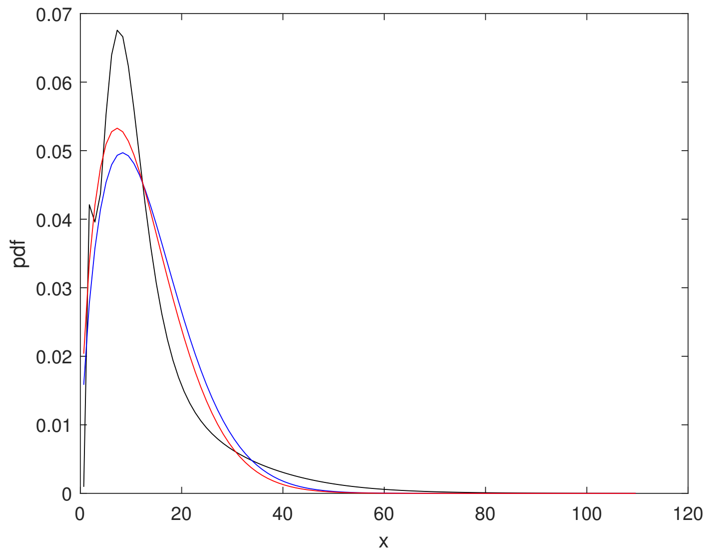


Figure 5: Plot of pdf's of X, Y_1, Y_2 in Example 8 (black, red and blue lines, respectively).

Example 9. Consider the crab dataset presented in Murphy and Aha (1994). We focus on the distribution of the width of female crabs, represented by the random variable X with pdf f; then, we have a sample of 100 units. Here again, we first estimate the density function through a kernel estimator with *MATLAB* function `ksdensity`. We now wish to compare the distribution of the data with Weibull and Lognormal distributions. Recall that if $Y_2 \sim Lognormal(\mu, \sigma)$, then the pdf is given by

$$g_2(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log x - \mu)^2}{2\sigma^2}\right), \quad x > 0.$$

In particular, by using the maximum likelihood estimation, we choose $Y_1 \sim W2(5.6162, 1.1953e\text{-}09)$ and $Y_2 \sim Lognormal(3.5559, 0.2192)$. Note that both distributions are accepted by using Kolmogorov-Smirnov test at a significance level of 5%. In Figure 6, we have plotted the estimated pdf of data and pdf's of Y_1, Y_2. With these distributions, we obtain

$$K(f : g_1) = 0.0381, \qquad K(f : g_2) = 0.0420,$$
$$VarK(f : g_1) = 0.1148, \quad VarK(f : g_2) = 0.0924.$$

Hence, we are in the case in which Y_1 has lower Kullback-Leibler divergence but higher $VarK$, and the difference between $K(f : g_2)$ and $K(f : g_1)$ is small enough. Then, in

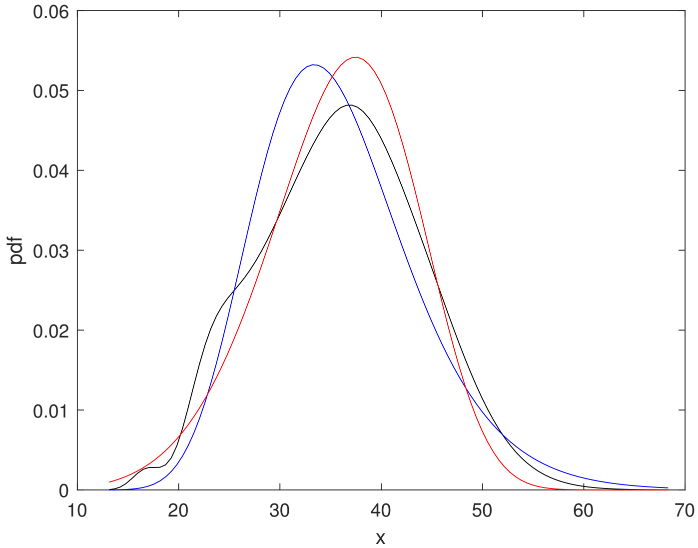


Figure 6: Plot of pdf's of X, Y_1, Y_2 in Example 9 (black, red and blue lines, respectively).

order to choose the most suitable distribution, we use the criterion in (18) and compute the difference

$$K(f:g_2) - \left(2 - \sqrt{\frac{VarK(f:g_2)}{VarK(f:g_1)}}\right) K(f:g_1) = -3.8085e\text{-}05.$$

Thus, the inequality in (18) is satisfied, and so we choose Y_2 as the distribution that fits the data the best.

6 Conclusion

In this paper, we have introduced new measures of variability for some measures of uncertainty, and specifically for the Kerridge inaccuracy measure and the Kullback-Leibler divergence. We have presented the generating functions of these measures and of Shannon entropy. We have defined a dispersion index based on the Kerridge inaccuracy, $VarI$, named varinaccuracy. We have discussed the effect of linear transformations and strict monotone functions on varinaccuracy and then have presented lower bounds. A dispersion index of Kullback-Leibler divergence, $VarK$, and a connection among varentropy, varinaccuracy and

$VarK$ have been given. As the Kullback-Leibler divergence is a measure of similarity between two distributions, $VarK$ has been used to compare two distributions chosen to fit the data. In order to obtain a criterion based on Kullback-Leibler divergence and its variance, we have used the mean-variance rule and some examples have then been presented to illustrate the results and methods developed here. Further analysis of these dispersion indices could be done for comparing distributions under different assumptions like shape characteristics and reliability properties.

Acknowledgements

Francesco Buono, Camilla Calì and Maria Longobardi are members of the research group GNAMPA of INdAM (Istituto Nazionale di Alta Matematica). Francesco Buono and Maria Longobardi are partially supported by MIUR - PRIN 2017, project "Stochastic Models for Complex Systems", no. 2017 JFFHSH. The present work was developed within the activities of the project 000009_ALTRI_CDA_75_2021_FRA_LINEA_B funded by "Programma per il finanziamento della ricerca di Ateneo - Linea B" of the University of Naples Federico II. The authors also express their sincere thanks to the reviewers for their useful comments and suggestions which resulted in this much improved version.

References

[1] Arizono, I., Ohta, H. 1989. *A test for normality based on Kullback–Leibler information.* The American Statistician, 43, 20–22.

[2] Balakrishnan, N., Buono, F., Longobardi, M. 2022. *A unified formulation of entropy and its application.* Physica A, 596, 127214.

[3] Balakrishnan, N., Rad, A.H., Arghami, N.R. 2007. *Testing exponentiality based on Kullback-Leibler information with progressively Type-II censored data.* IEEE Transactions on Reliability, 56, 349–356.

[4] Bedbur, S. Kamps, U. 2021. *On representations of divergence measures and related quantities in exponential families.* Entropy, 23, 726.

[5] Buono, F., Longobardi, M., Pellerey, F. 2022. *Varentropy of past lifetimes*. Mathematical Methods of Statistics, 31, 57–73.

[6] Fradelizi, M., Madiman, M., Wang, L. 2016. *Optimal concentration of information content for log-concave densities*. High Dimensional Probability VII, 45–60.

[7] Ghosh, A., Kundu, C. 2018. *On generalized conditional cumulative past inaccuracy measure*. Applications of Mathematics, 63, 167–193.

[8] Goodarzi, F., Amini, M., Borzadaran, G. 2017. *Characterizations of continuous distributions through inequalities involving the expected values of selected functions*. Applications of Mathematics, 62, 493–507.

[9] Kerridge, D. F. 1961. *Inaccuracy and inference*. J. R. Stat. Soc., Ser. B, 23, 184–194.

[10] Khorashadizadeh, M. 2018. *More results on dynamic cumulative inaccuracy measure*. Journal of the Iranian Statistical Society, 17, 89–108.

[11] Kullback, S., Leibler, R.A., 1951. *On information and sufficiency*. Ann. Math. Statist., 22, 79–86.

[12] Kumar, V., Taneja, H.C., Srivastava, R. 2011. *A dynamic measure of inaccuracy between two past lifetime distributions*. Metrika, 74, 1–10.

[13] Kundu, C., Di Crescenzo, A., Longobardi, M. 2016. *On cumulative residual (past) inaccuracy for truncated random variables*. Metrika, 79, 335–356.

[14] Liu, J. 2007. *Information Theoretic Content and Probability*. Ph.D. Thesis, University of Florida, Gainesville, FL, USA.

[15] Murphy, P.M., Aha, D.W. 1994. *UCI Repository of machine learning databases*, http://www.ics.uci.edu/ mlearn/MLRepository.html. Irvine, CA: University of California, Department of Information and Computer Science.

[16] Murthy, D.N.P., Xie, M., Jiang, R. 2004. *Weibull Models*. Hoboken: John Wiley & Sons.

[17] Park, S., Rao, M., Shin, D.W. 2012. *On cumulative residual Kullback-Leibler information*. Statistics and Probability Letters, 82, 2025–2032.

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

[19] Song, K.S. 2001. *Rényi information, log likelihood and an intrinsic distribution measure*. Journal of the Statistical Planning and Inference, 93, 51–69.

[20] Sunoj, S.M., Sankaran, P.G., Unnikrishnan Nair, N. 2017. *Quantile-based cumulative Kullback–Leibler divergence*. Statistics, 52, 1–17.

[21] Taneja, H.C., Kumar, V., Srivastava, R. 2009. *A dynamic measure of inaccuracy between two residual lifetime distributions*. International Mathematical Forum, 4, 1213–1220.

[22] Taneja, H.C., Tuteja, R.K. 1986. *Characterization of a quantitative-qualitative measure of inaccuracy.* Kybernetika, 22, 393–402.