British Journal of Mathematics & Computer Science
21(2): 1-16, 2017, Article no.BJMCS.32229
ISSN: 2231-0851

SCIENCEDOMAIN *international*
www.sciencedomain.org

Hellinger Distance Between Generalized Normal Distributions

C. P. Kitsos[1] **and T. L. Toulias**[2*]

[1]*Department of Informatics, Technological Educational Institute of Athens, Athens, Greece.*
[2]*Avenue Charbo 20, Schaerbeek 1030, Brussels, Belgium.*

Authors' contributions

This work was carried out in collaboration between both authors. Author CPK provided the generalized normal distribution and the information relative risk framework. Author TLT performed all the mathematical computations. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/BJMCS/2017/32229
Editor(s):
(1) Andrej V. Plotnikov, Department of Applied and Calculus Mathematics and CAD, Odessa State Academy of Civil Engineering and Architecture, Ukraine.
Reviewers:
(1) John Tumaku, Ho Technical University, Ghana.
(2) Anjali Munde, Amity University, India.
Complete Peer review History: http://www.sciencedomain.org/review-history/18235

Review Article

Received: 15th February 2017
Accepted: 9th March 2017
Published: 16th March 2017

Abstract

A relative measure of informational distance between two distributions is introduced in this paper. For this purpose the Hellinger distance is used as it obeys to the definition of a distance metric and, thus, provides a measure of informational "proximity" between of two distributions. Certain formulations of the Hellinger distance between two generalized Normal distributions are given and discussed. Motivated by the notion of Relative Risk we introduce a relative distance measure between two continuous distributions in order to obtain a measure of informational "proximity" from one distribution to another. The Relative Risk idea from logistic regression is then extended, in an information theoretic context, using an exponentiated form of Hellinger distance.

Keywords: Generalized γ-order normal distribution; kullback-Leibler divergence; hellinger distance; relative risk.

2010 Mathematics Subject Classification: 94A17, 97K50.

**Corresponding author: E-mail: th.toulias@gmail.com;*

1 Introduction

The discrimination, or information divergence, between two random variables (r.v.-s), say X and Y, refers, in principle, to an information-theoretic method that measures the increase, or decrease, of the amount information regarding an experiment. The term information distance is also used. However, the term "distance" it is not mathematically accurate, as the information divergence is often not a distance metric. In the context of Information Geometry, the search for information divergences that are also distance metrics on a statistical manifold is essential.

To measure the information distance between two distributions, the Kullback-Leibler (KL) divergence is one of the most commonly used measures, also known as *relative entropy*. The KL divergence serves as a simple quantification method of the amount of information "gained" when a given distribution (characterizing an I/O system) is substituted by another one. Recall that, for the discrete probability distributions P and Q, the KL divergence from Q to P is defined to be

$$D_{\mathrm{KL}}(P||Q) := \sum_{i=1}^{n} P_i \log \frac{P_i}{Q_i}, \tag{1.1}$$

which is the expectation of the logarithmic differences between probability values $P_i := P(i)$ and $Q_i := Q(i)$, $i = 1,2,\ldots,n$. Note that probability values such that $Q_i = 0$ are considered only if $P_i = 0$, $i \in \{1,2,\ldots,n\}$. In general, for probability measures P and Q on the same space, with p and q denoting probability densities (with respect to a common probability measure μ on their space), the KL information divergence is defined as, [1, 2],

$$D_{\mathrm{KL}}(P\|Q) := \int p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}\mu(x). \tag{1.2}$$

In principle $D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P)$. In this paper we shall adopt continuous probability distributions for the KL divergence as well as for all the other information measures/divergences.

Here is a well known example: The "distance", or better to say "how far is" the standard normal distribution $P \sim \mathcal{N}(0,1) \equiv P$ from the Laplace distribution $\mathscr{L}(0,1) \equiv Q$, is given by

$$D_{\mathrm{KL}}(P\|Q) = \tfrac{1}{2}\log\tfrac{2}{\pi} - \tfrac{1}{2} + \sqrt{\tfrac{1}{2}} = 0.07209 \ \text{ and } \ D_{\mathrm{KL}}(Q\|P) = 0.22579, \tag{1.3}$$

considering that

$$p(x) := (2\pi)^{-1/2} \mathrm{e}^{-x^2/2}, \quad x \in \mathbb{R}, \quad q(x) := \tfrac{1}{2}\mathrm{e}^{-|x|}, \quad x \in \mathbb{R}, \ \text{with}$$

$$\mathrm{E}_P(|X|) = \sqrt{2/\pi}, \ \text{ and } \ \mathrm{E}_Q(|X|) = 1,$$

with $\mathrm{E}_P(X^2) = 1$ and $\mathrm{E}_Q(X^2) = 2$.

Information distances are often used in practice as in Cryptography. Typical example being the set of 2^n binary words of length n, say W_n, with the distance $D(a,b)$ between words a and b, defined to be the number of bits in which word a differs from word b. As an example, $D(1010101, 1100100) = 3$. In such cases we are interested in evaluating $D_{\min}$, which is the minimum information distance between two distinct codewords in a given set. The above discussion shows that there is a reason for the experimenter to know how "close" can be

(from the informational point of view) two given distributions that correspond to some to Input/Output system.

If it is to investigate how close two distributions are, from the classical statistical point of view, the known Kolmogorov-Smirnov test can be simply applied. Recall that, as far as the estimation problems are concerned, the minimum distance property is also a statistical property, embedded in the notions of likelihood, regression, χ^2 etc. Therefore, adopting a measure of distance for given probability measures P and Q, it would be of interest to calculate the minimum distance. Moreover, [3] defined a conceptual distance metric which applied to bioassays, while in [4] a mini-max distance method was introduced, based on an entropy measure.

In this present work, the idea of Relative Risk (RR), fundamental to Logit methods, see [5], shall be transferred in a more information-theoretic framework: The well-known odds ratio, defined to measure a dichotomous exposure-outcome scenario, is eventually evaluated as the exponential of the logistic regression coefficient (to the given data), say b. This exponentiated function $\mathrm{e}^d - 1$ is a distance, provided d is a distance, see [6], and remains invariant under linear transformation; see [7, 8]. Therefore, evaluating the Hellinger distance for two γ-order Generalized Normal (γ-GN) distributions, the exponentiation (as above) of this distance, is also an informational distance metric, acts as a Relative Risk to what γ-GN we are closer.

Recall the γ-GN family of distributions, consisted of a three parameters exponential-power generalized form of the usual multivariate Normal distribution defined below; see [9, 10] for details.

Definition 1.1. The p-variate random variable X follows the γ-order Generalized Normal distribution, i.e. $X \sim \mathcal{N}_\gamma^p(\mu, \Sigma)$, with location parameter vector $\mu \in \mathbb{R}^p$, shape parameter $\gamma \in \mathbb{R} \setminus [0, 1]$ and positive definite scale parameter matrix $\Sigma \in \mathbb{R}^{p\times p}$, when the density function f_X of X is of the form

$$f_X(x) = f_X(x;\, \mu, \Sigma, \gamma, p) := C_X \exp\left\{-\tfrac{\gamma-1}{\gamma} Q_\theta(x)^{\frac{\gamma}{2(\gamma-1)}}\right\}, \quad x \in \mathbb{R}^p, \tag{1.4}$$

where Q_θ denotes the p-quadratic form $Q_\theta(x) := (x-\mu)\Sigma^{-1}(x-\mu)^{\mathrm{T}}$, $x \in \mathbb{R}^p$, $\theta := (\mu, \Sigma)$. with the normalizing factor C_X is defined as

$$C_X = C_X(\Sigma, \gamma, p) := \max f_X = \frac{\left(\frac{p}{2}+1\right)}{\pi^{p/2}\left(p\frac{\gamma-1}{\gamma}+1\right)\sqrt{|\Sigma|}} \left(\tfrac{\gamma-1}{\gamma}\right)^{p\frac{\gamma-1}{\gamma}}, \tag{1.5}$$

where $|\mathrm{A}| := \det \mathrm{A}$ denotes the determinant of any $\mathrm{A} \in \mathbb{R}^{p\times p}$.

The p.d.f. f_X as above shall be adopted for the probability densities p and q discussed in (1.2). Notice that the location parameter vector μ of X is essentially the mean vector of X, i.e. $\mu = \mu_X := \mathrm{E}(X)$. Moreover, for the shape parameter value $\gamma = 2$, $\mathcal{N}_2^p(\mu, \Sigma)$ is reduced to the well known multivariate normal distribution, where Σ is now the covariance of X, i.e. $\mathrm{Cov}\, X = \Sigma$.

The family of $\mathcal{N}_\gamma^p(\mu, \Sigma)$ distributions, i.e. the family of the elliptically contoured γ-order Generalized Normals, provides a smooth bridging between some important multivariate (and elliptically countered) distributions. Indeed, [11]:

Theorem 1.1. *For the elliptically contoured p-variate γ-order Normal distribution $\mathcal{N}_\gamma^p(\mu,\Sigma)$ with $\mu\in\mathbb{R}^p$ and $\Sigma\in\mathbb{R}^{p\times p}$, we obtain the following special cases:*

$\gamma := 0$. *For the limiting case of the shape parameter $\gamma\to 0^-$, the degenerate Dirac distribution $\mathcal{D}(\mu)$ with pole at μ is derived in dimensions $p := 1,2$, while for $p\geq 3$ the p.d.f. of $\mathcal{N}_0(\mu,\Sigma)$ is flattened (p.d.f. is everywhere zero).*

$\gamma := 1$. *For the limiting case of $\gamma\to 1^+$ the elliptically contoured Uniform distribution $\mathcal{U}^p(\mu,\Sigma)$ is obtained, which is defined over the p-ellipsoid $Q_\theta(x) = (x-\mu)\Sigma^{-1}(x-\mu)^{\mathrm{T}}\leq 1$, $x\in\mathbb{R}^p$.*

$\gamma := 2$. *For the "normality" case of $\gamma := 2$ the usual p-variate Normal distribution $\mathcal{N}^p(\mu,\Sigma)$ is obtained.*

$\gamma := \pm\infty$. *For the limiting case of $\gamma\to\pm\infty$ the elliptically contoured Laplace distribution $\mathcal{L}^p(\mu,\Sigma)$ is derived.*

One of the merits of the $\mathcal{N}_\gamma$ family of distributions is that it can provide "heavy-tailed" distributions as the shape parameter γ influences the "probability mass" at the tails; see [11, 10, 12].

The corresponding cumulative distribution function (c.d.f.) of the γ-GN, as in (1.4), is expressed as [12],

$$F_X(x) = 1 - \frac{1}{2\left(2\frac{\gamma-1}{\gamma}\right)}\left(\frac{\gamma-1}{\gamma},\frac{\gamma-1}{\gamma}\left(\frac{x-\mu}{\sigma}\right)^{\gamma/(\gamma-1)}\right),\quad x\in\mathbb{R}, \tag{1.6}$$

where $(\cdot,\cdot)$ denotes the upper (or complementary) incomplete gamma function. Alternatively, using positive arguments for the upper (complementary) incomplete gamma function (a,x), $x\in\mathbb{R}$, $a\in\mathbb{R}_+$ (which is more computationally oriented approach), it holds that

$$F_X(x) = \frac{1+\operatorname{sgn} z}{2} - (\operatorname{sgn} z)\frac{\left(g, g|z|^{1/g}\right)}{2(g)},\quad z = z(x;\,\mu,\sigma) := \frac{x-\mu}{\sigma},\quad x\in\mathbb{R}. \tag{1.7}$$

In such a case, the quantile function is then given by

$$Q_X(P) := \inf\left\{x\in\mathbb{R}:\ F_X(x)\geq P\right\} = \operatorname{sgn}(2P-1)\sigma\left[\tfrac{1}{g}^{-1}\left(g,|2P-1|\right)\right]^g,\quad P\in(0,1). \tag{1.8}$$

2 KL Divergence and the γ-GND

In this section the information divergence, between two γ-GN distributions of the same order and mean is obtained through the KL measure of information divergence. Recall that the KL divergence $D_{\mathrm{KL}}(X\|Y)$ from a r.v. Y to another r.v. X (of the same dimension) can be defined, through (1.2), as

$$D_{\mathrm{KL}}(X\|Y) := D_{\mathrm{KL}}(F_X\|F_Y), \tag{2.1}$$

where F_X and F_Y being the cumulative distribution functions of r.v.-s X and Y respectively. The KL divergence, which is the most frequently used information distance measure in practice, is not a genuine distance metric, as it is not non-negative in general, violating the "positive definiteness" or, alternatively, the "identity of indescernibles" property (i.e. is a distance *pseudo-metric*). Moreover, D_{KL} does not obeys the "subadditivity" property, also known as the triangle

inequality (i.e. is a distance *semi-metric*), while also violates symmetricity (i.e. is a distance *quasi-metric*). That is why we shall refer to it as a "divergence" rather than a "distance" measure. Note also that the KL divergence can be adopted to evaluate the transfer entropy between stationary processes whose continuous probability distributions are known. For the γ-GN case see [13].

Specifically, for two spherically contoured γ-order normally distributed r.v.-s with the same mean and shape, i.e. $X \in \mathcal{N}_\gamma\big(\mu_1, \sigma_1^2 \mathbb{I}_p\big)$, $Y \sim \mathcal{N}_\gamma\big(\mu_2, \sigma_2^2 \mathbb{I}_p\big)$, with $\mu := \mu_1 = \mu_2 \in \mathbb{R}^p$, the KL divergence of X over Y is given by, [14],

$$D_{\mathrm{KL}}(X\|Y) = p \log \frac{\sigma_2}{\sigma_1} - p\big(\tfrac{\gamma-1}{\gamma}\big)\Big[1 - \big(\tfrac{\sigma_1}{\sigma_2}\big)^{\frac{\gamma}{\gamma-1}}\Big], \tag{2.2}$$

while for $\mu_1 \neq \mu_2$ and $\gamma = 2$, it holds

$$D_{\mathrm{KL}}(X\|Y) = \frac{p}{2}\left[2\log\frac{\sigma_2}{\sigma_1} - 1 + \big(\frac{\sigma_1}{\sigma_2}\big)^2 + \frac{\|\mu_2 - \mu_1\|^2}{p\sigma_2^2}\right], \tag{2.3}$$

which is the usual KL divergence between two normally distributed r.v.-s. For the KL divergence of the γ-GN distribution over the Student's t-distribution see [15].

Remark 2.1. It is worth mentioning that the expression (2.2) is always positive definite as well as proportional to the dimension of X and Y for every $\sigma_1, \sigma_2 \in \mathbb{R}^+$ and $\gamma \in \mathbb{R} \setminus [0,1]$. Indeed, $D_{\mathrm{KL}}(X\|Y) \propto p \in \mathbb{N}^* := \mathbb{N} \setminus \{0\}$. Moreover, writing (2.2) as $D(s) := D_{\mathrm{KL}}(X\|Y) = p\log s - pg\big(1 - s^{-1/g}\big)$ where $s := \sigma_2/\sigma_1$ and $g := (\gamma-1)/\gamma$, the relation $D'(s) := \frac{\mathrm{d}}{\mathrm{d}s}D(s) = s^{-1} - s^{-(g+1)/g} = 0$ yields $s^{-1/g} = 1$, and hence $s = 1$ while $D(1) = 0$. It is then easy to see that $D'(s) > 0$ for $s > 1$, and $D'(s) < 0$ for $s < 1$. Therefore, (2.2) admits always a (global) minima when $\sigma_1 := \sigma_2$ or equivalently when $X = Y$. As a result, the global minima of D at 1 and the fact that $D(1) = 0$ implies the positiveness of $D_{\mathrm{KL}}(X\|Y)$, while the fact that $\sigma = 1$ is also a global minima implies that $X = Y$ when $D(X\|Y) = 0$ is assumed, i.e. the KL divergence between two spherically contoured γ-GN distributions of the same mean obeys positive definiteness. Moreover, (2.2) is also a non-bounded information divergence. In particular, concerning the limiting behavior of (2.2) with respect to the scale parameters σ_i, $i = 1,2$, it holds that $\lim_{\sigma_i \to +\infty} D_{\mathrm{KL}}(X\|Y) = +\infty$, $i = 1,2$. Indeed, it holds that $\lim_{\sigma_2 \to +\infty} D_{\mathrm{KL}}(X\|Y) = -pg + p\lim_{\sigma_2 \to +\infty} \log(\sigma_2/\sigma_1) = +\infty$. On the other hand, $\lim_{\sigma_1 \to +\infty} D_{\mathrm{KL}}(X\|Y) = -pg + p\lim_{s \to 0^+}\big(\log s + s^{-1/g}\big) = \lim_{s\to 0^+} s^{-1/g}\big(1 + s^{1/g}\log s\big) = \lim_{s\to 0^+} s^{-1/g}\big(1 - gs^{1/g}\big) = \lim_{s\to 0^+} s^{-1/g} = +\infty$, since $g \in \mathbb{R}^+$.

One can also notice that the relation (2.2) implies, for appropriate choices of the γ value (recall Theorem 1.1), that the KL divergence between two uniformly distributed r.v.-s $U_i \in \mathcal{U}^p\big(\mu, \sigma_i^2 \mathbb{I}_p\big)$, $i = 1,2$, is given by,

$$D_{\mathrm{KL}}\big(U_1\|U_2\big) = \lim_{\gamma \to 1^+} D_{\mathrm{KL}}(X\|Y) = \begin{cases} p\log\dfrac{\sigma_2}{\sigma_1}, & \sigma_1 \leq \sigma_2, \\ +\infty, & \sigma_1 > \sigma_2, \end{cases} \tag{2.4}$$

while the KL divergence between two Laplace distributed r.v.-s $L_i \in \mathcal{L}^p\big(\mu, \sigma_i^2 \mathbb{I}_p\big)$, $i = 1,2$, is given by

$$D_{\mathrm{KL}}\big(L_1\|L_2\big) = \lim_{\gamma \to +\pm\infty} D_{\mathrm{KL}}(X\|Y) = p\left(\log\frac{\sigma_2}{\sigma_1} - 1 + \frac{\sigma_1}{\sigma_2}\right). \tag{2.5}$$

From (2.2), it is easy to see that

$$D_{\mathrm{KL}}^{p}(\gamma) < D_{\mathrm{KL}}^{p+1}(\gamma), \quad p = 1,2,\ldots, \tag{2.6}$$

where $D_{\mathrm{KL}}^{p}(\gamma) := D_{\mathrm{KL}}(X\|Y)$, $X \sim \mathcal{N}_{\gamma}^{p}\left(\mu, \sigma_0^2 \mathbb{I}_p\right)$ and $Y \sim \mathcal{N}_{\gamma}^{p}\left(\mu, \sigma_1^2 \mathbb{I}_p\right)$. Similar inequalities hold also in case of Laplace probability function for $\gamma \to +\infty$. Note also that for given p, $\mu_1 = \mu_2$, and $\sigma_1 \neq \sigma_2$ the KL divergence in (2.2) appears in a strict descending order as $\gamma \in \mathbb{R} \setminus [0\,1]$ rises. In particular,

$$D_{\mathrm{KL}}^{p}(\gamma_1) > D_{\mathrm{KL}}^{p}(\gamma_2), \text{ for } \gamma_1 < \gamma_2. \tag{2.7}$$

Therefore with Laplace, $\gamma \to +\infty$, we obtain a lower bound, i.e.

$$D_{\mathrm{KL}}^{p}(\infty) < D_{\mathrm{KL}}^{p}(\gamma), \text{ for every } \gamma \text{ and } p \in \mathbb{N}. \tag{2.8}$$

3 Families of Information Divergences and Symmetry

Recall that the Fisher's entropy type information measure $I_{\mathrm{F}}(X)$ of an r.v. X with p.d.f. f on $\mathbb{R}^p$, is defined as the covariance of r.v. $\nabla \log f(X)$, i.e. $I_{\mathrm{F}}(X) := \mathrm{E}[\|\nabla \log f(X)\|^2]$, with $\mathrm{E}[\cdot]$ now denotes the usual expected value operator of a random variable with respect to the its p.d.f. Hence, $I_{\mathrm{F}}(X)$ can be written as

$$I_{\mathrm{F}}(X) = \int_{\mathbb{R}^p} f(x)\|\nabla \log f(x)\|^2 \,\mathrm{d}x = \int_{\mathbb{R}^p} f(x)^{-1}\|\nabla f(x)\|^2 \,\mathrm{d}x = \int_{\mathbb{R}^p} \nabla f(x)\cdot\nabla \log f(x)\,\mathrm{d}x = 4\int_{\mathbb{R}^p} \left\|\nabla \sqrt{f(x)}\right\|^2 \mathrm{d}x. \tag{3.1}$$

The Fisher's entropy type information of an r.v. X is a special case of information measures defined by the general form

$$I(X) := I(X; g, h) := g\Big(\mathrm{E}\left[h\big(U(X)\big)\right]\Big), \tag{3.2}$$

where g and h being real-valued functions and U being the score function, i.e. $U(X) := \left\|\nabla \log f(X)\right\|$. Indeed, letting $g :=$ i.d. and $h(X) := X^2$ we obtain the entropy type Fisher's information measure of X as in (3.1), i.e.

$$I_{\mathrm{F}}(X) = \mathrm{E}\left[\left\|\nabla \log f(X)\right\|^2\right]. \tag{3.3}$$

Other entropy type information measures, such as the Vajda's, Mathai's and Boeke's information measures, denoted with I_{V}, I_{M} and I_{B} respectively, are defined as follows:

$$I_{\mathrm{V}}(X) := I(X), \text{ with } \quad g := \mathrm{id}. \quad \text{and } h(u) := u^{\alpha}, \ \alpha \geq 1, \tag{3.4a}$$

$$I_{\mathrm{M}}(X) := I(X), \text{ with } g(x) := x^{1/\alpha} \text{ and } h(u) := u^{\alpha}, \ \alpha \geq 1, \tag{3.4b}$$

$$I_{\mathrm{B}}(X) := I(X), \text{ with } g(x) := x^{\alpha-1} \text{ and } h(u) := u^{\frac{\alpha}{\alpha-1}}, \ \alpha \in \mathbb{R}^{+} \setminus 1. \tag{3.4c}$$

For a generalisation of the Fisher's entropy type information measure see [16].

We define here a general formulation for information divergences of a p-variate r.v. X over a p-variate r.v. Y, which is given by

$$D_{\mathrm{KT}}(X\|Y) := g\left(\int_{\mathbb{R}^p} h(f_X, f_Y)\right), \tag{3.5}$$

where f_X and f_Y are the p.d.f.-s of X and Y respectively. We are then reduced to a series of known divergencies, [17, 18, 19, 20, 21], such as the Kullback-Leibler D_{KL}, exponential D_{e}, Vajda's D_{V}, Kagan (or χ^2) D_{χ^2}, Csiszár D_{C}, Rényi D_{R}, Tsallis D_{T}, Amari D_{A}, Chernoff α-divergence (of the first type) $D^{(\alpha)}_{\mathrm{Ch}}$, Chernoff α-divergence (of the second type) $D'^{(\alpha)}_{\mathrm{Ch}}$, and the Chernoff D_{Ch} divergence, as well as the distances, such as the Hellinger D_{H}, Bhattacharyya D_{B}, and total variation δ:

$$D_{\mathrm{KL}}(X\|Y):\ \text{With}\quad g:=\mathrm{id.}\quad \text{and}\quad h(f_X,f_Y):=f_X\log\frac{f_X}{f_Y}, \tag{3.6a}$$

$$D_{\mathrm{e}}(X\|Y):\ \text{With}\quad g:=\mathrm{id.}\quad \text{and}\quad h(f_X,f_Y):=f_X\left(\log\frac{f_X}{f_Y}\right)^2, \tag{3.6b}$$

$$D_{\mathrm{V}}(X\|Y):\ \text{With}\quad g:=\mathrm{id.}\quad \text{and}\quad h(f_X,f_Y):=f_X\left|1-\frac{f_Y}{f_X}\right|^\alpha,\ \alpha\geq 1, \tag{3.6c}$$

$$D_{\chi^2}(X\|Y):\ \text{With}\quad g:=\mathrm{id.}\quad \text{and}\quad h(f_X,f_Y):=\tfrac{1}{2}f_X\left(1-\frac{f_Y}{f_X}\right)^2, \tag{3.6d}$$

$$D_{\mathrm{C}}(X\|Y):\ \text{With}\quad g:=\mathrm{id.}\quad \text{and}\quad h(f_X,f_Y):=f_Y\phi\Big(\frac{f_X}{f_Y}\Big),\ \begin{matrix}\phi\ \text{convex},\\ \phi(1):=0,\end{matrix} \tag{3.6e}$$

$$D_{\mathrm{R}}(X\|Y):\ \text{With}\quad g:=\tfrac{1}{\alpha-1}\log\quad \text{and}\quad h(f_X,f_Y):=f_X^\alpha f_Y^{1-\alpha},\ 0<\alpha\neq 1, \tag{3.6f}$$

$$D_{\mathrm{T}}(X\|Y):\ \text{With}\quad g(u):=\tfrac{1-u}{1-\alpha}\quad \text{and}\quad h(f_X,f_Y):=f_X^{1-\alpha}f_Y^\alpha,\ 0<\alpha\neq 1, \tag{3.6g}$$

$$D_{\mathrm{A}}(X\|Y):\ \text{With}\quad g(u):=\tfrac{4(1-u)}{1-\alpha^2}\quad \text{and}\quad h(f_X,f_Y):=f_X^{(1-\alpha)/2}f_Y^{(1+\alpha)/2},\ \alpha\neq\pm 1, \tag{3.6h}$$

$$D^{(\alpha)}_{\mathrm{Ch}}(X\|Y):\ \text{With}\quad g:=\mathrm{id.}\quad \text{and}\quad h(f_X,f_Y):=f_X^\alpha f_Y^{1-\alpha},\ \alpha\in(0,1), \tag{3.6i}$$

$$D'^{(\alpha)}_{\mathrm{Ch}}(X\|Y):\ \text{With}\quad g(u):=\tfrac{1-u}{\alpha(1-\alpha)}\quad \text{and}\quad h(f_X,f_Y):=f_X^\alpha f_Y^{1-\alpha},\ \alpha\in(0,1), \tag{3.6j}$$

$$D_{\mathrm{Ch}}(X\|Y):\ \text{With}\quad g:=-\log\min_{\alpha\in(0,1)}\quad \text{and}\quad h(f_X,f_Y):=f_X^\alpha f_Y^{1-\alpha},\ \alpha\in(0,1), \tag{3.6k}$$

$$D_{\mathrm{H}}(X,Y):\ \text{With}\quad g(u):=\sqrt{u}\quad \text{and}\quad h(f_X,f_Y):=\tfrac{1}{2}\left(\sqrt{f_X}-\sqrt{f_Y}\right)^2, \tag{3.6l}$$

$$D_{\mathrm{B}}(X,Y):\ \text{With}\quad g:=-\log\quad \text{and}\quad h(f_X,f_Y):=\sqrt{f_Xf_Y}, \tag{3.6m}$$

$$\delta(X,Y):\ \text{With}\quad g:=\mathrm{id.}\quad \text{and}\quad h(f_X,f_Y):=\left|f_X-f_Y\right|. \tag{3.6n}$$

Alternatively, the above divergencies belong also to the wider Ali-Silvey class of information divergencies [18], expressed in the form

$$D_{\mathrm{AS}}(X\|Y):=g\Big(\mathrm{E}\left[\eta\big(\ell(Y)\big)\right]\Big), \tag{3.7}$$

where g is a non-decreasing function, η is convex, $\mathrm{E}[\cdot]$ is the expected value operator (with respect to the p.d.f. of Y), and ℓ denotes the likelihood ratio f_X/f_Y. For example, KL is an Ali-Silvey divergence as in (3.7) with $g:=\mathrm{i.d.}$ and $\eta(x):=x\log x$, $x\in\mathbb{R}^{*+}$. Chernoff α-divergence (of the first type) $D^{(\alpha)}_{\mathrm{Ch}}$ is also an Ali-Silvey divergence with $g:=\mathrm{i.d.}$ and $\eta(x):=-x^\alpha$, $x\in\mathbb{R}$. For the Chernoff divergence D_{Ch} we must adopt as $g(x):=-\log(-x)$, $x\in\mathbb{R}^{*-}$.

Note that Tsallis and Rényi divergences are related through the mappings

$$D_{\mathrm{T}}(X\|Y)=\tfrac{1}{1-\alpha}\left[1-\mathrm{e}^{(1-\alpha)D_{\mathrm{R}}(Y\|X)}\right],\ \text{ or }\ D_{\mathrm{R}}(X\|Y)=\tfrac{1}{\alpha-1}\log\big(1-D_{\mathrm{T}}(Y\|X)\big). \tag{3.8}$$

Regarding Bhattacharyya distance, Hellinger distance, and Chernoff α-divergence, between two multivariate normally distributed r.v.-s, the known computations are given in the following example:

Example 3.1. *Let $X \sim \mathcal{N}^p(\mu_X, \Sigma_X)$ and $Y \sim \mathcal{N}^p(\mu_Y, \Sigma_Y)$ with means $\mu_X, \mu_Y \in \mathbb{R}^p$, $\mu_X \neq \mu_Y$, and covariances $\Sigma_X, \Sigma_Y \in \mathbb{R}^{p\times p}$, $\Sigma_X \neq \Sigma_Y$. Then*

$$D_{\mathrm{B}}(X,Y) = \tfrac{1}{2}\log\frac{|\Sigma_X|}{\sqrt{|\Sigma_X||\Sigma_Y|}} + \tfrac{1}{8}(\mu_X-\mu_Y)\Sigma^{-1}(\mu_X-\mu_Y)^{\mathrm{T}}, \quad \Sigma := \tfrac{1}{2}(\Sigma_X+\Sigma_Y), \tag{3.9a}$$

$$D^2_{\mathrm{H}}(X,Y) = 1 - \frac{\sqrt[4]{|\Sigma_X|}\sqrt[4]{|\Sigma_Y|}}{\sqrt{|\Sigma|}}\exp\left\{\tfrac{1}{8}(\mu_X-\mu_Y)\Sigma^{-1}(\mu_X-\mu_Y)^{\mathrm{T}}\right\}, \tag{3.9b}$$

$$D^{(\alpha)}_{\mathrm{Ch}}(X\|Y) = \tfrac{1}{2}\log\frac{|\alpha\Sigma_X+(1-\alpha)\Sigma_Y|}{|\Sigma_X|^{\alpha}|\Sigma_Y|^{1-\alpha}}. \tag{3.9c}$$

For the univariate case of $X \sim \mathcal{N}\left(\mu_X, \sigma_X^2\right)$ and $Y \sim \mathcal{N}\left(\mu_Y, \sigma_Y^2\right)$ it holds that

$$D_{\mathrm{B}}(X,Y) = \tfrac{1}{4}\log\left\{\tfrac{1}{4}\left(\frac{\sigma_X^2}{\sigma_Y^2}+\frac{\sigma_Y^2}{\sigma_X^2}+2\right)\right\} + \frac{(\mu_X-\mu_Y)^2}{4\left(\sigma_X^2+\sigma_Y^2\right)}, \tag{3.10a}$$

$$D^2_{\mathrm{H}}(X,Y) = 1 - \sqrt{\frac{2\sigma_X\sigma_Y}{\sigma_X^2+\sigma_Y^2}}\exp\left\{-\frac{(\mu_X-\mu_Y)^2}{4\left(\sigma_X^2+\sigma_Y^2\right)}\right\}, \tag{3.10b}$$

$$D^{(\alpha)}_{\mathrm{Ch}}(X\|Y) = \tfrac{1}{2}\log\frac{\alpha\sigma_X^2+(1-\alpha)\sigma_Y^2}{\sigma_X^{2\alpha}\sigma_Y^{2(1-\alpha)}}. \tag{3.10c}$$

The KL divergence belongs also to a much wider class of divergences called *Bregman* divergence, [17], defined by

$$D_{\mathrm{Br}}(X\|Y) := \int \phi(f_X) - \phi(f_Y) - (f_X - f_Y)\phi'(f_X), \tag{3.11}$$

for a strictly convex and differentiable generator ϕ, where f_X and f_Y are the p.d.f.-s of r.v.-s X and Y. Indeed, for $\phi(x) := x\log x$, $x \in \mathbb{R}^{*+} := \mathbb{R}^+ \setminus \{0\}$, Bregman divergence is reduced to the KL divergence. Note that the Amari divergence D_{A} generalizes also the KL distance in the sense that $D_{\mathrm{A}}(X\|Y) = D_{\mathrm{KL}}(X\|Y)$ for the limiting case of $\alpha := -1$, while $D_{\mathrm{A}}(X\|Y) = D_{\mathrm{KL}}(Y\|X)$ for $\alpha := 1$; see [17]. Moreover, KL divergence belongs, in limit, to the Rényi and Tsallis divergence classes for $\alpha \to 1$. We try to summarize and clarify the existent relations between the different divergence measures, in the analytical form, since in technical/engineering interpretation have a different meaning and use. Given the one we can evaluate a series of other divergence measures.

4 Information Relative Risk

Consider the (upper) bounded Hellinger distance, $0 \leq D_{\mathrm{H}} \leq 1$, which obeys the triangle inequality. For the related Bhattacharyya distance, $D_{\mathrm{B}} = -\log\left(1 - D^2_{\mathrm{H}}\right)$, the triangle inequality does not hold, although both measures are usually referred as "distances" since both are symmetric. Therefore, like the total variation distance (which is usually referred as the statistical distance),

the Hellinger distance can be considered as a *statistical distance metric*, since it is essential a distance metric with the usual mathematical meaning, while the Bhattacharyya distance is considered to be a (statistical) *distance semi-metric*, as it is a distance metric not obeying the triangle inequality. Recall also the inequalities between the total variation and the Hellinger distance, i.e.

$$D_{\mathrm{H}}^2 \leq \delta \leq \sqrt{2} D_{\mathrm{H}}, \tag{4.1}$$

which follow clearly from the inequalities between the 1-norm and the 2-norm. Moreover, the Hellinger distance can be alternatively defined with the help of other divergences, through the mappings

$$D_{\mathrm{H}} = \mathrm{e}^{\frac{1}{4} D_{\mathrm{R}}(1/2)} = \sqrt{1 - \tfrac{1}{2} D_{\mathrm{T}}(1/2)} = \tfrac{1}{2}\sqrt{4 - D_{\mathrm{A}}(0)} = \sqrt{\mathrm{e}^{D_{\mathrm{B}}} - 1}, \tag{4.2}$$

where the parameter values 1/2, 1/2 and 0, in D_{R}, D_{T} and D_{A} respectively, corresponds to the parameter α of each divergence, e.g. $D_{\mathrm{R}}(1/2) := D_{\mathrm{R}}(X\|Y)\big|_{\alpha:=1/2}$. Alternatively,

$$D_{\mathrm{R}}(1/2) = 4\log D_{\mathrm{H}},\ D\mathrm{T}(1/2) = 2\big(1 - D_{\mathrm{H}}^2\big),\ D_{\mathrm{A}}(0) = 4\big(1 - D_{\mathrm{H}}^2\big),\ D_{\mathrm{B}} = \log\big(D_{\mathrm{H}}^2 - 1\big), \tag{4.3}$$

and thus, Rényi, Tsallis and Amari divergencies can be considered as generalized forms of the Hellinger distance.

Now, for two spherically contoured γ-GN r.v.-s with the same mean and shape parameter γ, their corresponding Hellinger distance is given in the following:

Proposition 4.1. *Let $X \sim \mathcal{N}_\gamma^p\big(\mu, \sigma_1^2 \mathbb{I}_p\big)$ and $Y \sim \mathcal{N}_\gamma^p\big(\mu, \sigma_2^2 \mathbb{I}_p\big)$, with mean vector $\mu \in \mathbb{R}^p$ and $\sigma_1, \sigma_2 \in \mathbb{R}^+$. The Hellinger distance between X and Y is then given by*

$$D_{\mathrm{H}}(X,Y) = \sqrt{1 - \frac{2^{p\frac{\gamma-1}{\gamma}} s^{p/2}}{\Big(1 + s^{\frac{\gamma}{\gamma-1}}\Big)^{p\frac{\gamma-1}{\gamma}}}}, \quad s := \frac{\sigma_2}{\sigma_1}. \tag{4.4}$$

Proof. Consider the p.d.f. (1.4) for r.v.-s X and Y. Substituting them into (3.6l) we obtain consecutively

$$\begin{aligned} D_{\mathrm{H}}^2(X,Y) &= 1 - \int \sqrt{f_X f_Y} = 1 - \sqrt{C_X C_Y} \int_{\mathbb{R}^p} \exp\left\{ -\tfrac{g}{2} \left\| \frac{x-\mu}{\sigma_1} \right\|^{1/g} - \tfrac{g}{2} \left\| \frac{x-\mu}{\sigma_2} \right\|^{1/g} \right\} \mathrm{d}x \\ &= 1 - \sqrt{C_X C_Y} \int_{\mathbb{R}^p} \exp\Big\{ -\sigma g \|x-\mu\|^{1/g} \Big\} \mathrm{d}x = 1 - \sqrt{C_X C_Y} \int_{\mathbb{R}^p} \mathrm{e}^{-\sigma g \|x\|^{1/g}} \mathrm{d}x, \end{aligned}$$

where $g := (\gamma - 1)/\gamma$ and $\sigma := \frac{1}{2}\Big(\sigma_1^{-1/g} + \sigma_2^{-1/g}\Big)$. Switching to hyperspherical coordinates, it holds that

$$D_{\mathrm{H}}^2(X,Y) = 1 - \omega_{p-1}\sqrt{C_X C_Y} \int_{\mathbb{R}^+} \rho^{p-1} \mathrm{e}^{-\sigma g \rho^{1/g}} \mathrm{d}\rho,$$

where $\omega_{p-1} := 2\pi^{p/2}/(p/2)$ denotes the volume of the unitary $(p-1)$-sphere. Applying the variable transformation $\mathrm{d}u = \mathrm{d}u(\rho) := \mathrm{d}\big(\sigma g \rho^{1/g}\big) = \sigma \rho^{(1-g)/g} \mathrm{d}\rho$, the above can be written successively as

$$D_{\mathrm{H}}^2(X,Y) = 1 - \sigma^{-1}\omega_{p-1}\sqrt{C_X C_Y} \int_{\mathbb{R}^+} \rho^{\frac{(p-1)g+g-1}{g}} \mathrm{e}^{-u} \mathrm{d}u$$

$$= 1-\sigma^{-pg}g^{1-pg}\omega_{p-1}\sqrt{C_XC_Y}\int_{\mathbb{R}^+}\left(\sigma g\rho^{1/g}\right)^{pg-1}\mathrm{e}^{-u}\,\mathrm{d}u$$
$$= 1-\sigma^{-pg}g^{1-pg}\omega_{p-1}\sqrt{C_XC_Y}\int_{\mathbb{R}^+}u^{pg-1}\mathrm{e}^{-u}\,\mathrm{d}u$$
$$= 1-\sigma^{-pg}g^{1-pg}\omega_{p-1}(pg-1)\sqrt{C_XC_Y}.$$

By substitution of ω_{p-1} and the normalizing factors C_X and C_Y, as in (1.5), we get

$$D^2_{\mathrm{H}}(X,Y) = 1-pg\sigma^{-pg}\frac{(pg)}{(pg+1)}(\sigma_1\sigma_2)^{-p/2}. \tag{4.5}$$

Utilizing the additivity of the gamma function, i.e. $(x+1)=x(x)$, $x\in\mathbb{R}$, (4.5) can be written as

$$D^2_{\mathrm{H}}(X,Y) = 1-\sigma^{-pg}(\sigma_1\sigma_2)^{-p/2}, \tag{4.6}$$

and by substitution of σ and g into (4.6), the expression (4.4) is finally derived.

□

It is easy to see that $D_{\mathrm{H}}(X,Y)\to 1$ as $\sigma_i\to+\infty$, $i=1,2$, as, i.e. the Hellinger distance approaches its (upper) bound as the scale parameters of either X or Y are getting larger. The lower bound 0 is achieved only when $X=Y$, or $\sigma_1=\sigma_2$.

Corollary 4.1. *The Hellinger distance between two spherically contoured normally, or Laplace, or uniformly distributed r.v.-s of the same mean, are given by*

$$D_{\mathrm{H}}(X_1,X_2) = \sqrt{1-\left(\frac{2s}{s^2+1}\right)^{p/2}}, \quad X_i\sim\mathscr{N}^p\left(\mu,\sigma_i^2\mathbb{I}_P\right),\ i=1,2, \tag{4.7a}$$
$$D_{\mathrm{H}}(X_1,X_2) = \sqrt{1-\left(\frac{2\sqrt{s}}{s+1}\right)^{p}}, \quad X_i\sim\mathscr{L}^p\left(\mu,\sigma_i^2\mathbb{I}_p\right),\ i=1,2, \tag{4.7b}$$
$$D_{\mathrm{H}}(X_1,X_2) = \sqrt{1-s^{\mathrm{sgn}(1-s)p/2}}, \quad X_i\sim\mathscr{U}^p\left(\mu,\sigma_i^2\mathbb{I}_p\right),\ i=1,2, \tag{4.7c}$$

respectively, where $s:=\sigma_2/\sigma_1$. *For the univariate cases, it holds*

$$D_{\mathrm{H}}(X_1,X_2) = \sqrt{1-\sqrt{\frac{2s}{s^2+1}}}, \quad X_i\sim\mathscr{N}\left(\mu,\sigma_i^2\right),\ i=1,2, \tag{4.8a}$$
$$D_{\mathrm{H}}(X_1,X_2) = \frac{\left|\sqrt{s}-1\right|}{\sqrt{s+1}}, \quad X_i\sim\mathscr{L}(\mu,\sigma_i),\ i=1,2, \tag{4.8b}$$
$$D_{\mathrm{H}}(X_1,X_2) = \sqrt{1-\sqrt{s^{\mathrm{sgn}(1-s)}}}, \quad X_i\sim\mathscr{U}(\mu-\sigma_i,\mu+\sigma_i),\ i=1,2. \tag{4.8c}$$

Proof. For the normality case as well as for the (limiting) Laplace case, the corresponding expressions (4.7a), (4.8a), and (4.7b), (4.8b), are derived straightforward from (4.4) with $\gamma:=2$ and with $\gamma\to\pm\infty$ respectively.

For the (limiting) uniformity case of $X_i\sim\mathscr{U}^p\left(\mu,\sigma_i^2\mathbb{I}_p\right)$, $i=1,2$, consider the γ-GN r.v.-s $X_\gamma\sim\mathscr{N}_\gamma^p\left(\mu,\sigma_1^2\mathbb{I}_p\right)$ and $Y_\gamma\sim\mathscr{N}_\gamma^p\left(\mu,\sigma_2^2\mathbb{I}_p\right)$. Recalling Theorem 1.1 and expression (4.4), it holds that

$$D^2_{\mathrm{H}}(X_1,X_2) := \lim_{\gamma\to1^+}D^2_{\mathrm{H}}(X_\gamma,Y_\gamma) = 1-s^{p/2}\lim_{g\to0^+}\left(1+s^{1/g}\right)^{-pg}, \quad g:=\tfrac{\gamma-1}{\gamma}. \tag{4.9}$$

The following three cases are then distinguished:

Case $\sigma_2 > \sigma_1$***, or*** $s > 1$**.** In such a case (4.9) can be written as

$$D^2_{\mathrm{H}}(X_1, X_2) = 1 - s^{p/2}\,\mathrm{e}^{-p\,\ell(s)}, \tag{4.10}$$

where

$$\ell(s) := \lim_{g\to 0^+} \frac{\log\left(1+s^{1/g}\right)}{1/g} = \lim_{g\to 0^+} \frac{s^{1/g}\left(-g^{-2}\right)\log s}{-g^{-2}\left(1+s^{1/g}\right)} = (\log s)\lim_{g\to 0^+}\frac{s^{1/g}}{1+s^{1/g}} = \log s,$$

and hence $D^2_{\mathrm{H}}(X_1, X_2) = 1 - s^{-p/2}$.

Case $\sigma_2 < \sigma_1$***, or*** $s < 1$**.** Then relation (4.9) yields that $D^2_{\mathrm{H}}(X_1, X_2) = 1 - s^{p/2}$.

Case $\sigma_2 = \sigma_1$***, or*** $s = 1$**.** Then relation (4.9) yields easily that $D^2_{\mathrm{H}}(X_1, X_2) = 1 - 1 = 0$ as it was expected, since in this case $X_1 = X_2$.

From the above discussion we conclude that (4.7c) holds.

Finally, the corresponding univariate expressions (4.8a) and (4.8b) are obtained straightforward from (4.8a) and (4.8b) respectively.

For the univariate case of $X_i \sim \mathscr{U}^p\left(\mu, \sigma_i^2 \mathbb{I}_p\right) = \mathscr{U}^1\left(\mu, \sigma_i^2\right)$, $i = 1,2$, recall that, in general, the (univariate) Uniform distribution $\mathscr{U}^1\left(\mu, \sigma^2\right)$, $\sigma \in \mathbb{R}^+$, can be written in its classical form $\mathscr{U}(a,b)$ with $a := \mu - \sigma$ and $b := \mu + \sigma$, and thus (4.8c) is obtained. □

Remark 4.1. Note that (4.7a) and (4.8a) are in accordance with (3.9b) and (3.10b) respectively, for $\mu_X := \mu_Y$. For the Uniform case, (4.8c) can be expressed alternatively, considering r.v.-s $X_i \sim \mathscr{U}(a_i, b_i)$, $i = 1,2$, where $a_1 + b_1 := a_2 + b_2$ (common mean assumption), in the form

$$D_{\mathrm{H}}(X_1, X_1) = \sqrt{1 - \sqrt{(\sigma_2/\sigma_1)^{\mathrm{sgn}(\sigma_2-\sigma_1)}}} = \sqrt{1 - \sqrt{\left(\frac{b_2-a_2}{b_1-a_1}\right)^{\mathrm{sgn}(b_1+a_2-b_2-a_1)}}}, \tag{4.11}$$

since $\mathscr{U}^1\left(\mu, \sigma^2\right) := \mathscr{N}_1^1\left(\mu, \sigma^2\right)$ is written in the Uniform's usual notation $\mathscr{U}(a,b)$, with $a := \mu - \sigma$ and $b := \mu + \sigma$, or alternatively, when $\mu := (a+b)/2$ and $\sigma := (b-a)/2$.

Remark 4.2. For the degenerate case of the zero-order Normal distributions, i.e. for $X_i \sim \mathscr{N}_0^p\left(\mu, \sigma_i^2 \mathbb{I}_p\right)$, $i = 1,2$, the Hellinger distance between X_1 and X_2, is then given, through (4.4), by

$$D^2_{\mathrm{H}}(X_1, X_2) = \lim_{\gamma\to 0^-} D^2_{\mathrm{H}}(Y_1, Y_2) = 1 - s^{p/2} \lim_{g\to+\infty}\left(\frac{2}{1+s^{1/g}}\right)^{pg} = 1 - s^{p/2}\,\mathrm{e}^{pL(s)}, \tag{4.12}$$

where $Y_i \sim \mathscr{N}_\gamma\left(\mu, \sigma_i^2 \mathbb{I}_p\right)$, $i = 1,2$, $s := \sigma_2/\sigma_1$, and

$$\begin{aligned} L(s) &:= \lim_{g\to+\infty} g \log\frac{2}{1+s^{1/g}} = \lim_{g\to+\infty}\frac{\log\left(2/\left(1+s^{1/g}\right)\right)}{1/g} = \lim_{g\to+\infty}\frac{\left(1+s^{1/g}\right)\frac{\mathrm{d}}{\mathrm{d}g}\left(1+s^{1/g}\right)^{-1}}{-g^{-2}} \\ &= -\lim_{g\to+\infty}\frac{s^{1/g}\left(-g^{-2}\right)\log s}{-g^{-2}\left(1+s^{1/g}\right)} = -\tfrac{1}{2}\log s. \end{aligned} \tag{4.13}$$

Therefore, by substitution of (4.13) into (4.12), we finally derive that $D^2_{\mathrm{H}}(X_1, X_2) = 1 - 1 = 0$, which is expected as $\mathscr{N}_0^p\left(\mu, \sigma_i^2 \mathbb{I}_p\right)$, $i = 1,2$, coincide both with the multivariate degenerate

Dirac distribution $\mathscr{D}^p(\mu)$ with pole at $\mu \in \mathbb{R}^p$ for $p = 1,2$, or are both flattened for $p \geq 3$; recall Theorem 1.1.

Note that we can also derive the Bhattacharyya distance between two γ-GN distributions of the same order and mean, with the following:

Corollary 4.2. *The Bhattacharyya distance between two spherically contoured γ-order normally distributed r.v.-s $X_i \sim \mathscr{N}_\gamma^p\left(\mu, \sigma_i^2 \mathbb{I}_p\right)$, $i = 1,2$, of the same order $\gamma \in \mathbb{R} \setminus [0,1]$ and mean $\mu \in \mathbb{R}^p$, is given by*

$$D_{\mathrm{B}}(X_1, X_2) = p\frac{\gamma-1}{\gamma}\log\frac{\sigma_1^{\frac{\gamma}{\gamma-1}} + \sigma_2^{\frac{\gamma}{\gamma-1}}}{2(\sigma_1\sigma_2)^{\frac{\gamma}{2(\gamma-1)}}}. \tag{4.14}$$

For the special cases of Normal, Laplace and Uniform distributions, it holds

$$D_{\mathrm{B}}(X_1, X_2) = \frac{p}{2}\log\frac{\sigma_1^2 + \sigma_2^2}{2\sigma_1\sigma_2}, \qquad X_i \sim \mathscr{N}^p\left(\mu, \sigma_i^2 \mathbb{I}_P\right),\ i = 1,2, \tag{4.15a}$$

$$D_{\mathrm{B}}(X_1, X_2) = p\log\frac{\sigma_1 + \sigma_2}{2\sqrt{\sigma_1\sigma_2}}, \qquad X_i \sim \mathscr{L}^p\left(\mu, \sigma_i^2 \mathbb{I}_p\right),\ i = 1,2, \tag{4.15b}$$

$$D_{\mathrm{B}}(X_1, X_2) = \operatorname{sgn}(\sigma_2 - \sigma_1)\frac{p}{2}\log\frac{\sigma_2}{\sigma_1}, \quad X_i \sim \mathscr{U}^p\left(\mu, \sigma_i^2 \mathbb{I}_p\right),\ i = 1,2, \tag{4.15c}$$

Proof. Since the Hellinger and Bhattacharyya distances are related via $D_{\mathrm{B}} = -\log\left(1 - D_{\mathrm{H}}^2\right)$, Proposition 4.1 implies that

$$D_{\mathrm{B}}(X_1, X_2) = -p\frac{\gamma-1}{\gamma}\log\frac{2(\sigma_2/\sigma_1)^{\frac{\gamma}{2(\gamma-1)}}}{1 + (\sigma_2/\sigma_1)^{\frac{\gamma}{\gamma-1}}}, \tag{4.16}$$

or, equivalently, (4.14), while Corollary 4.1 yields (4.15a)-(4.15c).

□

One can notice that the Bhattacharyya distance between X_1 and X_2 as in (4.14), similar to their corresponding KL divergence as in (2.2), is always proportional to the dimension of X_1 and X_2, i.e. $D_{\mathrm{B}}(X_1, X_2) \propto p \in \mathbb{N}^*$. Especially for the univariate and uniformly distributed r.v.-s $X_i \sim \mathscr{U}(a_i, b_i)$, $i = 1,2$, where $a_1 + b_1 := a_2 + b_2$ is assumed (common mean), relation (4.11) gives

$$D_{\mathrm{B}}(X_1, X_2) = \operatorname{sgn}(b_2 + a_1 - b_1 - a_2)\log\sqrt{\frac{b_2 - a_2}{b_1 - a_1}}. \tag{4.17}$$

For applications of distance measures in Statistics, see [4] and [6] among others. Now we state and proof the following lemma which will be helpful to apply the line of thought we expressed in introduction concerning the "informational Relative Risk" idea.

Lemma 4.3. *The "exponentiated" metric $d_{\mathrm{e}} := \mathrm{e}^d - 1$ of a bounded distance metric $0 \leq d \leq 1$ is also a bounded distance metric.*

Proof. The new defined metric $d_{\mathrm{e}} := \mathrm{e}^{d} - 1$ of a metric $0 \leq d \leq 1$ which is defined on a set $\mathbb{A}$, is a positive-definite and symmetric metric since d is a distance metric. Moreover, d_{e} is also a distance metric, as the triangle inequality is also satisfied. Indeed, for three arbitrary elements $x, y, z \in \mathbb{A}$, the exponential identity $\mathrm{e}^{x} \geq (1 + x/n)^{n}$, $x \in \mathbb{R}$, for $n := 3$, gives

$$d_{\mathrm{e}}(x,y) + d_{\mathrm{e}}(y,z) = \mathrm{e}^{d(x,y)} + \mathrm{e}^{d(y,z)} - 2 \geq \left[1 + \tfrac{1}{3}d(x,y)\right]^{3} + \left[1 + \tfrac{1}{3}d(y,z)\right]^{3} - 2,$$

and using the simplified notations $a := d(x,y)$, $b := d(y,z)$ and $c := d(x,z)$,

$$\begin{aligned} d_{\mathrm{e}}(x,y) + d_{\mathrm{e}}(y,z) &\geq \tfrac{1}{27}\left(a^3 + b^3\right) + \tfrac{1}{3}\left(a^2 + b^2\right) + a + b \\ &= \tfrac{1}{27}(a+b)^3 - \tfrac{1}{9}ab(a+b) + \tfrac{1}{3}(a+b)^2 - \tfrac{2}{3}ab + a + b \\ &\geq \tfrac{1}{27}(a+b)^3 - \tfrac{1}{36}(a+b)^3 + \tfrac{1}{3}(a+b)^2 - \tfrac{1}{6}(a+b)^2 + a + b \\ &\geq \tfrac{1}{3}c^3 + \tfrac{1}{6}c^2 + c, \end{aligned} \tag{4.18}$$

where the triangle inequality of metric d was used as well as the inequality $\sqrt{ab} \leq \frac{1}{2}(a+b)$, $a, b \in \mathbb{R}^{+}$. From the definition of d_{e}, expressing d in terms of d_{e}, relation (4.18) yields

$$d_{\mathrm{e}}(x,y) + d_{\mathrm{e}}(y,z) \geq \tfrac{1}{3}\log^3\left(1 + d_{\mathrm{e}}(x,z)\right) + \tfrac{1}{6}\log^2\left(1 + d_{\mathrm{e}}(x,z)\right) + \log\left(1 + d_{\mathrm{e}}(x,z)\right). \tag{4.19}$$

Consider now the function $\varphi(x) := \frac{1}{3}\log^3(1+x) + \frac{1}{6}\log^2(1+x) + \log(1+x) - x$, $x \in \mathbb{R}^{+}$. Assuming that $\varphi' \leq 0$, we obtain $\log^2(1+x) + \frac{1}{3}\log(1+x) - x \leq 0$, where through the logarithm identity $\log x \geq (x-1)/x$, $x \in \mathbb{R}^{*+}$, gives $4x^2 - 2x - 3 \leq 0$, which holds for $x \geq x_0 := \frac{1}{4}\left(2 + \sqrt{28}\right) \approx 1.822$. Therefore, φ has a global maxima at x_0, and as $x_1 = 0 = \varphi(0)$ is one of the two roots $x_1, x_2 \in \mathbb{R}^{+}$ of φ, the fact that $0 = x_1 \leq x_0$ means that $\varphi(x) \geq 0$ for $x \in [0, x_2]$, where $x_2 \approx 3.5197$ (numerically computed). Therefore from (4.19) the requested triangle inequality $d_{\mathrm{e}}(x,y) + d_{\mathrm{e}}(y,z) \geq d_{\mathrm{e}}(x,z)$ is obtained since $0 \leq d_{\mathrm{e}} \leq \mathrm{e} - 1 \approx 1.718 < x_2$. □

Corollary 4.4. *As an immediate result of Lemma 4.3, the "exponentiated" Hellinger distance, i.e.* $D_{\mathrm{eH}} := e^{D_{\mathrm{H}}} - 1$ *is indeed a distance metric, since* $0 \leq D_{\mathrm{H}} \leq 1$.

Slightly more general, in the proof of the lemma above, note that if $0 \leq d < k := \log(1 + x_2) \approx 1.03775$ is assumed, then d_{e} continuous to be a distance metric of d since the triangle inequality, derived from (4.19) and the non-negativeness of function $\varphi(x)$, $x \in [0, x_2]$, is satisfied since $0 \leq d_{\mathrm{e}} \leq \mathrm{e}^{k} - 1 < x_2$.

Note also that $0 \leq D_{\mathrm{eH}} \leq \mathrm{e} - 1$ since $0 \leq D_{\mathrm{H}} \leq 1$. Therefore, a "standardized" form of the exponentiated Hellinger distance can be defined as

$$D_{\mathrm{eH}} := \frac{\mathrm{e}^{D_{\mathrm{H}}} - 1}{\mathrm{e} - 1}. \tag{4.20}$$

It is known that the Least Square Method is a minimum distance method. So is the Relative Risk method, see [6]. Recall that the odds ratio, [5], is the exponential of a Least Square estimate (i.e. a distance oriented estimator) of the slope for a dichotomous exposure-outcome scenario, modeled by a typical simple logistic regression; see [7] among others. From the

affine-geometric point of view, the Relative Risk method remains invariant under linear transformations; see [8]. For a dichotomous outcome, say $x = 0$ or $x = 1$, the odds ratio values are crucial because they are the only measure which depicts "how much more likely" is for the outcome to be present among "those results with $x = 1$" than among "those with $x = 0$". For example, an odds ratio equals to 2 declares that an outcome occurs twice as often among exposed to a possible Risk than among the non-exposed ones to Risk.

Having the above in mind, the normalized and standardized exponentiated Hellinger distance can considered acting as as a Statistical Relative Risk, although it does not related to "odds ratio", [5]. Therefore, we suggest to refer to it with the term "Information Relative Risk" (IRR). This IRR index, having such a strong information background, offers to researcher a relative "distance associated measure" between two distributions. That is, a distance metric is provided, which has the information-geometric property of distance that the KL divergence does not have, which is also a relative measure of informational comparison between two distributions with different information context. In particular, for IRR values within, say $[0, 0.3)$, we shall conclude that no difference, in terms of the provided information exist between the two distributions (due to the standardized exponentiated Hellinger distance) and, therefore, the "offered information" from both distributions is very close. For IRR values within $[0.3, 0.7)$ our knowledge concerning how far (i.e. how different) one distribution is from the other is inconclusive, and for values within $[0.7, 1]$ we can say the two distributions differ, as they offer different information content for the experiment under investigation.

5 Conclusions

Distance methods have an essential improvement concerning Statistical Estimation Methods in Data Analysis, since the pioneering work of Blyth in [22], where she pointed out that the known estimations methods are trying to minimize the distance of the estimated value from the unknown parameter. Through the concept of distance between two sets [3] defined the normalized distance of two objects as a criterion of "how" different the objects under investigation are. It was also proved that this distance was adopting values within $[0, 1]$. Therefore, a measure of informational "proximity" between two continuous distributions was given.

A probabilistic background, comparing two different outcomes, is not provided with the theory we tried to develop here. We simply tried to investigate if the information provided from two variables following different distributions are "significantly" different. This "significantly" needs an explanation which were addressed by asking what was the "additional information is offered", when a r.v. is coming from the one or the other distribution. The Kullback-Leibler information divergence is a rather popular but also weak method for the quantification of the "gained" information between two r.v.-s, as far as its distance metric formalization is concerned. Therefore, by adopting the Hellinger distance, we can conclude that the normalized exponentiated Hellinger distance (which is also proved to be a distance metric) follows the odds ratio (as well as the log-odds ratio) line of though in Logit methods. That is why the name Information Relative Risk (IRR) was adopted. The proposed IRR measure is recommenced to the experimenter in order to provide him with additional "informational insight" when he has to choose between two comparative distributions both describing the problem under investigation; for example when there is an inconclusive graph between their c.d.f. curves. Based on the above, part of our future work is to give certain numerical results through simulation studies.

Acknowledgement

The authors would like to thank the reviewers for there helpful comments promoting the final version of this paper.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Cover TM, Thomas JA. Elements of Information Theory, 2nd ed. John Wiley and Sons; 2006.

[2] Kullback S, Leibler RA. On information and sufficiency. Ann. Math. Stat. 1951; 22:79-86.

[3] Kitsos CP, Sotiropoulos M. Distance methods for bioassays. In: Kitsos CP., Rigas A., Bieber, K-E. (Eds) Cancer Risk Assessment. Biometrie und Mediziniche Informatik - Greifswalder Seminarberichte. 2009; 15:55–74.

[4] Kitsos CP, Toulias TL. A min-max entropy algorithm for relative risk problems. In: Oliveira T., Biebler K-E., Oliveira A., J ager B. (Eds.) Statistical and Biometrical Challenges, Theory and Applications. Biometrie und Mediziniche Informatik - Greifswalder Seminarberichte. 2014; 23:191–201.

[5] Breslow NE, Day NE. Statistical Methods in Cancer Research. IARC Pub. No. 32, Lyon, France; 1980.

[6] Kitsos CP, Toulias TL, Papageorgiou E. Statistical invariance for the logit and probit model. In: Oliveira T., Biebler K-E., Oliveira A., J ager B. (Eds.) Statistical and Biometrical Challenges, Theory and Applications. Biometrie und Mediziniche Informatik - Greifswalder Seminarberichte. 2014; 23:203–216.

[7] Kitsos CP. On the logit methods for Ca problems (design and fit). Communications in Dependability and Quality Management. 2007; 10(2):88–95.

[8] Kitsos CP. Invariant canonical form for the multiple logistic regression. Mathematics in Engineering Science and Aerospace. 2011; 38(1):267–275.

[9] Kitsos CP, Tavoularis NK. Logarithmic Sobolev inequalities for information measures. IEEE Trans. Inform. Theory. 2009; 55(6):2554–2561.

[10] Kitsos CP, Toulias TL. Inequalities for the Fisher's information measures. In: Rassias MT (ed.) Handbook of Functional Equations: Functional Inequalities, Springer. 2014;281–313.

[11] Kitsos CP, Toulias TL, Trandafir P-C. On the multivariate γ-ordered normal distribution. Far East J. of Theoretical Statistics. 2012; 38(1):49–73.

[12] Kitsos CP, Vassiliadis VG, Toulias TL. MLE for the γ-order generalized normal distribution. Discussiones Mathematicae Probability and Statistics. 2014; 34:143–158.

[13] Hlavácková-Schindler K., Toulias TL, Kitsos CP. Evaluating transfer entropy for normal and γ-order normal distributions. British Journal of Mathematics and Computer Science. 2016; 17(5):1–20.

[14] Kitsos CP, Toulias TL. New information measures for the generalized normal distribution. Information. 2010; 1:13–27.

[15] Toulias TL, Kitsos CP. Kullback-Leibler divergence of the γ-ordered normal over t-distribution. British Journal Of Mathematics And Computer Science. 2012; 2(4):198–212.

[16] Toulias TL, Kitsos CP. Generalizations of entropy and information measures. In: Darras NJ, Rassias MT. (eds.) Computation, Cryptography and Network Security. Springer. 2015;495–526.

[17] Amari S., Nagaoka H. Methods of Information Geometry. A. M. Society, Ed., Oxford University Press; 2000.

[18] Ali SM, Silvey SD. A general class of coefficients of divergence of one distribution from another. JRSS B, 1966; 28(1):131–142.

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

[20] Chernoff H. Measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Ann. Math. Stat.; 1952; 23:493–507.

[21] Cichocki A., Amari S. Families of alpha- beta- and gamma-divergences: flexible and robust measures of similarities. Entropy. 2010; 12(6):1532–1568.

[22] Blyth CR. On the inference and decision models of Statistics. Ann. Math. Stat. 1970; 41:1034–1058.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://sciencedomain.org/review-history/18235