Statistics
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713682269

On divergences of finite measures and their applicability in statistics and information theory

Wolfgang Stummer [ab]; Igor Vajda [c]
[a] Department of Mathematics, University of Erlangen-Nürnberg, Erlangen, Germany [b] School of Business and Economics, University of Erlangen-Nürnberg, Erlangen, Germany [c] Academy of Sciences of the Czech Republic, Institute of Information Theory and Automation, Czech Republic

First Published on: 17 June 2009

To cite this Article Stummer, Wolfgang and Vajda, Igor(2009)'On divergences of finite measures and their applicability in statistics and information theory',Statistics,99999:1,

To link to this Article: DOI: 10.1080/02331880902986919

URL: http://dx.doi.org/10.1080/02331880902986919

Statistics, Vol. 00, No. 0, Month 2009, 1–19

Taylor & Francis
Taylor & Francis Group

On divergences of finite measures and their applicability in statistics and information theory

Wolfgang Stummer[a,b]* and Igor Vajda[c]

[a]*Department of Mathematics, University of Erlangen–Nürnberg, Bismarckstrasse 1 1/2, D-91054 Erlangen, Germany;* [b]*School of Business and Economics, University of Erlangen–Nürnberg, Erlangen, Germany;* [c]*Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Pod Vodárenskou Věží 4, 182 08 Praha 8 – Libeň, Czech Republic*

(*Received 23 November 2005; final version received 18 February 2009*)

Modifications of the classical ϕ-divergences $D_\phi(\mu,\nu)=\int q\phi(p/q)\,\mathrm{d}\lambda$ of finite measures μ,ν on a σ-finite measure space $(\mathcal{X},\mathcal{A},\lambda)$ with Radon–Nikodym densities $p=\mathrm{d}\mu/\mathrm{d}\lambda$, $q=\mathrm{d}\nu/\mathrm{d}\lambda$ are introduced by the formula $\mathfrak{D}_\phi(\mu,\nu)=\int q\tilde{\phi}(p/q)\,\mathrm{d}\lambda$ using the nonnegative convex functions $\tilde{\phi}(t)=\phi(t)-\phi'_+(1)\,(t-1)$. Basic properties of the modified ϕ-divergences are investigated, such as the range of values, symmetry and a decomposition into local and global components. A general ϕ-divergence formula for right-censored observations illustrates the statistical applicability. The Pinsker inequality for finite measures and the generalized Ornstein distance of stationary random processes are among the illustrations of applicability in the information theory.

Keywords: divergences of finite measures; local and global divergences of finite measures; divergences of σ-finite measures; statistical censoring; Pinsker's inequality for finite measures; differential power entropies

AMS Subject Classification: 62B10; 94A17; 62N01; 62C10

1. Introduction

Csiszár [1] introduced ϕ-divergences $D_\phi(\mu,\nu)$ of probability measures μ,ν for real-valued convex functions $\phi(t)$, $t>0$. They satisfy the natural nonnegativity condition if $\phi(1)=0$ even if $\phi(t)$ is negative on some intervals. Liese and Vajda [2] extended ϕ-divergences $D_\phi(\mu,\nu)$ to finite and even infinite measures μ,ν. Their definition admitted negativity of the extended ϕ-divergences resulting from the possible negativity of $\phi(t)$. In this paper, we avoid this inconvenience by replacing the general convex $\phi(t)$ with $\phi(1)=0$ by the nonnegative linear transform $\tilde{\phi}(t)=\phi(t)-\phi'_+(1)\,(t-1)$, where $\phi'_+(1)$ stands for the right-hand derivative of ϕ at $t=1$. The paper investigates in a systematic and rigorous manner the basic properties of the nonnegative ϕ-divergences $\mathfrak{D}_\phi(\mu,\nu):=D_{\tilde{\phi}}(\mu,\nu)$ of finite measures and illuminates the importance of these divergences by selected applications in statistics and information theory.

*Corresponding author. Email: stummer@mi.uni-erlangen.de

ISSN 0233-1888 print/ISSN 1029-4910 online

DOI: 10.1080/02331880902986919
http://www.informaworld.com

The objects of interest in computer-based decisions, machine learning, classification, speech and image compression, automatic information retrieval and other rapidly developing areas of modern computer science seem to be more and more often characterized by various histograms, spectral densities and distribution functions not normalized to 1 and thus representing measures which are typically finite. Relations between such objects are preferred to be based on divergences between the corresponding distributions. The aim of this paper is to provide a rigorous, sufficiently general and practically applicable definition of such divergences (namely the abovementioned $\mathfrak{D}_\phi(\mu, \nu)$) and to propose a simple universal method for the derivation of their properties from the similar properties of divergences of probability measures which are well established in the existing literature.

Section 2 presents detailed definitions and basic properties of the modified ϕ-divergences of finite measures and introduces their extension to the σ-finite measures. Theorem 2.4 enables one to transfer or transform the results concerning probability measures to the more general models with finite measures, and partially also to the σ-finite measure context. Section 3 illustrates the applicability in the statistical theory. Section 4 illustrates the applicability in the information theory.

2. Definition and basic properties

Let $\mathbf{\Phi}$ be the class of convex functions $\phi : (0, \infty) \longmapsto \mathbb{R}$ which are strictly convex at 1 and satisfy the condition $\phi(1) = 0$. For every $\phi \in \mathbf{\Phi}$, we put

$$\phi(0) = \lim_{t \downarrow 0} \phi(t). \tag{1}$$

It is easy to see that if $\phi \in \mathbf{\Phi}$, then the $*$-adjoint function defined by

$$\phi^*(t) = t\phi\left(\frac{1}{t}\right), \quad t > 0, \tag{2}$$

belongs to $\mathbf{\Phi}$ too.

Let $\mathcal{X}$ be an observation space with a given σ-algebra of subsets $\mathcal{A} \neq \{\emptyset, \mathcal{X}\}$ and $\mathcal{M}$ the class of all finite measures on this space, which are not identically zero. If $\mathcal{X}$ is a metric space, then $\mathcal{A}$ is assumed to be the corresponding Borel σ-algebra. By Definition 1.1 and Remark 1.2 in Liese and Vajda [2], the ϕ-divergence of measures $\mu, \nu \in \mathcal{M}$ is well defined for every $\phi \in \mathbf{\Phi}$ by the integral

$$D_\phi(\mu, \nu) = \int_{\mathcal{X}} q\phi\left(\frac{p}{q}\right) \mathrm{d}\lambda, \tag{3}$$

where λ is a σ-finite measure on $(\mathcal{X}, \mathcal{A})$ dominating $\{\mu, \nu\}$ (in symbols $\mu \ll \lambda$, $\nu \ll \lambda$) and $p = \mathrm{d}\mu/\mathrm{d}\lambda$, $q = \mathrm{d}\nu/\mathrm{d}\lambda$ are the corresponding Radon–Nikodym densities. Behind the integral, it is assumed that

$$0\phi\left(\frac{p}{0}\right) = \lim_{t \downarrow 0} t\,\phi\left(\frac{p}{t}\right) = p\phi^*(0) \quad \text{(cf. (1), (2))} \tag{4}$$

with the convention $0\phi^*(0) = 0$ even if $\phi^*(0)$ is infinite. As argued by Liese and Vajda, the value of the classical ϕ-divergence (3) is independent of the choice of the dominating measure λ.

For the special case $\mu(\mathcal{X}) = \nu(\mathcal{X})$, it is straightforward to show (e.g. from the discussion below Definition 2.1) that $D_\phi(\mu, \nu) \geq 0$, and that $D_\phi(\mu, \nu) = 0$ if and only if the measures μ and ν totally coincide on $\mathcal{A}$. However, in general, the values of the divergence (3) may be negative or $D_\phi(\mu, \nu)$ may be zero even if the measures μ and ν differ on $\mathcal{A}$. This can be seen, for instance, from the following:

Example 2.1 Let μ, ν be the measures on the interval $\mathcal{X} = (0, 1)$ with the densities

$$p(x) = \mathrm{e}^{-x}, \quad q(x) = \frac{\mathrm{e}^{1-x}}{\mathrm{e}-1}.$$

Then it is easy to verify that for $\phi_1(t) = t \ln t$, we get the negative information divergence $D_{\phi_1}(\mu, \nu) < 0$. On the other hand, let μ, ν be measures on $\mathcal{X} = (0, \infty)$ with the densities

$$p(x) = \mathbf{1}_{]0,1[}(x)\mathrm{e}^{-x}, \quad q(x) = \mathrm{e}^{-x},$$

where $\mathbf{1}_A(\cdot)$ stands here and in the sequel for the indicator function of a set A. Then $\mu \neq \nu$ but $D_\phi(\mu, \nu) = 0$ not only for the above considered $\phi = \phi_1$ but for all $\phi \in \mathbf{\Phi}$ such that $\phi(0) = 0$.

For the rest of this paper, $D_\phi(\mu, \nu)$ defined by Equation (3) is referred to as *classical ϕ-divergence of measures μ, ν*. To avoid the paradoxes demonstrated in Example 2.1, we slightly modify definition (3), and the modified concept is referred to simply as *ϕ-divergence of measures μ, ν*. As mentioned in the introduction, the modification consists in the replacement of the functions $\phi \in \mathbf{\Phi}$ by their nonnegative transforms $\tilde{\phi} \in \mathbf{\Phi}$, where

$$\tilde{\phi}(t) = \phi(t) - \phi'_+(1)(t-1) \begin{cases} > 0, & \text{for } t \neq 1, \\ = 0, & \text{for } t = 1, \end{cases} \tag{5}$$

and $\phi'_+(1)$ denotes the right-hand derivative $\phi'_+(t)$ of $\phi(t)$ at $t = 1$.

Since $\phi(1) = 0$ and $\phi(1) + \phi'_+(1)(t-1)$ is the support line of $\phi(t)$ at $t = 1$, it follows from the assumed strict convexity of $\phi(t)$ at $t = 1$ that the equality as well as the strict inequality of Equation (5) hold.

DEFINITION 2.1 *For every $\phi \in \mathbf{\Phi}$, the ϕ-divergence of measures $\mu, \nu \in \mathcal{M}$ is defined by the formula*

$$\mathfrak{D}_\phi(\mu, \nu) = \int_{\mathcal{X}} q\, \tilde{\phi}\left(\frac{p}{q}\right) \mathrm{d}\lambda, \tag{6}$$

where p, q, λ are the same as in Equation (3)*, $\tilde{\phi} \in \mathbf{\Phi}$ is defined by Equation* (5) *and the conventions* (4) *are applied to $\tilde{\phi}$.*

The nonnegative transforms $\tilde{\phi} \in \mathbf{\Phi}$ of the functions $\phi \in \mathbf{\Phi}$ were used previously by Liese and Vajda [2] in the model with probability measures $(\mu, \nu) = (P, Q)$. These authors frequently applied the equality $D_{\tilde{\phi}}(P, Q) = D_\phi(P, Q)$.

Like the classical ϕ-divergence $D_\phi(\mu, \nu)$, also the ϕ-divergence $\mathfrak{D}_\phi(\mu, \nu)$ does not depend on the concrete choice of dominating measure λ but, contrary to Equation (3), the expression (6) is always nonnegative. From Equations (5) and (6), we get the relation

$$\mathfrak{D}_\phi(\mu, \nu) = D_\phi(\mu, \nu) + \phi'_+(1)[\nu(\mathcal{X}) - \mu(\mathcal{X})] \tag{7}$$

between the two concepts of ϕ-divergence, which shows in particular that this nonnegativity is achieved by adding an appropriate compensation term to $D_\phi(\mu, \nu)$. Furthermore, Equation (7) implies also the coincidence

$$\mathfrak{D}_\phi(P, Q) = D_\phi(P, Q) \quad \text{for all probability measures } P, Q \in \mathcal{M} \tag{8}$$

between the ϕ-divergence and the classical ϕ-divergence.

Example 2.2 For every $\alpha \in \mathbb{R}$ with $\alpha(\alpha-1) \neq 0$, the power functions

$$\phi_\alpha(t) = \frac{t^\alpha - 1}{\alpha(\alpha-1)} \tag{9}$$

belong to $\mathbf{\Phi}$, and their transforms corresponding to Equation (5) are

$$\tilde{\phi}_\alpha(t) = \frac{t^\alpha - \alpha(t-1) - 1}{\alpha(\alpha-1)}. \tag{10}$$

Their limits for α tending to 1 resp. 0 are

$$\tilde{\phi}_1(t) = t\ln t - t + 1 \quad \text{resp.} \quad \tilde{\phi}_0(t) = -\ln t + t - 1, \tag{11}$$

which are the (5)-type transforms of the functions

$$\phi_1(t) = t\ln t \quad \text{resp.} \quad \phi_0(t) = -\ln t \tag{12}$$

belonging to $\mathbf{\Phi}$ too. For the classical power divergences $D_{\phi_\alpha}(\mu, \nu)$ (defined by Equations (3), (9) and (12)) and the power divergences $\mathfrak{D}_{\phi_\alpha}(\mu, \nu)$ (defined by Equations (6), (10) and (11)), we use the simplified notation

$$D_\alpha(\mu, \nu) = D_{\phi_\alpha}(\mu, \nu) \quad \text{and} \quad \mathfrak{D}_\alpha(\mu, \nu) = \mathfrak{D}_{\phi_\alpha}(\mu, \nu), \quad \alpha \in \mathbb{R}. \tag{13}$$

Furthermore, in accordance with the common notation $I(P, Q)$ for the classical information divergence (Kullback–Leibler divergence) $D_1(P, Q)$ of probability measures P, Q, the corresponding divergence $\mathfrak{D}_1(\mu, \nu)$ of finite measures μ, ν is denoted by $\mathfrak{I}(\mu, \nu)$ and called information divergence (Kullback–Leibler divergence) of finite measures μ, ν. In other words,

$$\mathfrak{I}(\mu, \nu) \equiv \mathfrak{D}_1(\mu, \nu) = D_1(\mu, \nu) + \nu(\mathcal{X}) - \mu(\mathcal{X}) \quad \text{for all } \mu, \nu \in \mathcal{M}, \tag{14}$$

which corresponds to Equation (7). For $\alpha \neq 1$, we get from Equations (6) and (7)

$$\mathfrak{D}_\alpha(\mu, \nu) = D_{\tilde{\phi}_\alpha}(\mu, \nu) = D_\alpha(\mu, \nu) + \frac{\nu(\mathcal{X}) - \mu(\mathcal{X})}{\alpha - 1}. \tag{15}$$

The first two general results deal with symmetry properties of the ϕ-divergences (6).

THEOREM 2.1 *If ϕ belongs to $\mathbf{\Phi}$, then the function*

$$\varphi(t) = \phi(t) + c\,(t-1) \qquad \textit{where } c \in \mathbb{R} \tag{16}$$

belongs to $\mathbf{\Phi}$ too, and for all $\mu, \nu \in \mathcal{M}$,

$$\mathfrak{D}_\varphi(\mu, \nu) = \mathfrak{D}_\phi(\mu, \nu) \tag{17}$$

holds. Conversely, if for some $\varphi, \phi \in \mathbf{\Phi}$ the equality (17) *holds for all $\mu, \nu \in \mathcal{M}$, then φ and ϕ must satisfy* (16).

Proof Equation (17) is obvious from the definition (6), as Equation (16) implies $\tilde{\varphi}(t) = \tilde{\phi}(t)$. On the other hand, if Equation (17) holds for some $\varphi, \phi \in \mathbf{\Phi}$ and for all $\mu, \nu \in \mathcal{M}$, then it follows from Equation (8) that

$$D_\varphi(P, Q) = D_\phi(P, Q)$$

for all probability measures $P, Q \in \mathcal{M}$. Therefore, Equation (16) follows from Proposition 1.13 in Liese and Vajda [2]. ■

THEOREM 2.2 *If ϕ belongs to $\mathbf{\Phi}$, then the function*

$$\varphi(t) = t\phi\left(\frac{1}{t}\right) + c(t-1) \quad \textit{for } c \in \mathbb{R} \tag{18}$$

belongs to $\mathbf{\Phi}$ too, and if additionally ϕ is differentiable at $t = 1$, then for all $\mu, \nu \in \mathcal{M}$,

$$\mathfrak{D}_\varphi(\mu, \nu) = \mathfrak{D}_\phi(\nu, \mu) \tag{19}$$

holds. Conversely, if for some $\varphi, \phi \in \mathbf{\Phi}$ the equality (19) *holds for all $\mu, \nu \in \mathcal{M}$, then φ and ϕ must satisfy Equation* (18).

Proof Similar to the previous proof, the first assertion is easily verifiable directly from Equation (18). If $\phi \in \mathbf{\Phi}$, then, by Equations (7) and (18), for all $\mu, \nu \in \mathcal{M}$

$$\mathfrak{D}_\varphi(\mu, \nu) = D_\phi(\nu, \mu) + (\varphi'_+(1) - c)[\nu(\mathcal{X}) - \mu(\mathcal{X})]$$

and

$$\mathfrak{D}_\phi(\nu, \mu) = D_\phi(\nu, \mu) + \phi'_+(1)[\mu(\mathcal{X}) - \nu(\mathcal{X})]$$

hold. But

$$\varphi'_+(1) = -\phi'_-(1) + c, \tag{20}$$

so that in the case of $\phi'_-(1) = \phi'_+(1)$ the desired equality (19) holds. The last assertion follows from Proposition 1.13 in Liese and Vajda [2] too, because its assumptions imply the relation

$$D_\varphi(P, Q) = D_\phi(Q, P)$$

for all probability measures $P, Q \in \mathcal{M}$. ■

Example 2.3 Since $\phi_{1-\alpha}(t) = t\phi_\alpha(1/t) + (t-1)/[\alpha(\alpha-1)]$ when $\alpha(\alpha-1) \neq 0$ and $\phi_{1-\alpha}(t) = t\phi_\alpha(1/t)$ when $\alpha(\alpha-1) = 0$, we get from Theorem 2.2 the skew symmetry

$$\mathfrak{D}_\alpha(\nu, \mu) = \mathfrak{D}_{1-\alpha}(\mu, \nu), \quad \alpha \in \mathbb{R}. \tag{21}$$

Notice also that without the differentiability assumption on ϕ, the symmetry (19) may break down. For instance, take $\phi(t) = |t-1|$ and $c = 0$, and thus $\varphi(t) = \phi(t)$ by Equation (18). Accordingly, by Equation (7) and $\varphi'_+(1) = \phi'_+(1) = 1$, one gets the total variations

$$\mathfrak{D}_\varphi(\mu, \nu) = \mathfrak{V}(\mu, \nu) = \int_{\mathcal{X}} |p - q| \, \mathrm{d}\lambda + \mu(\mathcal{X}) - \nu(\mathcal{X}) \tag{22}$$

and

$$\mathfrak{D}_\phi(\nu, \mu) = \mathfrak{V}(\nu, \mu) = \int_{\mathcal{X}} |q - p| \, \mathrm{d}\lambda + \nu(\mathcal{X}) - \mu(\mathcal{X}),$$

which are unequal unless the total masses $\mu(\mathcal{X})$ and $\nu(\mathcal{X})$ coincide.

The following assertion follows directly from the assertions of Theorem 2.2 and Equation (20).

COROLLARY 2.1 *Let ϕ belong to $\mathbf{\Phi}$ and be differentiable at $t = 1$. Then the ϕ-divergence* (6) *is symmetric in the sense*

$$\mathfrak{D}_\phi(\mu, \nu) = \mathfrak{D}_\phi(\nu, \mu) \quad \textit{for all } \mu, \nu \in \mathcal{M}$$

if and only if

$$\phi(t) = t\phi\left(\frac{1}{t}\right) + 2\phi'(1)(t-1) \quad \textit{for all } t \in (0, \infty). \tag{23}$$

For instance, the choice $\phi_{1/2}(t) = 4(1 - \sqrt{t})$ in Example 2.2 satisfies Equation (23) and leads to the symmetry of the Hellinger divergence

$$\mathfrak{D}_{1/2}(\mu, \nu) = 2\int_{\mathcal{X}} (\sqrt{p} - \sqrt{q})^2 \, \mathrm{d}\lambda = \mathfrak{D}_{1/2}(\nu, \mu) \tag{24}$$

for all $\mu, \nu \in \mathcal{M}$. This is also consistent with Equation (21).

Next, we derive directly from Definition 2.1 some upper bounds for the divergences $\mathfrak{D}_\phi(\mu, \nu)$, $\phi \in \mathbf{\Phi}$. As a first step, from Equations (3), (4) and (6) and from the assumptions $\phi(1) = \tilde{\phi}(1) = 0$, we get

$$\begin{aligned} \mathfrak{D}_\phi(\mu, \nu) &= \int_{\{p \neq q\}} q\tilde{\phi}\left(\frac{p}{q}\right) \mathrm{d}\lambda \\ &= \int_{\{p<q\}} q\tilde{\phi}\left(\frac{p}{q}\right) \mathrm{d}\lambda + \int_{\{q<p\}} p\tilde{\phi}^*\left(\frac{q}{p}\right) \mathrm{d}\lambda, \end{aligned} \tag{25}$$

where

$$\tilde{\phi}^*(t) = \phi^*(t) + \phi'_+(1)\,(t-1)$$

is adjoint to $\tilde{\phi}$ in the sense of Equation (2). By the generalized Taylor formula for convex functions proved in Theorem 1 of Liese and Vajda [3], for every $\phi \in \mathbf{\Phi}$ with $\phi'_+(1) = 0$ and every $0 \leq t \leq 1$,

$$\phi(t) = \int \mathbf{1}_{]t,1]}(s)(s-t)\,\mathrm{d}\lambda_\phi(s),$$

holds, where λ_ϕ is the unique extension of the measure

$$\lambda_\phi(]a, b]) = \phi'_+(b) - \phi'_+(a)$$

of intervals $]a, b] \subset (0, \infty)$ on the Borel subsets of $\mathbb{R}$. The strict convexity of ϕ at 1 implies that λ_ϕ is strictly positive in the neighbourhood of $s = 1$, so that the strict monotonicity $\mathbf{1}_{]t_1,1]}(s)(s-t_1) > \mathbf{1}_{]t_2,1]}(s)(s-t_2)$ of integrands for any $0 \leq t_1 < t_2 \leq 1$ implies the monotonicity

$$0 = \phi(1) \leq \phi(t_2) < \phi(t_1) \leq \phi(0).$$

Applying this result to $\phi = \tilde{\phi}$ and $\phi = \tilde{\phi}^*$, we get from Equation (25)

$$\begin{aligned} 0 = \phi(1) \leq \mathfrak{D}_\phi(\mu, \nu) &= \int_{\{p<q\}} \tilde{\phi}\left(\frac{p}{q}\right) \mathrm{d}\nu + \int_{\{q<p\}} \tilde{\phi}^*\left(\frac{q}{p}\right) \mathrm{d}\mu \\ &\leq \nu(\mathcal{X})\,\tilde{\phi}(0) + \mu(\mathcal{X})\,\tilde{\phi}^*(0), \end{aligned} \tag{26}$$

where the left inequality is strict unless $\nu(\{p < q\}) = \mu(\{q < p\}) = 0$, and the right inequality is strict unless $\nu(\{p = 0\}) = \nu(\mathcal{X})$ and $\mu(\{q = 0\}) = \mu(\mathcal{X})$, except for the cases where one of

the involved integrals is infinite. Since $\tilde{\phi}(0) = \phi(0) + \phi'_+(1)$, $\tilde{\phi}^*(0) = \phi^*(0) - \phi'_+(1)$, we have proved the following result.

THEOREM 2.3 *The divergence* (6) *satisfies for all* $\mu, \nu \in \mathcal{M}$ *the inequalities*

$$0 \leq \mathfrak{D}_\phi(\mu, \nu) \leq \nu(\mathcal{X})\,\phi(0) + \mu(\mathcal{X})\phi^*(0) + \phi'_+(1)[\nu(\mathcal{X}) - \mu(\mathcal{X})], \tag{27}$$

where the left equality holds if and only if $\mu = \nu$ *and the right equality holds if* $\mu \perp \nu$ (*singularity*). *For* $\phi(0) + \phi^*(0) < \infty$, *the right equality holds if and only if* $\mu \perp \nu$.

If $\phi(0) = \infty$ or $\phi^*(0) = \infty$, then the upper bound of $\mathfrak{D}_\phi(\mu, \nu)$ in Inequality (27) is ∞. If $\phi(0) = \infty$ holds, then one can see from Inequality (26) that already the condition $\nu(p = 0) > 0$, i.e. $\nu \not\ll \mu$ (and not only the more restrictive condition $\mu \perp \nu$ stated in Theorem 2.3) implies $\mathfrak{D}_\phi(\mu, \nu) = \infty$. Similarly, if $\phi^*(0) = \infty$ holds, then already the condition $\mu \not\ll \nu$ (and not only the condition $\mu \perp \nu$) implies $\mathfrak{D}_\phi(\mu, \nu) = \infty$.

Example 2.4 Let us apply Theorem 2.3 to the power divergences of Example 2.2. (i) In the case $\alpha \in]0, 1[$, one gets $\phi_\alpha(0) = 1/[\alpha(1-\alpha)]$, $\phi^*_\alpha(0) = 0$, $\phi'_{\alpha+}(1) = 1/(\alpha - 1)$, and thus from Inequality (27)

$$0 \leq \mathfrak{D}_\alpha(\mu, \nu) \leq \frac{\nu(\mathcal{X})}{\alpha} + \frac{\mu(\mathcal{X})}{1-\alpha}, \quad \text{if } \alpha \in]0, 1[,$$

where the left equality holds if and only if $\mu = \nu$ and the right equality holds if and only if $\mu \perp \nu$. (ii) In the case $\alpha \leq 0$, it follows $\phi_\alpha(0) = \infty$ and $\phi^*_\alpha(0) = 0$, and thus from Inequality (27)

$$0 \leq \mathfrak{D}_\alpha(\mu, \nu) \leq \infty, \tag{28}$$

where the left equality holds if and only if $\mu = \nu$ and $\mathfrak{D}_\alpha(\mu, \nu) = \infty$ not only if $\mu \perp \nu$ (as stated in Theorem 2.3) but already if $\nu \not\ll \mu$. (iii) The case $\alpha \geq 1$ leads to finite $\phi_\alpha(0)$ as well as to $\phi^*_\alpha(0) = \infty$, and thus from Inequality (27) one obtains Inequality (28) where the left equality holds if and only if $\mu = \nu$ and $\mathfrak{D}_\alpha(\mu, \nu) = \infty$ not only if $\mu \perp \nu$ (as stated in Theorem 2.3) but already if $\mu \not\ll \nu$. It is easy to verify that the paradoxes of Example 2.1 disappear if $D_\phi(\mu, \nu)$ is replaced by the divergence $\mathfrak{D}_\phi(\mu, \nu)$.

Further properties of the divergences (6) can be obtained from the following theorem which relates the ϕ-divergence of measures to the ϕ-divergences of probability measures. To formulate this, put for arbitrary $\mu, \nu \in \mathcal{M}$ with $\mu(\mathcal{X})\nu(\mathcal{X}) > 0$ and $\phi \in \mathbf{\Phi}$

$$R_{\mu,\nu} = \frac{\mu(\mathcal{X})}{\nu(\mathcal{X})}, \quad \phi_{\mu,\nu}(t) = \phi(R_{\mu,\nu}t) - \phi(R_{\mu,\nu}) \tag{29}$$

and

$$P_\mu = \frac{\mu}{\mu(\mathcal{X})}, \quad P_\nu = \frac{\nu}{\nu(\mathcal{X})}. \tag{30}$$

THEOREM 2.4 *The function* $\phi_{\mu,\nu}$ *of Equation* (29) *belongs to* $\mathbf{\Phi}$, *the normalized measures* P_μ *and* P_ν *of Equation* (30) *are probability measures, and the* ϕ*-divergence of* μ, ν *decomposes as follows*:

$$\mathfrak{D}_\phi(\mu, \nu) = \nu(\mathcal{X})[\mathbf{\Lambda}_\phi(\mu, \nu) + \mathbf{\Gamma}_\phi(\mu, \nu)],$$

where the nonnegative components

$$\mathbf{\Lambda}_\phi(\mu, \nu) = D_{\phi_{\mu,\nu}}(P_\mu, P_\nu) \quad \textit{resp.} \quad \mathbf{\Gamma}_\phi(\mu, \nu) = \phi(R_{\mu,\nu}) - \phi'_+(1)(R_{\mu,\nu} - 1)$$

can be interpreted as local resp. global ϕ*-divergence of measures* μ, ν.

Proof Let $\tilde{\phi}$ be defined by Equation (5). Then $\phi_{\mu,\nu}$ of Equation (29) satisfies the relation

$$\phi_{\mu,\nu}(t) = \tilde{\phi}(R_{\mu,\nu}t) - \tilde{\phi}(R_{\mu,\nu}) + R_{\mu,\nu}\phi'_+(1)(t-1)$$

and belongs to $\mathbf{\Phi}$. Hence, by Equation (3)

$$\begin{aligned} D_{\phi_{\mu,\nu}}(P_\mu, P_\nu) &= \int_{\mathcal{X}} \left[\tilde{\phi}\left(\frac{R_{\mu,\nu}\nu(\mathcal{X})p}{\mu(\mathcal{X})q}\right) - \tilde{\phi}(R_{\mu,\nu})\right]\frac{q}{\nu(\mathcal{X})}\mathrm{d}\lambda \\ &= \frac{1}{\nu(\mathcal{X})}\left[\int_{\mathcal{X}} q\tilde{\phi}\left(\frac{p}{q}\right)\mathrm{d}\lambda - \nu(\mathcal{X})\tilde{\phi}(R_{\mu,\nu})\right]. \end{aligned}$$

By using Equation (6), we obtain from here

$$\mathfrak{D}_\phi(\mu,\nu) = \nu(\mathcal{X})D_{\phi_{\mu,\nu}}(P_\mu, P_\nu) + \nu(\mathcal{X})\tilde{\phi}(R_{\mu,\nu}),$$

where by Equation (5) $\tilde{\phi}(R_{\mu,\nu}) = \phi(R_{\mu,\nu}) - \phi'_+(1)(R_{\mu,\nu}-1)$, which completes the proof. ■

Example 2.5 Let us apply Theorem 2.4 to the power divergences of Example 2.2. We get for all $\alpha \in \mathbb{R}$

$$\mathfrak{D}_\alpha(\mu,\nu) = \nu(\mathcal{X})[\mathbf{\Lambda}_\alpha(\mu,\nu) + \mathbf{\Gamma}_\alpha(\mu,\nu)], \tag{31}$$

where $\mathbf{\Lambda}_\alpha(\mu,\nu)$ is the local and $\mathbf{\Gamma}_\alpha(\mu,\nu)$ the global power divergence of measures μ, ν given by the formulas

$$\mathbf{\Lambda}_\alpha(\mu,\nu) = (R_{\mu,\nu})^\alpha D_\alpha(P_\mu, P_\nu) \quad \text{and} \quad \mathbf{\Gamma}_\alpha(\mu,\nu) = \tilde{\phi}_\alpha(R_{\mu,\nu}) \tag{32}$$

for $R_{\mu,\nu}$ given by Equation (29), P_μ, P_ν given by Equation (30) and $\tilde{\phi}_\alpha$ given by Equation (10). In particular,

$$\mathfrak{I}(\mu,\nu) = \nu(\mathcal{X})[\mathbf{\Lambda}(\mu,\nu) + \mathbf{\Gamma}(\mu,\nu)],$$

where

$$\mathbf{\Lambda}(\mu,\nu) = \mathbf{\Lambda}_1(\mu,\nu) = R_{\mu,\nu}I(P_\mu, P_\nu) \tag{33}$$

is the local and

$$\mathbf{\Gamma}(\mu,\nu) = \mathbf{\Gamma}_1(\mu,\nu) = R_{\mu,\nu}\ln R_{\mu,\nu} - R_{\mu,\nu} + 1 \tag{34}$$

the global information divergence of measures μ, ν.

Theorem 2.4 enables one to reformulate all the theorems concerning the classical ϕ-divergences $D_\phi(P, Q)$ of probability measures P, Q established in the previous literature to the ϕ-divergences $\mathfrak{D}_\phi(\mu,\nu)$ of finite measures μ, ν. An illustration will be given in Theorems 4.1 and 4.2 in Section 4.

3. Statistical applicability

It is well known that ϕ-divergences of probability distributions play an important role in the statistical inference (see, e.g. minimum divergence estimation and testing in Liese and Vajda [3], minimum divergence testing in Morales *et al.* [4,5] and divergence-based decisions in Stummer and Vajda [6], Stummer [7,8]). It is easy to find situations where the evaluation of ϕ-divergences of probability distributions reduces to the evaluation of ϕ-divergences of finite measures. Classical examples can be found in Liese and Vajda [2] and references therein. For example, by

[2, Theorem 3.30], the Rényi divergence of the order $\alpha \in \mathbb{R}$ of two Poisson processes with intensity measures ν_1, ν_2 is the power divergence $\mathfrak{D}_\alpha(\nu_1, \nu_2)$ defined by Equation (13) when ν_1, ν_2 are finite. The power divergences $\mathfrak{D}_\alpha(\nu_1, \nu_2)$ play a similar role also in [2, Theorem 4.24] dealing with the Rényi divergences of Lévy processes with characteristic triplets (a_1, b_1, ν_1) and (a_2, b_2, ν_2) (these divergences were used for goodness-of-fit testing, for example, in Morales *et al.* [4]).

In this section, we look in more detail at the statistical model with censored observations (see, e.g. Miller [9]) where various measures of uncertainty, informativity and divergence were studied previously, e.g. by Hollander *et al.* [10], Stute [11,12] and Tsairidis *et al.* [13,14]. We illustrate the applicability of the ϕ-divergences of finite measures by a rigorous evaluation of the general ϕ-divergence of probability distributions of randomly right-censored observations.

In more detail, let us consider two independent real-valued random variables X and Y defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with distribution functions F and G, where X is supposed to have the Lebesgue density $f = \mathrm{d}F/\mathrm{d}x$.

In the randomly right-censored statistical model, one assumes the observation space $\mathcal{X} = \mathcal{X}_1 \otimes \mathcal{X}_2 = \mathbb{R} \otimes \{0, 1\}$ and the $(\mathbb{R} \otimes \{0, 1\})$-valued observed variable

$$W = (Z, U),$$

where

$$Z = \min\{X, Y\} \quad \text{and} \quad U = \mathbf{1}_{(-\infty, Y]}(X)$$

are $\mathbb{R}$-valued and $\{0, 1\}$-valued random variables, respectively. The probability distribution P of W on $\mathcal{X}$ is defined on the σ-algebra $\mathcal{A}$ generated by the products $A \otimes B$ of Borel sets $A \subset \mathbb{R}$ and arbitrary subsets $B \subset \{0, 1\}$ by the condition $P(A \times B) = \mathbb{P}(Z \in A, U \in B)$. P is dominated by the σ-finite measure λ defined on $\mathcal{A}$ by the condition

$$\lambda(A \times B) = \boldsymbol{\delta_0}(B) \int_A \mathrm{d}G(y) + \boldsymbol{\delta_1}(B) \int_A \mathrm{d}x, \tag{35}$$

where, as usual, $\boldsymbol{\delta_u}$ is the Dirac probability measure with all mass concentrated at $u \in \{0, 1\}$. The density of P with respect to λ has the form

$$\frac{\mathrm{d}P}{\mathrm{d}\lambda}(x, u) = \mathbf{1}_{\{0\}}(u)(1 - F(x)) + \mathbf{1}_{\{1\}}(u)(1 - G(x))f(x), \quad (x, u) \in \mathbb{R} \otimes \{0, 1\} \tag{36}$$

illustrated in Example 3.1 below. This density follows from the formulas

$$\begin{aligned} P(]-\infty, z] \times \{0\}) &= \mathbb{P}(Z \le z, U = 0) = \mathbb{P}(Y \le z, Y < X) \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} \mathbf{1}_{]-\infty, z]}(y) \mathbf{1}_{]y, \infty[}(x)\, \mathrm{d}F(x)\, \mathrm{d}G(y) \\ &= \int_{]-\infty, z]} \mathbb{P}(X > y)\, \mathrm{d}G(y) = \int_{]-\infty, z]} (1 - F(x))\, \mathrm{d}G(x) \qquad (37) \\ &= \int_{]-\infty, z] \times \{0\}} [\mathbf{1}_{\{0\}}(u)(1 - F(x)) + \mathbf{1}_{\{1\}}(u)(1 - G(x))f(x)]\, \mathrm{d}\lambda(x, u) \end{aligned}$$

for the marginal distribution corresponding to $u = 0$ and

$$\begin{aligned} P(]-\infty, z] \times \{1\}) &= \mathbb{P}(Z \le z, U = 1) = \mathbb{P}(X \le z, X \le Y) \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} \mathbf{1}_{]-\infty,z]}(x) \mathbf{1}_{[x,\infty[}(y)\, \mathrm{d}F(x)\, \mathrm{d}G(y) \\ &= \int_{]-\infty,z]} \mathbb{P}(Y \ge x)\, \mathrm{d}F(x) = \int_{]-\infty,z]} (1 - G(x)) f(x)\, \mathrm{d}x \\ &= \int_{]-\infty,z]\times\{1\}} [\mathbf{1}_{\{0\}}(u)(1 - F(x)) + \mathbf{1}_{\{1\}}(u)(1 - G(x)) f(x)]\, \mathrm{d}\lambda(x, u) \end{aligned} \tag{38}$$

corresponding to $u = 1$. The fact that this density (36) integrates to 1 can be verified by the integration-by-parts,

$$1 = \int_{\mathbb{R}} \mathrm{d}(F \cdot G)(x) = \int_{\mathbb{R}} G(x) f(x)\, \mathrm{d}x + \int_{\mathbb{R}} F(x)\, \mathrm{d}G(x). \tag{39}$$

Indeed, by the above computations, one gets

$$\begin{aligned} \int_{\mathbb{R}\otimes\{0,1\}} \frac{\mathrm{d}P}{\mathrm{d}\lambda}(x, u)\, \mathrm{d}\lambda(x, u) &= \int_{\mathbb{R}} (1 - G(x))\, \mathrm{d}F(x) + \int_{\mathbb{R}} (1 - F(x))\, \mathrm{d}G(x) \\ &= 2 - \int_{\mathbb{R}} G(x) f(x)\, \mathrm{d}x - \int_{\mathbb{R}} F(x)\, \mathrm{d}G(x) = 1 \qquad \text{(cf. (39))}. \end{aligned} \tag{40}$$

Example 3.1 Figures 1–3 illustrate for various types of censoring Y the marginal distributions $P(]-\infty, z] \times \{u\})$ of W for $u = 0$ and $u = 1$.

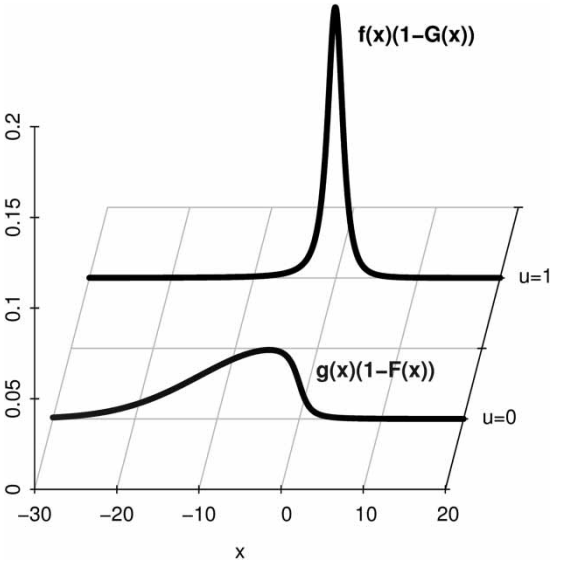

Figure 1. The marginal distributions of Example 3.1(i).

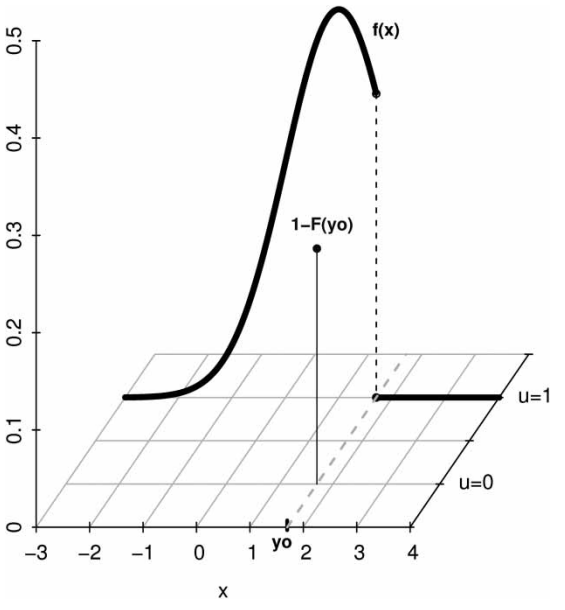

Figure 2. The marginal distributions of Example 3.1(ii).

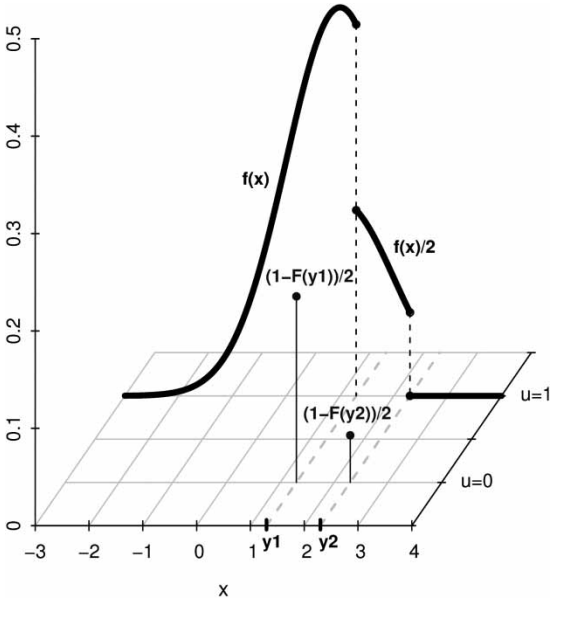

Figure 3. The marginal distributions of Example 3.1(iii).

(i) Let the censoring Y be continuous with a Lebesgue density $g = \mathrm{d}G/\mathrm{d}x$. Accordingly, Figure 1 presents at the coordinate $u = 0$ the Lebesgue density of the distribution

$$P(]-\infty, z] \times \{0\}) = \mathbb{P}(Z \leq z, U = 0) = \int_{]-\infty,z]} (1 - F(x))g(x)\,\mathrm{d}x$$

(cf. Equation (37)) and at the coordinate $u = 1$ the Lebesgue density of $P(]-\infty, z] \times \{1\})$ given in Equation (38) (exemplarily, we have used Student's t-distribution function F with two degrees of freedom, and the normal distribution function G with mean -2 and variance 100).

(ii) Let the censoring Y be constant, i.e. let $\mathbb{P}(Y = y_0) = 1$ for some fixed $y_0 \in \mathbb{R}$. Then $G(x) = \mathbf{1}_{[y_0,\infty[}(x)$, and Equation (37) implies

$$P(]-\infty, z] \times \{0\}) = \int_{]-\infty,z]} (1 - F(x))\, \mathrm{d}G(x) = (1 - F(y_0))\mathbf{1}_{[y_0,\infty[}(z).$$

The corresponding discrete density of this marginal distribution function is drawn at the coordinate $u = 0$ in Figure 2 (exemplarily, we have used $y_0 = 1.7$ and the normal distribution function F with mean 1 and variance 1). Furthermore, the formula (38) reduces to $P(]-\infty, z] \times \{1\}) = \int_{]-\infty,z]} \mathbf{1}_{]-\infty,y_0[}(x) f(x)\, \mathrm{d}x$. The corresponding Lebesgue density is drawn at the coordinate $u = 1$ in Figure 2.

(iii) Let the censoring Y be discrete and take for simplicity the binomial case where $\mathbb{P}(Y = y_1) = \mathbb{P}(Y = y_2) = 1/2$ for some fixed real values $y_1 < y_2$. Then $G(x) = (1/2)\mathbf{1}_{[y_1,\infty[}(x) + \mathbf{1}_{[y_2,\infty[}(x)$, and Equation (37) implies

$$\begin{aligned} P(]-\infty, z] \times \{0\}) &= \int_{]-\infty,z]} (1 - F(x))\, \mathrm{d}G(x) \\ &= \frac{1 - F(y_1)}{2}\, \mathbf{1}_{[y_1,\infty[}(z) + \frac{1 - F(y_2)}{2}\, \mathbf{1}_{[y_2,\infty[}(z). \end{aligned}$$

The corresponding discrete density is drawn at the coordinate $u = 0$ in Figure 3 (exemplarily, we have used $y_1 = 1.3$, $y_2 = 2.3$ and the normal distribution function F with mean 1 and variance 1). Moreover, the formula (38) gives the absolutely continuous distribution

$$P(]-\infty, z] \times \{1\}) = \int_{]-\infty,z]} \left(\mathbf{1}_{]-\infty,y_1[}(x) f(x) + \mathbf{1}_{[y_1,y_2[}(x) \frac{f(x)}{2} \right) \mathrm{d}x,$$

with the corresponding Lebesgue density visible at the coordinate $u = 1$ in Figure 3.

Suppose that one has to decide between two different statistical situations where the observations are distributed by F_1 or F_2 with densities f_1, f_2 on $\mathbb{R}$, but practically available only in the form randomly right-censored according to a general distribution function G considered above. Then one has in fact to decide between the statistical models characterized by versions P_1, P_2 of the probability measure P defined on $\mathbb{R} \otimes \{0, 1\}$ above, with F and f replaced by F_1, F_2 and f_1, f_2. Discernability between these models can be evaluated by the ϕ-divergence $D_\phi(P_1, P_2)$, which is characterized in the next theorem. We see from Equation (40) that the measures μ_i, ν_i considered in this theorem satisfy the relation $\mu_i(\mathbb{R}) + \nu_i(\mathbb{R}) = 1$ ($i \in \{1, 2\}$) so that they cannot be probability measures simultaneously.

THEOREM 3.1 *The ϕ-divergence of P_1 and P_2 is the sum*

$$D_\phi(P_1, P_2) = \mathfrak{D}_\phi(\mu_1, \mu_2) + \mathfrak{D}_\phi(\nu_1, \nu_2), \tag{41}$$

where μ_i, ν_i ($i \in \{1, 2\}$) are the finite measures on $\mathbb{R}$ defined by

$$\mu_i(A) = \int_A (1 - G(x)) f_i(x)\, \mathrm{d}x, \quad \nu_i(A) = \int_A (1 - F_i(x))\, \mathrm{d}G(x), \quad A \subset \mathbb{R} \text{ Borel}.$$

Proof From Equation (36), we get the densities

$$h_i(x,u) = \frac{\mathrm{d}P_i}{\mathrm{d}\lambda}(x,u) = \mathbf{1}_{\{0\}}(u)\,(1-F_i(x)) + \mathbf{1}_{\{1\}}(u)\,(1-G(x))\,f_i(x), \quad i=1,2,$$

with respect to λ given by Equation (35). Therefore, by Equations (6) and (7),

$$\begin{aligned} D_\phi(P_1,P_2) = \mathfrak{D}_\phi(P_1,P_2) &= \int_{\mathbb{R}\times\{0,1\}} h_2\tilde{\phi}\left(\frac{h_1}{h_2}\right)\mathrm{d}\lambda \\ &= \int_{\mathbb{R}} h_2(x,0)\,\tilde{\phi}\left(\frac{h_1(x,0)}{h_2(x,0)}\right)\mathrm{d}G(x) + \int_{\mathbb{R}} h_2(x,1)\,\tilde{\phi}\left(\frac{h_1(x,1)}{h_2(x,1)}\right)\mathrm{d}x \\ &= \int_{\mathbb{R}} (1-F_2(x))\,\tilde{\phi}\left(\frac{1-F_1(x)}{1-F_2(x)}\right)\mathrm{d}G(x) \\ &\quad + \int_{\mathbb{R}} (1-G(x))f_2(x)\tilde{\phi}\left(\frac{(1-G(x))f_1(x)}{(1-G(x))f_2(x)}\right)\mathrm{d}x. \end{aligned}$$

The rest follows from the fact that

$$\mathfrak{D}_\phi(\mu_1,\mu_2) = \int_{\mathbb{R}} f_2(x)\tilde{\phi}\left(\frac{f_1(x)}{f_2(x)}\right)(1-G(x))\,\mathrm{d}x \tag{42}$$

and

$$\mathfrak{D}_\phi(\nu_1,\nu_2) = \int_{\mathbb{R}} (1-F_2(x))\tilde{\phi}\left(\frac{1-F_1(x)}{1-F_2(x)}\right)\mathrm{d}G(x). \tag{43}$$

■

An illustration of this theorem is given by the following.

Example 3.2 We apply the formulas (42) and (43) to the concrete censorings (i)–(iii) of Example 3.1, where f can now be either f_1 or f_2, and thus F is either F_1 or F_2, respectively.

(i) The divergence $\mathfrak{D}_\phi(\mu_1,\mu_2)$ is given by Equation (42), and (43) reduces to

$$\mathfrak{D}_\phi(\nu_1,\nu_2) = \int_{\mathbb{R}} (1-F_2(x))\tilde{\phi}\left(\frac{1-F_1(x)}{1-F_2(x)}\right)g(x)\,\mathrm{d}x.$$

(ii) Due to Equations (41), (42) and (43), one gets for this 'nonrandomly' right-censored context

$$\begin{aligned} \mathfrak{D}_\phi(\mu_1,\mu_2) &= \int_{-\infty}^{y_0} f_2(x)\tilde{\phi}\left(\frac{f_1(x)}{f_2(x)}\right)\mathrm{d}x, \\ \mathfrak{D}_\phi(\nu_1,\nu_2) &= (1-F_2(y_0))\tilde{\phi}\left(\frac{1-F_1(y_0)}{1-F_2(y_0)}\right). \end{aligned}$$

We see that if the censoring sharpens in the sense $y_0 \to -\infty$, then $\mathfrak{D}_\phi(\mu_1,\mu_2) + \mathfrak{D}_\phi(\nu_1,\nu_2) \to 0$. Contrarily, if the censoring relaxes to complete disappearance in the sense $y_0 \to +\infty$, then $\mathfrak{D}_\phi(\nu_1,\nu_2) \to 0$, and both $\mathfrak{D}_\phi(\mu_1,\mu_2)$ and $D_\phi(P_1,P_2)$ tend to the ϕ-divergence $D_\phi(F_1,F_2)$ of alternative distributions F_1 and F_2 of the uncensored observation X.

(iii) In this censoring,

$$\mathfrak{D}_\phi(\mu_1, \mu_2) = \int_{-\infty}^{y_1} f_2(x)\tilde{\phi}\left(\frac{f_1(x)}{f_2(x)}\right) \mathrm{d}x + \frac{1}{2}\int_{y_1}^{y_2} f_2(x)\tilde{\phi}\left(\frac{f_1(x)}{f_2(x)}\right) \mathrm{d}x,$$

$$\mathfrak{D}_\phi(\nu_1, \nu_2) = \frac{1}{2}(1 - F_2(y_1))\,\tilde{\phi}\left(\frac{1 - F_1(y_1)}{1 - F_2(y_1)}\right) + \frac{1}{2}(1 - F_2(y_2))\tilde{\phi}\left(\frac{1 - F_1(y_2)}{1 - F_2(y_2)}\right).$$

We see that if the censoring on the one hand sharpens in the sense $y_1 \to -\infty$ and at the same time relaxes in the sense $y_2 \to +\infty$, then both $\mathfrak{D}_\phi(\mu_1, \mu_2)$ and $D_\phi(P_1, P_2)$ tend to the half of the ϕ-divergence $D_\phi(F_1, F_2)$ of alternative distributions F_1 and F_2 of the uncensored observation X. This agrees with the intuition since in this case one gets with probability $1/2$ either no observation or the uncensored observation X.

4. Information-theoretic applicability

An important role in information theory plays the so-called *Pinsker inequality*

$$I(P, Q) \geq \frac{V(P, Q)^2}{2} \tag{44}$$

between the classical information divergence $I(P, Q)$ of probability measures P, Q and the classical total variation

$$V(P, Q) = \int_{\mathcal{X}} |p - q|\, \mathrm{d}\lambda$$

(cf. (22)) of these measures (for this inequality and its various sharpenings, see Fedotov *et al.* [15]). In Example 2.2, we introduced the nonnegative extension $\mathfrak{I}(\mu, \nu)$ of the classical information divergence $I(P, Q)$ to the case of finite measures $\mu, \nu \in \mathcal{M}$. The corresponding extension of the Pinsker inequality follows easily from Theorem 2.4.

THEOREM 4.1 *For all measures $\mu, \nu \in \mathcal{M}$, the information divergence $\mathfrak{I}(\mu, \nu)$ given by Equation* (14) *satisfies the generalized Pinsker inequality*

$$\frac{\mathfrak{I}(\mu, \nu)}{\mu(\mathcal{X})} \geq \frac{V(P_\mu, P_\nu)^2}{2} + \frac{\mathbf{\Gamma}(\mu, \nu)}{R_{\mu,\nu}},$$

*where $\mathbf{\Gamma}(\mu, \nu)/R_{\mu,\nu}$ is the global information divergence (*34*) normalized by factor $R_{\mu,\nu}$.*

Proof Clear from the consequence (34) of Theorem 2.4 and the original inequality (44). ■

According to Inequality (44) and Equation (33), the term $V(P_\mu, P_\nu)^2/2$ is a lower bound on the normalized local divergence $I(P_\mu, P_\nu) = \mathbf{\Lambda}(\mu, \nu)/R_{\mu,\nu}$ of measures μ, ν on $\mathcal{X}$. This component of the bound is increased by the contribution of the normalized global divergence $\mathbf{\Gamma}(\mu, \nu)/R_{\mu,\nu}$.

Another applicability, which deserves to be mentioned here, is connected with *differential Shannon entropies* $H(p)$ and *Burg entropies* $\tilde{H}(p)$ of random observations X with probability densities p on finite or σ-finite measure spaces $(\mathcal{X}, \mathcal{A}, \lambda)$ (see Cover & Thomas [16]). In terms

of Equation (13), these entropies are defined by the formulas

$$H(p) = -D_1(P, \lambda) = -\int_{\mathcal{X}} p \ln p \, \mathrm{d}\lambda \tag{45}$$

and

$$\tilde{H}(p) = -D_0(P, \lambda) = \int_{\mathcal{X}} \ln p \, \mathrm{d}\lambda \tag{46}$$

for probability densities $p = \mathrm{d}P/\mathrm{d}\lambda$ w.r.t. a dominating finite or σ-finite measure λ such that the corresponding integrals exist in the extended real line $[-\infty, \infty]$. These integrals exist in the extensions $(-\infty, \infty]$ or $[-\infty, \infty)$ of $\mathbb{R}$ if the positive or negative parts of the corresponding integrands are λ-integrable, respectively. For finite measures λ, the Shannon entropy $H(p)$ always exists in $(-\infty, \infty]$. In the remaining cases, sufficient conditions for the existence are needed.

If the dominating measure λ is finite, then one gets the following representations by means of ϕ-divergences $\mathfrak{D}(\cdot, \cdot)$

$$H(p) = -\mathfrak{D}_1(P, \lambda) + \lambda(\mathcal{X}) - 1 \quad \text{(cf. (14))}, \tag{47}$$

and

$$\tilde{H}(p) = -\mathfrak{D}_0(P, \lambda) - \lambda(\mathcal{X}) + 1 \quad \text{(cf. (15))}. \tag{48}$$

Let us mention at least two of the classical concrete applications of the Shannon and Burg entropies.

Example 4.1 Consider an information source which generates a message $\boldsymbol{X} = (X_1, X_2, \ldots)$ formally described as a real stationary zero mean Gaussian process with autocorrelation function $R(k) = EX_jX_{j+k}$ and let λ be the Lebesgue measure on the interval $\mathcal{X} =]-\pi, \pi]$. Suppose for simplicity that the process is normalized so that the spectral measure P defined on $\mathcal{X}$ by the spectral density

$$p(x) = \sum_{k=-\infty}^{\infty} R(k) \, \mathrm{e}^{-ikx}, \quad x \in \mathcal{X},$$

is a probability measure, i.e.

$$\int_{\mathcal{X}} p \, \mathrm{d}\lambda = \int_{-\pi}^{\pi} p(x) \, \mathrm{d}x = 1.$$

According to Kolmogorov [17], the Shannon entropy $H(q_n)$ of the probability density q_n of the first n terms $(X_1, \ldots, X_n)$ of $\boldsymbol{X}$ w.r.t. the Lebesgue measure on $\mathbb{R}^n$ exists and is related to the spectral Burg entropy

$$\tilde{H}(p) = -D_0(P, \lambda) = \int_{-\pi}^{\pi} \ln p(x) \, \mathrm{d}x$$

in the asymptotic sense

$$h(\boldsymbol{X}) \equiv \lim_{n\to\infty} \frac{H(q_n)}{n} = \frac{\ln(2\pi e) + \tilde{H}(p)/2\pi}{2},$$

where $h(\boldsymbol{X})$ is an important information-theoretic parameter called *entropy rate* of the process $\boldsymbol{X}$. The quantity $\sigma_\infty^2 = \exp\{\tilde{H}(p)/2\pi)\}$ is the limit for $n \to \infty$ of the variance in the best estimate of X_n based on the past history $X_1, \ldots, X_{n-1}$, called also *one-step prediction error or gain* in the literature on speech and image coding (see, e.g. Markel and Gray [18], Buzo *et al.* [19]).

Example 4.2 In the literature on image and speech coding, one frequently meets the so-called Itakura–Saito distance. This is a measure of distortion of a stationary signal with spectral measure μ when it is replaced by a stationary signal with spectral measure ν (see [19, Section II] and the references therein). For spectral measures with finite $\mu(\mathcal{X}) = \nu(\mathcal{X}) > 0$ on the spectral space $\mathcal{X} = (-\pi, \pi)$ with continuous densities $p(x)$, $q(x)$, this distance is given as

$$\mathrm{IS}(\mu, \nu) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[-\ln \frac{p(x)}{q(x)} + \frac{p(x)}{q(x)} - 1 \right] \mathrm{d}x$$

(cf. [19, (10)]). We see from Equations (13), (11) and (6) that it differs only by the uniform norming from our zero-order power divergence

$$\mathfrak{D}_0(\mu, \nu) = \int_{-\pi}^{\pi} \left[-\ln \frac{p(x)}{q(x)} + \frac{p(x)}{q(x)} - 1 \right] q(x)\, \mathrm{d}x.$$

Example 4.3 Sturm [20,21] applied the concept of the Shannon entropy in the research on the geometry of metric measure spaces $(\mathcal{X}, \mathcal{A}, \lambda)$. He introduced *relative entropies* $\mathrm{Ent}(P, \lambda)$ of probability measures P on $(\mathcal{X}, \mathcal{A})$ w.r.t. dominating locally finite λ by the formula

$$\mathrm{Ent}(P, \lambda) = \lim_{\epsilon \downarrow 0} \int_{\{p > \epsilon\}} p \ln p \, \mathrm{d}\lambda$$

for $p = \mathrm{d}P/\mathrm{d}\lambda$. It is easy to see that $\mathrm{Ent}(P, \lambda)$ exists if and only if the Shannon entropy $H(p) = -D_1(P, \lambda)$ exists, and then $\mathrm{Ent}(P, \lambda) = -H(p)$.

If in an open neighbourhood N of $\alpha = 1$ the function $M(\alpha) = \int_{\mathcal{X}} p^\alpha \, \mathrm{d}\lambda$ is finite and differentiable in the sense $M(\alpha)' = \int_{\mathcal{X}} (p^\alpha)' \, \mathrm{d}\lambda$, then the Shannon entropy $H(p)$ of Equation (45) can be imbedded as a special limit case for $\alpha = 1$ in the family of *alternative differential power entropies*

$$H_\alpha(p) = \frac{1}{\alpha - 1} \left(1 - \int_{\mathcal{X}} p^\alpha \, \mathrm{d}\lambda \right), \qquad \alpha \in N. \tag{49}$$

If in addition λ is finite, then one gets for $\alpha \in N \backslash \{0, 1\}$ the representation

$$H_\alpha(p) = -\alpha \mathfrak{D}_\alpha(P, \lambda) + \lambda(\mathcal{X}) - 1. \tag{50}$$

Let us give an example which demonstrates that the entropies from the class (49) have been over a century successfully applied in various areas of science.

Example 4.4 If the observation space is finite, $\mathcal{X} = \{1, \ldots, n\}$, and λ is the counting measure on $\mathcal{X}$ then the density p of a probability measure P on $\mathcal{X}$ reduces to a discrete distribution $p = (p_1, \ldots, p_n)$ and (49) reduces to the formula

$$H_\alpha(p) = \frac{1}{\alpha - 1} \left(1 - \sum_{i=1}^{n} p_i^\alpha \right), \quad \alpha \in N = \mathbb{R}, \tag{51}$$

where $H_1(p)$ stands for the limit

$$H_1(p) = \lim_{\alpha \to 1} H_\alpha(p) = -\sum_{i=1}^{n} p_i \ln p_i. \tag{52}$$

The quadratic entropy

$$H_2(p) = 1 - \sum_{i=1}^{n} p_i^2 \tag{53}$$

has been used as an econometric measure of uniformity of income ever since Dalton [22,23] and as a measure of diversity in biology, ecology and sociology ever since Gini [24] and Simpson [25]. The term 'quadratic entropy' was coined by Vajda [26] who proposed $H_2(p)$ as an approximation to the Bayes error $e(p) = 1 - \max_i\ p_i$. In this role, the quadratic entropy is frequently applied in pattern recognition, cf., e.g. Devroye *et al.* [27]. For the applications of the 'Gini–Simpson index of diversity' $H_2(p)$ in genetics and medicine, see, e.g. Zvárová and Vajda [28]. The whole subclass of the entropies (51) of positive powers $\alpha > 0$ was introduced axiomatically by Havrda and Charvát [29], but this subclass is just an exponential rescaling of the class

$$R_\alpha(p) = \frac{1}{1-\alpha} \ln \sum_{i=1}^{n} p_i^\alpha \tag{54}$$

introduced axiomatically earlier by Rényi [30]. The entropies (51) of the nonpositive powers $\alpha \leq 0$ were introduced recently by Vajda and Zvárová [31] on pragmatic grounds: they proved that $H_{2-\beta}(p)$ are the generalized informations of positive orders $\beta > 0$ obtained by a direct observation of outputs of a discrete information source $(\mathcal{X}, p)$.

If we replace Equation (49) by

$$\tilde{H}_\alpha(p) = \frac{1}{\alpha}\left(1 - \int_{\mathcal{X}} p^\alpha \, \mathrm{d}\lambda\right), \quad \alpha \in N,$$

for a neighbourhood N of $\alpha = 0$, and correspondingly modify the assumptions concerning N before Equation (49), then we obtain a family of power extensions of the Burg entropy $\tilde{H}(p)$ of (46) which is now obtained as the limit for $\alpha \to 0$.

To demonstrate another, information-theoretic applicability of power divergences of finite measures, consider two real stationary zero mean Gaussian processes $\boldsymbol{X} = (X_1, X_2, \ldots)$, $\boldsymbol{Y} = (Y_1, Y_2, \ldots)$, with autocorrelation functions

$$R(k) = EX_jX_{j+k}, \quad S(k) = EY_jY_{j+k}$$

and spectral densities

$$p(x) = \sum_{k=-\infty}^{\infty} R(k)\, \mathrm{e}^{-ikx}, \quad q(x) = \sum_{k=-\infty}^{\infty} S(k)\, \mathrm{e}^{-ikx}$$

on the interval $\mathcal{X} =]-\pi, \pi]$. Denote by λ the finite measure on the Borel subsets of $\mathcal{X}$ with the constant density $1/4\pi$ w.r.t. the Lebesgue measure on $\mathcal{X}$, and let μ and ν be the finite measures on the Borel subsets of $\mathcal{X}$ with the Radon–Nikodym densities p and q w.r.t. λ. Then the Hellinger divergence

$$\begin{aligned}\mathfrak{D}_{1/2}(\mu, \nu) &= 2\int_{\mathcal{X}} \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mathrm{d}\lambda(x) \quad \text{(cf. (24))} \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mathrm{d}x\end{aligned}$$

is a generalized Ornstein distance of the processes $\boldsymbol{X}$, $\boldsymbol{Y}$ with important applications in the information theory and also in system identification and modelling (see Gray *et al.* [32, Section 4]).

An analogue of the generalized Pinsker inequality for $\mathfrak{D}_{1/2}(\mu, \nu)$ instead of $\mathfrak{I}(\mu, \nu) = \mathfrak{D}_1(\mu, \nu)$ is given in the next theorem where it is assumed for simplicity that $\mu(\mathcal{X}) = \nu(\mathcal{X})$. The extension to $\mu(\mathcal{X}) \neq \nu(\mathcal{X})$ is obvious.

THEOREM 4.2 *If μ, ν are finite measures with the property $\mu(\mathcal{X}) = \nu(\mathcal{X})$, then the Hellinger divergence* (24) *satisfies the inequality*

$$\mathfrak{D}_{1/2}(\mu, \nu) \geq 4\mu(\mathcal{X}) \left\{ 1 - \left(1 - \frac{V(P_\mu, P_\nu)^2}{4} \right)^{1/2} \right\},$$

where P_μ, P_ν are the normalized versions of μ, ν given by Equation (30).

Proof By (2.38) in [2],

$$D_{1/2}(P_\mu, P_\nu) \geq 4 \left\{ 1 - \left(1 - \frac{V(P_\mu, P_\nu)^2}{4} \right)^{1/2} \right\},$$

and by Equations (31) and (32), the divergence (24) satisfies the relation

$$\mathfrak{D}_{1/2}(\mu, \nu) = \mu(\mathcal{X}) D_{1/2}(P_\mu, P_\nu). \qquad \blacksquare$$

Acknowledgements

We are very grateful to two unknown referees for their valuable suggestions and comments. We also thank for the partial support by the MŠMT grant 1M0572 and the GAČR grant 102/07/1131.

References

[1] I. Csiszár, *Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten*, Publ. Math. Inst. Hung. Acad. Sci. Ser. A 8 (1963), pp. 85–108.
[2] F. Liese and I. Vajda, *Convex Statistical Distances*, Teubner, Leipzig, 1987.
[3] F. Liese and I. Vajda, *On divergences and informations in statistics and information theory*, IEEE Trans. Inform. Theory 52(10) (2006), pp. 4394–4412.
[4] D. Morales, L. Pardo, M.C. Pardo, and I. Vajda, *Rényi statistics for testing composite hypotheses in general exponential models*, Statistics 38(2) (2004), pp. 133–147.
[5] D. Morales, L. Pardo, and I. Vajda, *Rényi statistics in directed families of exponential experiments*, Statistics 34(1) (2000), pp. 151–174.
[6] W. Stummer and I. Vajda, *Optimal statistical decisions about some alternative financial models*, J. Econom. 137 (2007), pp. 441–471.
[7] W. Stummer, *On a statistical information measure for a generalized Samuelson–Black–Scholes model*, Statist. Decisions 19 (2001), pp. 289–314.
[8] W. Stummer, *Exponentials, Diffusions, Finance, Entropy and Information*, Shaker, Aachen, 2004.
[9] R. Miller, *Survival Analysis*, Wiley, New York, 1981.
[10] M. Hollander, F. Proschan, and J. Sconing, *Information, censoring, and dependence*, in Topics in Statistical Dependence H.W. Block, A.R. Sampson, and T.H. Savits, eds., IMS, Hayward, 1990, pp. 257–268.
[11] W. Stute, *Strong consistency of the MLE under random censoring*, Metrika 39 (1992), pp. 257–267.
[12] W. Stute, *The sequential probability ratio test under random censorship*, Metrika 44 (1996), pp. 1–8.
[13] Ch. Tsairidis, K. Ferentinos, and T. Papaioannou, *Information and random censoring*, Inform. Comput. Sci. 92 (1996), pp. 159–174.
[14] Ch. Tsairidis, K. Zografos, K. Ferentinos, and T. Papaioannou, *Information in quantal response data and random censoring*, Ann. Inst. Statist. Math. 53(3) (2001), pp. 528–542.
[15] A.A. Fedotov, P. Harremoës, and F. Topsøe, *Refinements of Pinsker's inequality*, IEEE Trans. Inform. Theory 49(6) (2003), pp. 1491–1498.
[16] T.M. Cover and J.B. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.
[17] A.N. Kolmogorov, *On the Shannon theory of information transmission in the case of continuous signals*, IRE Trans. Inform. Theory 2 (1956), pp. 102–108.
[18] J.D. Markel and A.H. Gray, Jr., *Linear Prediction of Speech*, Springer, Berlin, 1976.
[19] A. Buzo, A.H. Gray, Jr., R.M. Gray, and J.D. Markel, *Speech coding based upon vector quantization*, IEEE Trans. Acoust Speech Signal Process. 28(5) (1980), pp. 562–574.
[20] K.-Th. Sturm, *On the geometry of metric measure spaces I*, Acta Math. 196 (2006), pp. 65–131.
[21] K.-Th. Sturm, *On the geometry of metric measure spaces II*, Acta Math. 196 (2006), pp. 133–177.
[22] H. Dalton, *The measurement of the inequality of incomes*, Econ. J. 30 (1920), pp. 348–361.

[23] H. Dalton, *Some Aspects of the Inequality of Incomes in Modern Communities*, 2nd ed. Routledge and Kegan Paul, London, 1925.
[24] C. Gini, *Variabilità e Mutabilità*, Studo Economico-Giuridici della R. Univ. di Cagliari 3(2) (1912), p. 80.
[25] E.H. Simpson, *Measurement of diversity*, Nature 163 (1949), p. 688.
[26] I. Vajda, *Bounds on the minimal error probability and checking a finite or countable number of hypotheses*, Inform. Transmis. Probl. 4 (1968), pp. 9–17.
[27] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, Berlin, 1996.
[28] J. Zvárová and I. Vajda, *On genetic information, diversity and distance*, Methods Inform. Med. 2 (2006), pp. 173–179.
[29] J. Havrda and F. Charvát, *Concept of structural a-entropy*, Kybernetika 3 (1967), pp. 30–35.
[30] A. Rényi, *On measures of entropy and information*, in Proceedings of Fourth Berkeley Symposium on Probability and Statistics, Vol. 1, University of California Press, Berkeley, 1961, pp. 547–561.
[31] I. Vajda and J. Zvárová, *On generalized entropies, Bayesian decisions and statistical diversity*, Kybernetika 43(5) (2007), pp. 675–696.
[32] R.M. Gray, D.L. Neuhoff, and P.C. Shields, *A generalization of Ornstein's $\overline{d}$ distance with applications to information theory*, Ann. Probab. 3(2) (1975), pp. 315–328.