Metrika (2016) 79:303–333
DOI 10.1007/s00184-015-0556-6

On local divergences between two probability measures

G. Avlogiaris[1] · A. Micheas[2] · K. Zografos[1]

Received: 11 July 2014 / Published online: 24 July 2015

Abstract A broad class of local divergences between two probability measures or between the respective probability distributions is proposed in this paper. The introduced local divergences are based on the classic Csiszár ϕ-divergence and they provide with a pseudo-distance between two distributions on a specific area of their common domain. The range of values of the introduced class of local divergences is derived and explicit expressions of the proposed local divergences are also derived when the underlined distributions are members of the exponential family of distributions or they are described by multivariate normal models. An application is presented to illustrate the behavior of local divergences.

Keywords ϕ-divergence · Kullback–Leibler divergence · Cressie and Read power divergence · Local divergence · Exponential family

Mathematics Subject Classification 62B10 · 62F99

1 Introduction

The concept of divergence is of fundamental importance, not only in mathematics but in almost all branches of science and engineering. This concept has also a prominent role in probability theory and mathematical statistics. The Kolmogorov–Smirnov test is based, for instance, on a divergence measure between the empirical distribution function and the respective function which is specified by the null hypothesis. The

✉ K. Zografos
kzograf@uoi.gr

[1] Department of Mathematics, University of Ioannina, 451 10 Ioannina, Greece

[2] Department of Statistics, University of Missouri-Columbia, Columbia, USA

classical chi-square goodness-of-fit test is based on a divergence measure between the theoretic probabilities and the expected ones. Many other statistical procedures base their origins on a divergence measure between probability distributions.

The most important attempt to define a broad class of divergence measures between two probability measures or between the respective Radon-Nikodym derivatives was made by Csiszár (1963, 1967) and independently by Ali and Silvey (1966). Following these authors, if P and Q are two probability measures on the measurable space $(\mathcal{X}, \mathcal{A})$ and μ is a σ-finite measure on the same measurable space, such that $P \ll \mu$ and $Q \ll \mu$, then for p and q the respective Radon-Nikodym derivatives, $p = \frac{dP}{d\mu}$ and $q = \frac{dQ}{d\mu}$, a broad class of divergence measures between P and Q, or between p and q, is defined by the following integral,

$$D_\phi(P, Q) = D_\phi(p, q) = \int\limits_{\mathcal{X}} \phi\left(\frac{dP}{dQ}\right) dQ = \int\limits_{\mathcal{X}} q(x)\phi\left(\frac{p(x)}{q(x)}\right) d\mu(x), \qquad (1)$$

where ϕ is a real valued convex function, satisfying appropriate conditions (cf. Csiszár 1967). These conditions will be discussed and extended later on.

An important property of $D_\phi(P, Q)$ is that if ϕ is strictly convex at 1 with $\phi(1) = 0$, then (cf. Pardo 2006, p. 9),

$$D_\phi(p, q) = 0 \text{ if and only if } p(x) = q(x), \text{ a.e. } x \in \mathcal{X}. \qquad (2)$$

This is the reason why $D_\phi(P, Q)$ has been established in the literature as a measure of divergence between the probability measures P and Q, or between the respective densities p and q, and it is referred to as Csiszár ϕ-divergence, or simply, as ϕ-divergence. As defined, $D_\phi(P, Q)$ is not symmetric, but can be expressed as a symmetric measure by taking $D_{\tilde{\phi}}(P, Q) = D_{\tilde{\phi}}(Q, P) = D_\phi(P, Q) + D_\phi(Q, P)$, for the convex functions $\tilde{\phi}(u) = \phi(u) + u\phi(\frac{1}{u})$, $u > 0$ (cf. Liese and Vajda 1987, p. 14). Moreover, $D_\phi(P, Q)$ is not a distance in the usual sense of a metric since it does not satisfy in general the triangular inequality. We can think of divergence measures as distances, in the same way we treat a loss function in a decision theoretic problem; It simply tells us if two probability measures are the same or not and the closer the value of D_ϕ to 0, the closer P and Q are.

Following Csiszár (1967, p. 301), ϕ-divergence extends, in essence, the "information for discrimination" or I-divergence, introduced by Kullback and Leibler (1951) and the "information gain" or I-divergence of order α, introduced by Rényi (1960). The Kullback–Leibler divergence measure is obtained when the convex function ϕ is of the form $\phi(x) = x \log x$ or $\phi(x) = x \log x - x + 1, x > 0$, and Rényi's divergence is obtained, as a function of Csiszár ϕ-divergence, for $\phi(x) = sgn(\alpha - 1)x^\alpha$, $x > 0$ and $\alpha > 0$. Other choices of the convex function ϕ lead to important measures of divergence (cf. Pardo 2006, p. 6, where measures of divergence are tabulated for specific choices of ϕ). Among these measures of divergence the Cressie and Read or λ-divergence, introduced independently by Cressie and Read (1984) and Liese and Vajda (1987), plays a prominent role in the development of goodness of fit and λ-divergence tests. It is obtained from (1), for $\phi(x) = \phi_\lambda(x) = \frac{x^{\lambda+1} - x - \lambda(x-1)}{\lambda(\lambda+1)}$, $\lambda \neq 0, -1$. Kullback–Leibler divergence, $D_0(f, g) = \int_{\mathcal{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right) d\mu(x)$, is the limiting case of Cressie and Read λ -divergence, as $\lambda \to 0$, that is, $\lim_{\lambda \to 0} D_{\phi_\lambda}(p, q) = D_0(p, q)$.

After Csiszár (1963, 1967) pioneering work in the subject, a plethora of papers and books have been published. Some of them are concentrated on the characterization and the study of the properties of ϕ-divergence, while others on generalizations of ϕ-divergence. A large portion of this literature is concerned with applications of ϕ-divergence to formulate and solve a great variety of problems in probability and statistics and in almost every branch of science and engineering. The books and monographs by Kullback (1959), Csiszár and Korner (1981), Liese and Vajda (1987), Vajda (1989, 1995), Pardo (2006) and Basu et al. (2011) and the review papers by Papaioannou (1986, 2001), Ullah (1996), Soofi (2000), Ebrahimi et al. (2010) and the references in these works constitute a basis of the existing literature on ϕ-divergence measures.

The ϕ-divergence, as it is defined by (1), quantifies the difference between the arguments P and Q, or p and q, in a domain $\mathcal{X}$. However, there are situations in practice where the interest is focused on the differences between two probability distributions in a subset of the domain $\mathcal{X}$. For example, suppose that a researcher is interested in inferring about the homogeneity or similarity of two populations, regarding a joint characteristic of their members. Consider a population of men and women and suppose that the joint characteristic under study is the level of blood cholesterol of the members of both populations. The joint characteristic of the two populations is described by two probabilistic models, one for each population, and the homogeneity or similarity of the two populations can be quantified by a divergence measure between the two probabilistic models (densities) which describe the populations, with values close to zero indicating equality of the two densities.

However, a divergence measure, resulting from (1) for a specific choice of ϕ, quantifies the similarity of the populations in the whole domain. Hence, application of (1) provides a misconception about the differences of the two populations if the interest is focused on similarity in a subset of the whole domain, say, if the researcher is focused in the investigation of whether the populations of men and women exhibit the same behavior for high or low levels of blood cholesterol. A first solution to this problem can be achieved if we use a measure, based on (1), by integrating over the desired subset of $\mathcal{X}$, instead of the entire space $\mathcal{X}$. However, this option leads to intractable measures of divergence and, mainly, based on Lemma 1.1 of Csiszár (1967), they lead to measures of divergence that violate (2), which is essential in the characterization of (1) as a divergence between probability measures. A second solution could be to replace the probability distributions in (1), by the respective truncated distributions, over the desired subset of $\mathcal{X}$. However, this second approach is based on the divergence of the truncated models, which are not necessarily the proper models to use to describe the data under consideration. Consequently, a measure of divergence should be defined that helps overcome these problems and in addition provide an indication about the similarity of the two probability distributions in a subset of their common domain.

Based on the above discussion, the main aim of this paper is to introduce a measure of the local divergence between two probability measures or probability distributions and to study its range of values. Ergo, in the next Sect. 2 local ϕ-divergences are introduced and some numerical examples that illustrate their behavior are given. The range of values of the introduced divergences will be investigated in this section. Section 3 concentrates on explicit forms for a particular case of the local ϕ -divergence between members of the exponential family of distributions. The case of multivariate normal distributions will also be considered. An application is presented in Sect. 4, in order to illustrate the usefulness of the methodology introduced. Section 5 presents some concluding remarks.

2 Local ϕ-divergence and its properties

The aim of this section is to present a measure of local divergence between two probability measures or the respective probability distributions. This measure has its origins on Csiszár ϕ-divergence, which is defined by (1). The local ϕ-divergence will be introduced in the next subsection, while the subsequent subsection provides with the range of values of the introduced local divergences. Special cases of local ϕ-divergence have been studied in the past, see for example McElroy and Holan (2009) for an application of the Kullback–Leibler divergence.

2.1 Local ϕ-divergence

Following Csiszár (1967, p. 299) and Pardo (2006, p. 5), consider the class Φ^* of all real convex functions ϕ defined on the interval $[0, \infty)$, such that $\phi(1) = 0, 0\phi\left(\frac{0}{0}\right) = 0$ and $0\phi\left(\frac{u}{0}\right) = u \lim_{v \to +\infty} \frac{\phi(v)}{v}$, where the last two conditions are necessary in order to avoid meaningless expressions in what follows. Moreover, it is assumed that ϕ is strictly convex at 1. It should be noted, at this point, that all the convex functions ϕ that lead to important particular cases of Csiszár ϕ-divergences, like Kullback and Leibler (1951) divergence, Kagan (1963) divergence ($\phi(u) = (u-1)^2, u > 0$), Vajda (1973) divergence ($\phi(u) = |1-u|^\alpha, u > 0, \alpha \geq 1$), Cressie and Read (1984) λ-power divergence, etc. satisfy all the above conditions.

Motivated by Csiszár ϕ-divergence, given by (1), a measure of local divergence between two probability measures P and Q, or between the respective Radon-Nikodym derivatives p and q, can be defined by means of (1), if an additional function, say $A(\cdot, \omega)$, would be inserted in Csiszár ϕ-divergence in order to shift the mass of the integral (1) in the desired subset of $\mathcal{X}$. The function $A(\cdot, \omega)$ plays the role of a kernel and in complete analogy with Csiszár divergence (1), a measure of local divergence can be defined as follows,

$$D_\phi^A(P, Q) = \int_{\mathcal{X}} A(x, \omega)\phi\left(\frac{dP}{dQ}\right) dQ = \int_{\mathcal{X}} A(x, \omega)q(x)\phi\left(\frac{p(x)}{q(x)}\right) d\mu(x).$$

Notice that if $A(x, \omega) = 1$, then $D_\phi^1(P, Q) = D_\phi(P, Q)$. The introduction of the kernel $A(x, \omega)$ weighs differently the distance between P and Q providing the ability to focus on specific areas of the domain $\mathcal{X}$ that may be of particular interest. In practice, the kernel $A(\cdot, \cdot)$ can be thought of as a window that can be calibrated to highlight specific features of P and Q and how they differ.

In what follows and in order to avoid problems related to the existence of the above integral, we will restrict on functions $A(x, \omega)$ which are related to a probability measure, say R, in the same measurable space $(\mathcal{X}, \mathcal{A})$ and in particular, the function $A(x, \omega)$ will be considered to be the Radon-Nikodym derivative of R with respect to μ, with μ a σ-finite measure on $(\mathcal{X}, \mathcal{A})$. In this setting the definition of the local ϕ-divergence is formulated as follows.

Definition 1 Let P, Q and R three probability measures on the measurable space $(\mathcal{X}, \mathcal{A})$, dominated by a σ-finite measure μ which is defined on the same measurable space. Let p, q and r denote the respective Radon-Nikodym derivatives $p = \frac{dP}{d\mu}$, $q = \frac{dQ}{d\mu}$ and $r = \frac{dR}{d\mu}$ and ϕ a convex function belonging to the class of convex functions Φ^*, defined above. Then, the local ϕ-divergence between P and Q, driven by R, is defined by

$$D_\phi^R(P, Q) = \int\limits_{\mathcal{X}} \frac{dR}{d\mu} \phi\left(\frac{dP}{dQ}\right) dQ = \int\limits_{\mathcal{X}} r(x) q(x) \phi\left(\frac{p(x)}{q(x)}\right) d\mu(x). \tag{3}$$

Remark 1 (i) The definition can be modified in such a way as to be valid on a parametric family of probability measures. Consider the measurable space $(\mathcal{X}, \mathcal{A})$ and let $\{P_\theta : \theta \in \Theta \subseteq R^M\}$ be a parametric family of probability measures on $(\mathcal{X}, \mathcal{A})$. Let μ be a σ-finite measure on the same measurable space, such that $P_\theta \ll \mu$, for $\theta \in \Theta$. Denote by $f_\theta = \frac{dP_\theta}{d\mu}$, the Radon-Nikodym derivative, and consider a convex function ϕ belonging to the class of convex functions Φ^*. Further consider a probability measure $P_\omega \ll \mu$, $\omega \in \Theta$, on $(\mathcal{X}, \mathcal{A})$, with Radon-Nikodym derivative $f_\omega = \frac{dP_\omega}{d\mu}$, for $\omega \in \Theta$. The local ϕ-divergence between two members of the class $\{P_\theta : \theta \in \Theta \subseteq R^M\}$, P_{θ_1} and P_{θ_2}, or between the respective Radon-Nikodym derivatives f_{θ_1} and f_{θ_2}, with kernel the function f_ω, is defined as follows,

$$D_\phi^\omega(\theta_1, \theta_2) = \int\limits_{\mathcal{X}} \frac{dP_\omega}{d\mu} \phi\left(\frac{dP_{\theta_1}}{dP_{\theta_2}}\right) dP_{\theta_2} = \int\limits_{\mathcal{X}} f_\omega(x) f_{\theta_2}(x) \phi\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right) d\mu(x). \tag{4}$$

The local ϕ-divergence, as defined is a measure of divergence between two members of the above family, and is governed by another measure of the family that determines the weights and the area over which the divergence is calculated. In the latter definition, f_ω depends on a parameter ω that drives the window over which the integral is computed. Calculation of the measure (4) in a closed form is accomplished easier when the driving measure P_ω or the corresponding density f_ω is in the same parametric family of probability measures $\{P_\theta : \theta \in \Theta \subseteq R^M\}$, but that need not be the case in practice. Consequently, the distribution f_ω can be chosen in such a way in order to smooth or exemplify certain features of the area over which the integral is calculated.

(ii) If $\mathcal{X}$ is finite (or countable), $\mathcal{X}$ =$\{1, 2, ..., n\}$, and P, Q, R are represented by the discrete probability distributions $\mathbf{p} = (p_1, ..., p_n)$, $\mathbf{q} = (q_1, ..., q_n)$ and $\mathbf{r} =$

$(r_1, ..., r_n)$, respectively, then the local ϕ-divergence between $\mathbf{p}$ and $\mathbf{q}$, driven by $\mathbf{r}$, is defined, in view of (3), by

$$D_{\phi}^{\mathbf{r}}(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} r_i q_i \phi\left(\frac{p_i}{q_i}\right).$$

This last measure is known in the literature as the weighted ϕ -divergence and it has been studied in the papers by Landaburu and Pardo (2000, 2003), Landaburu et al. (2005) and the references therein.

(iii) An extension of the local ϕ-divergence, defined by (3) or (4), can be obtained by a quite similar argument as that of Pardo (2006, p. 8). More precisely, if h is a differentiable increasing real function, then the local (h, ϕ)-divergence is defined by $D_{h,\phi}^{\omega}(P, Q) = h\left(D_{\phi}^{\omega}(P, Q)\right)$. This last measure allows us to define more general measures of local divergence, for several choices of the functions h and ϕ. However, the main reason for the above transformation of $D_{\phi}^{\omega}(P, Q)$ is that it allows us to obtain Rényi's local divergence by means of the local ϕ -divergence.

(iv) In general, we cannot obtain $D_{\phi}(P, Q)$ from $D_{\phi}^{R}(P, Q)$, unless R is, for example, a uniform measure over $\mathcal{X}$, in which case $D_{\phi}^{R}(P, Q)$ is a multiple of $D_{\phi}(P, Q)$. Notice that $D_{\phi}^{R}(P, Q) = D_{\phi}(P, Q)$ when $E_q\left[(1 - r(X))\phi\left(\frac{p(X)}{q(X)}\right)\right] = 0$.

The local ϕ-divergence, defined by (3) or (4) above, is quite similar to the one defined by (1). The only difference is the distribution function r or f_{ω} that enters into the expression of the classic Csiszár ϕ-divergence and in particular the additional parameter $\omega \in \Theta$. The role of the parameter ω is decisive in the above definition, and exactly this role will be investigated in the following examples. In the first example normal distributions will be used in order to investigate how definition (4) actually quantifies the divergence between two normal models in a subset of their domain. In this example, we clarify the role of the parameter ω in the definition of D_{ϕ}^{ω}.

Example 1 **Normal Distributions.** Let $P_{\theta}, \theta \in \Theta = \left\{(\mu, \sigma^2) : \mu, \sigma^2 \in R,\ \sigma^2 > 0\right\}$ be the univariate normal distribution. For three cases of the parameter $\theta, \theta_1 = (\mu_1, \sigma_1^2)$, $\theta_2 = (\mu_2, \sigma_2^2)$ and $\omega = (\mu, \sigma^2)$, denote by f_{θ_1}, f_{θ_2} and f_{ω} the respective univariate normal densities. Consider Cressie and Read (1984) λ-power divergence and more specifically its local version, as it is obtained from (4), for $\phi(u) = \phi_{\lambda}(u) = \frac{u^{\lambda+1} - u - \lambda(u-1)}{\lambda(\lambda+1)}$, $\lambda \neq 0, -1$. The explicit form of this local divergence $D_{\phi_{\lambda}}^{\omega}(\theta_1, \theta_2)$ between f_{θ_1} and f_{θ_2}, driven by f_{ω}, is given by the following expression,

$$D_{\phi_{\lambda}}^{\omega}(\theta_1, \theta_2) = \frac{1}{\lambda(\lambda+1)}\left\{K_{\lambda,\omega}(\theta_1, \theta_2) - (\lambda+1)E_{f_{\theta_1}}\left[f_{\omega}(X)\right] + \lambda E_{f_{\theta_2}}\left[f_{\omega}(X)\right]\right\}, \tag{5}$$

where

$$\begin{aligned} K_{\lambda,\omega}(\theta_1, \theta_2) &= \int_{\mathcal{X}} f_{\omega}(x) f_{\theta_1}^{\lambda+1}(x) f_{\theta_2}^{-\lambda}(x) d\mu(x) \\ &= \frac{(2\pi)^{-1/2}\sigma_1^{-\lambda}\sigma_2^{\lambda+1}}{\left(\sigma_1^2\sigma_2^2 + (\lambda+1)\sigma^2\sigma_2^2 - \lambda\sigma^2\sigma_1^2\right)^{1/2}} \exp\left\{-\frac{1}{2}\left(B_1 + B_2\right)\right\}, \end{aligned}$$

with

$$B_1 = -\frac{\lambda(\lambda+1)\,(\mu_1-\mu_2)^2}{(\lambda+1)\sigma_2^2-\lambda\sigma_1^2}\,,\ B_2 = (\mu-\widetilde{\mu})^2\,\frac{(\lambda+1)\sigma_2^2-\lambda\sigma_1^2}{\sigma_1^2\sigma_2^2+(\lambda+1)\sigma^2\sigma_2^2-\lambda\sigma^2\sigma_1^2},$$

and

$$\widetilde{\mu} = \frac{(\lambda+1)\mu_1\sigma_2^2-\lambda\mu_2\sigma_1^2}{(\lambda+1)\sigma_2^2-\lambda\sigma_1^2}.$$

Moreover,

$$E_{f_{\theta_i}}\left[f_\omega(X)\right] = \left(2\pi(\sigma^2+\sigma_i^2)\right)^{-1/2}\exp\left\{-\frac{(\mu-\mu_i)^2}{2(\sigma^2+\sigma_i^2)}\right\},\ i=1,2.$$

The aforementioned expressions can be obtained as particular cases of the local ϕ-divergence between members of the exponential family of distributions, which will be obtained in a subsequent section. Using (5), we present $D^{\omega}_{\phi_{2/3}}(\theta_1,\theta_2)$, $D^{\omega}_{\phi_{2/3}}(\theta_2,\theta_1)$ and the symmetric version $D^{\omega}_{\phi_{2/3}}(\theta_1,\theta_2)$ $+D^{\omega}_{\phi_{2/3}}(\theta_2,\theta_1)$, for $\theta_1=(0,1)$, $\theta_2=(0,2)$ and several values of the parameter $\omega=(\mu,\sigma^2)$, in Table 1. We concentrate on the value $\lambda=2/3$ because this choice for the power λ is considered ideal in many statistical applications of the classic Cressie and Read power divergence, which is obtained from (1). Table 1 also includes the classic Cressie and Read power divergence $D_{\phi_{2/3}}(\theta_1,\theta_2)=\int_{\mathcal{X}} f_{\theta_2}(x)\phi_{2/3}\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right)d\mu(x)$ and values of the integral

$$I_{i,j} = \int\limits_{\mathcal{X}} I_A(x)f_{\theta_j}(x)\phi_{2/3}\left(\frac{f_{\theta_i}(x)}{f_{\theta_j}(x)}\right)d\mu(x),\ i,j=1,2,\ i\neq j,$$

Table 1 Values of $D_{\phi_{2/3}}(\theta_1,\theta_2), D_{\phi_{2/3}}(\theta_2,\theta_1)$, $D^{\omega}_{\phi_{2/3}}(\theta_1,\theta_2), D^{\omega}_{\phi_{2/3}}(\theta_2,\theta_1)$, $D^{\omega}_{\phi_{2/3}}(\theta_1,\theta_2)$ + $D^{\omega}_{\phi_{2/3}}(\theta_2,\theta_1)$, $I_{1,2}$, $I_{2,1}$ and $I_{1,2}+I_{2,1}$ for normal distributions with parameters $\theta_1=(\mu_1,\sigma_1^2)=(0,1)$, $\theta_2=(\mu_2,\sigma_2^2)=(0,2)$ and several values of $\omega=(\mu,\sigma^2)$

$\omega=(\mu,\sigma^2)$	$D^{\omega}_{\phi_{2/3}}(\theta_1,\theta_2), D^{\omega}_{\phi_{2/3}}(\theta_2,\theta_1)$, $D^{\omega}_{\phi_{2/3}}(\theta_1,\theta_2)+D^{\omega}_{\phi_{2/3}}(\theta_2,\theta_1)$ $\theta_1=(0,1),\ \theta_2=(0,2)$	$A\subseteq X$ $A=(\alpha,\beta)$	$I_{1,2}, I_{2,1}, I_{1,2}+I_{2,1}$
(0, 0.1)	0.0195, 0.0151, 0.0346	(−0.5, 0.5)	0.0199, 0.0154, 0.0353
(1, 0.1)	0.0040, 0.0035, 0.0075	(0.5, 1.5)	0.0037, 0.0032, 0.0069
(2, 0.1)	0.0115, 0.0217, 0.0332	(1.5, 2.5)	0.0116, 0.0216, 0.0332
(3, 0.1)	0.0120, 0.0509, 0.0629	(2.5, 3.5)	0.0121, 0.0513, 0.0634
(4, 0.1)	0.0033, 0.0448, 0.0481	(3.5, 4.5)	0.0032, 0.0450, 0.0482
(5, 0.1)	0.0004, 0.0247, 0.0251	(4.5, 5.5)	0.0004, 0.0246, 0.0250
(6, 0.1)	0.0000, 0.0104, 0.0104	(5.5, 6.5)	0.0000, 0.0103, 0.0103
$D_{\phi_{2/3}}(\theta_1,\theta_2)=0.082$, $D_{\phi_{2/3}}(\theta_2,\theta_1)=0.3373$		$(-3\sigma_2,3\sigma_2)$	0.0804, 0.2383

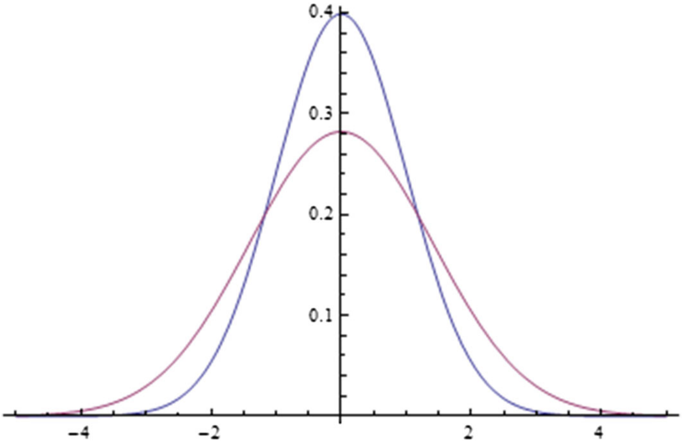

Fig. 1 Plot for normal distributions with parameters $\theta_1 = (0, 1)$ and $\theta_2 = (0, 2)$

which is, in essence, Csiszár classic ϕ-divergence, restricted to the set $A \subseteq \mathcal{X}$. Based on this table, the Cressie and Read λ-divergence between two univariate normal models, $N(0, 1)$ and $N(0, 2)$, is equal to $D_{\phi_{2/3}}(\theta_1, \theta_2) = 0.082$ on the whole domain $\mathcal{X} = R$. Its value is significantly reduced if the interest is focused on specific subsets $A = (\alpha, \beta)$ of $\mathcal{X}$, as it is quantified by the integral $I_{i,j}$. This exemplifies the role of the density f_ω, in (3), and more specifically the role of the parameter ω which adjusts the subset of $\mathcal{X}$ over which the divergence between the normal models, $N(0, 1)$ and $N(0, 2)$ is evaluated. Notice that when we focus on the tails (outside the interval $[-5, 5]$) of the distributions the more similar the two densities become, as shown in Fig. 1. The choice of λ is of great importance, and some measures will not be able to adequately capture the divergence between two distributions.

The next example examines the behavior of the local ϕ-divergence when the symmetric normal models are replaced by skewed models and more specifically by skew normal models.

Example 2 **Skew Normal Distributions.** Consider the standard skew-normal model with parameter α and density $2\phi(x)\Phi(\alpha x)$, where ϕ and Φ are used to denote the p.d.f. and the c.d.f. of the standard normal distribution. The next table presents values of the Cressie and Read (1984) λ-power divergence $D^{\omega}_{\phi_{2/3}}(\alpha_1, \alpha_2)$, $D^{\omega}_{\phi_{2/3}}(a_2, a_1)$ and the symmetric version $D^{\omega}_{\phi_{2/3}}(a_1, a_2) + D^{\omega}_{\phi_{2/3}}(a_2, a_1)$, between two skew-normal models with parameters $\alpha_1 = 2$ and $\alpha_2 = -1$. The density f_ω, of the local λ-power divergence, defined by (3) and (4) for $\phi(u) = \phi_\lambda(u) = \frac{u^{\lambda+1} - u - \lambda(u-1)}{\lambda(\lambda+1)}$, $\lambda \neq 0, -1$, is that of the univariate normal distribution with parameters $\omega = (\mu, \sigma^2)$. Table 2 leads to quite similar conclusions as these of the previous example. It illustrates that the divergence between two probability distributions in the whole domain $\mathcal{X}$ differs significantly in some subsets of $\mathcal{X}$ where the kernel density f_ω centers the main mass of the integral (4). Figure 2 helps us visualize the two skew normal distributions and

Table 2 Values of $D_{\phi_{2/3}}(\alpha_1,\alpha_2)$, $D_{\phi_{2/3}}(\alpha_2,\alpha_1)$, $D^{\omega}_{\phi_{2/3}}(\alpha_1,\alpha_2)$, $D^{\omega}_{\phi_{2/3}}(\alpha_2,\alpha_1)$, $D^{\omega}_{\phi_{2/3}}(\alpha_1,\alpha_2) + D^{\omega}_{\phi_{2/3}}(\alpha_2,\alpha_1)$, $I_{1,2}$, $I_{2,1}$, and $I_{1,2}+I_{2,1}$ for standard skew-normal distributions with parameters $\alpha_1 = 2$ and $\alpha_2 = -1$ and several values of $\omega = (\mu, \sigma^2)$

$\omega = (\mu, \sigma^2)$	$D^{\omega}_{\phi_{2/3}}(\alpha_1,\alpha_2)$, $D^{\omega}_{\phi_{2/3}}(\alpha_2,\alpha_1)$ $D^{\omega}_{\phi_{2/3}}(\alpha_1,\alpha_2) + D^{\omega}_{\phi_{2/3}}(\alpha_2,\alpha_1)$ $\alpha_1 = 2,\ \alpha_2 = -1$	$A \subseteq X$ $A = (\alpha, \beta)$	$I_{1,2}, I_{2,1}, I_{1,2}+I_{2,1}$
(0, 0.1)	0.0878, 0.1314, 0.2192	(−0.5, 0.5)	0.0819, 0.1015, 0.1834
(1, 0.1)	0.7249, 0.1665, 0.8914	(0.5, 1.5)	0.7284, 0.1693, 0.8977
(2, 0.1)	1.0133, 0.0671, 1.0804	(1.5, 2.5)	1.0191, 0.0666, 1.0857
(3, 0.1)	0.6426, 0.0075, 0.6501	(2.5, 3.5)	0.6425, 0.0071, 0.6496
(4, 0.1)	0.2521, 0.0003, 0.2524	(3.5, 4.5)	0.2502, 0.0003, 0.2505
(5, 0.1)	0.0678, 5.29×10^{-6}, 0.0678	(4.5, 5.5)	0.0668, 4.05×10^{-6}, 0.0668
(6, 0.1)	0.0129, 3.57×10^{-8}, 0.0129	(5.5, 6.5)	0.0126, 2.27×10^{-8}, 0.0126
$D_{\phi_{2/3}}(\alpha_1,\alpha_2) = 3.0839$, $D_{\phi_{2/3}}(\alpha_2,\alpha_1) = 2.32617 \times 10^6$		(−4, 4)	3.0786, 503072.48

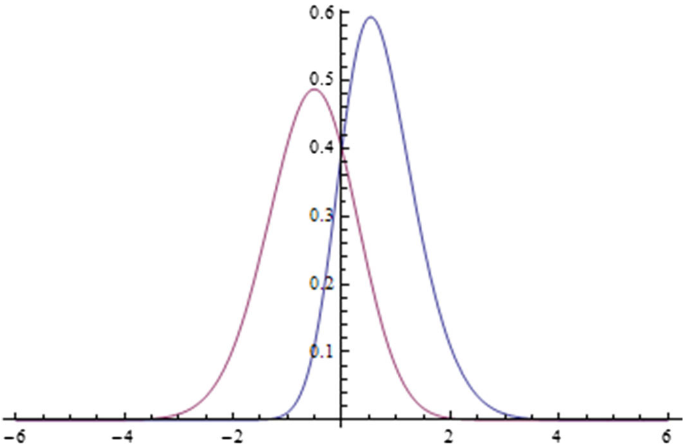

Fig. 2 Plot for standard skew-normal distributions with parameters $\alpha_1 = 2$ and $\alpha_2 = -1$

how the values of Table 2 are exemplifying the different of the two distributions. Notice that globally the measure suggests divergence while locally and in particular near the tails, the distributions are similar.

In this subsection, the definition of the local ϕ-divergence was given and its use as a measure of divergence or quasi distance between two probability distributions has been illustrated by two examples. However, the usefulness of a new proposed measure is assessed by the properties which it satisfies. The aim of the next subsection is to investigate the range of values of the local ϕ-divergence between two distributions.

2.2 Range of values of local ϕ-divergence

There is a vast list of properties that Csiszár classic ϕ -divergence, defined by (1), can satisfy. Some are of a mathematical and statistical nature, while others are motivated by particular problems of the research areas where the classic ϕ-divergence is applied. Discussions on the properties of Csiszár ϕ-divergence, are provided in the review papers by Papaioannou (1986, 2001), in the recent paper by Liese and Vajda (2006) and in the books by Liese and Vajda (1987) and Vajda (1989), to name a few.

Typically, measures are non-negative quantities. Hence, in order to avoid negativity of the local ϕ-divergence, defined by (3) or (4), interest is restricted to real convex functions ϕ which are defined on the interval $[0, \infty)$ and belong to the class of convex functions

$$\Phi^* = \left\{\phi : \phi \text{ is strictly convex at } 1, \phi(1) = 0, 0\phi\left(\frac{0}{0}\right) = 0, 0\phi\left(\frac{u}{0}\right) = u \lim_{v\to\infty} \frac{\phi(v)}{v}\right\}. \tag{6}$$

In addition, following Stummer and Vajda (2010, p. 171), for a function $\phi \in \Phi^*$, the function

$$\overline{\phi}(u) = \phi(u) - \phi'_+(1)(u-1), \tag{7}$$

belongs to the class Φ^* and it moreover satisfies,

$$\overline{\phi}(1) = \overline{\phi}'(1) = 0, \tag{8}$$

where ϕ'_+ is used to denote the right hand derivative of ϕ at the point 1. Based on Stummer and Vajda (2010, p. 171), it holds $\overline{\phi}(u) \geq 0$, for $u \geq 0$. Taking into account that $\overline{\phi}(u) \geq 0$, for $u \geq 0$ and based on (3) and (7),

$$\begin{aligned} 0 \leq D^R_{\overline{\phi}}(P, Q) &= \int_{\mathcal{X}} r(x)q(x)\overline{\phi}\left(\frac{p(x)}{q(x)}\right) d\mu(x) \\ &= \int_{\mathcal{X}} r(x)q(x)\left(\phi\left(\frac{p(x)}{q(x)}\right) - \phi'_+(1)\left(\frac{p(x)}{q(x)} - 1\right)\right) d\mu(x) \\ &= D^R_{\phi}(P, Q) - \phi'_+(1) \int_{\mathcal{X}} r(x)\left(p(x) - q(x)\right) d\mu(x). \end{aligned} \tag{9}$$

It is now clear, from (9), that the local divergence which is defined by means of the convex function $\overline{\phi}$, given in (7), is always non-negative and hence it can be considered as a measure of a local divergence between two probability distributions. Thus, motivated by Stummer and Vajda (2010, p. 171), we refine the definition of the local ϕ-divergence as follows.

Definition 2 Let P, Q and R three probability measures on the measurable space $(\mathcal{X}, \mathcal{A})$, dominated by a σ-finite measure μ which is defined on the same measurable space. If p, q and r denote the respective Radon-Nikodym derivatives and $\phi \in \Phi^*$, then the local ϕ-divergence between p and q, driven by r, is defined by

$$\widetilde{D}_{\phi}^{R}(P,Q) = D_{\overline{\phi}}^{R}(P,Q) = D_{\phi}^{R}(P,Q) - \phi'_{+}(1) \int\limits_{\mathcal{X}} r(x)\,(p(x)-q(x))\,d\mu(x), \quad (10)$$

where $D_{\overline{\phi}}^{R}(P,Q)$ is defined by (3).

Based on (9), it is clear that the two divergences $\widetilde{D}_{\phi}^{R}(P,Q)$ and $D_{\phi}^{R}(P,Q)$ coincide if we include the property $\phi'(1) = 0$ in the class Φ^*. Thus, if we consider local divergences, defined by (3) and (4), in the set of convex functions

$$\Phi = \Phi^* \cap \{\phi : \phi'(1) = 0\}, \quad (11)$$

then they are always non-negative (see also Pardo 2006, p. 6). It should be noted at this point, that all convex functions ϕ that lead to important particular cases of Csiszár ϕ-divergences, like Kullback and Leibler (1951) divergence ($\phi(u) = u \log u - u + 1$, $u > 0$), Kagan (1963) divergence ($\phi(u) = (u-1)^2$, $u > 0$), Cressie and Read (1984) λ -power divergence $\left(\phi_{\lambda}(u) = \frac{u^{\lambda+1} - u - \lambda(u-1)}{\lambda(\lambda+1)}, \lambda \neq 0, -1\right)$, and many more, belong to the set Φ, defined by (11).

The theorem that follows investigates the range of values of the local ϕ-divergence, as defined by (10). The detailed proof is given in "Appendix 1".

Theorem 1 *(a) For $\phi \in \Phi^*$, the local ϕ-divergence, as defined by* (10)*, satisfies,*

$$0 \leq \widetilde{D}_{\phi}^{R}(P,Q) \leq \phi(0)\xi_0 + \phi^*(0)\xi_1 + \phi'_{+}(1)\,(\xi_0 - \xi_1)\,,$$

with $\xi_0 = \int_{\mathcal{X}} r(x)q(x)d\mu(x)$, $\xi_1 = \int_{\mathcal{X}} r(x)p(x)d\mu(x)$ and $\phi^* \in \Phi^*$, with ϕ^* the adjoint function defined by $\phi^*(u) = u\phi\left(\frac{1}{u}\right)$, $u > 0$.

(b) $\widetilde{D}_{\phi}^{R}(P,Q) = 0$ if and only if $P = Q$.

(c) $\widetilde{D}_{\phi}^{R}(P,Q) = \phi(0)\xi_0 + \phi^*(0)\xi_1 + \phi'_{+}(1)\,(\xi_0 - \xi_1)$ if $P \perp Q$, where $\perp$ denotes singularity of probability measures. In addition, if $\phi(0) + \phi^*(0) < \infty$ and $\widetilde{D}_{\phi}^{R}(P,Q) = \phi(0)\xi_0 + \phi^*(0)\xi_1 + \phi'_{+}(1)\,(\xi_0 - \xi_1)$, then $P \perp Q$.

3 Local ϕ-divergence for exponential family of distributions

Measures of entropy or divergence have been widely applied in several disciplines and contexts not only in statistics, classic and contemporary, but in almost every branch of science and engineering. Consequently, it is of great importance to tabulate expressions for entropies or divergences for specific families of distributions. This tabulation is very useful for the development of information theoretic concepts and methods. There is an extensive literature where expressions are derived for Shannon entropy and hence for mutual information, a particular case of Kullback–Leibler divergence. For more details we refer to Soofi and Retzer (2002), Zografos and Nadarajah (2005), Zografos (2008) and the references therein.

Expressions for particular cases of Csiszár ϕ-divergence between two members of the exponential family of distributions have been obtained in Liese andVajda (1987, p. 43) and they have been utilized in testing statistical hypotheses in Morales et al. (2000, 2004). The exponential family of distributions is a broad family which includes the majority of the well known and used, in practice, statistical distributions.

Consider the exponential family of distributions with probability densities of the form

$$f_C(x,\theta) = \exp\left\{\theta^t T(x) - C(\theta) + h(x)\right\},\ x \in \mathcal{X}, \tag{12}$$

with natural parameters $\theta \in \Theta \subseteq R^k$ and $T(x) = (T_1(x), ..., T_k(x))^t$, $x \in \mathcal{X}$, where the superscript ^{t} is used to denote the transpose of a vector or a matrix.

For two members of this family, $f_C(x,\theta_i)$, $\theta_i \in \Theta \subseteq R^k$, $i = 1, 2$, the Cressie and Read local power divergence is defined, taking into account (4), for $\phi(u) = \phi_\lambda(u) = \frac{u^{\lambda+1}-u-\lambda(u-1)}{\lambda(\lambda+1)}$, $\lambda \neq 0, -1$, by

$$D^\omega_{\phi_\lambda}(\theta_1,\theta_2) = \frac{1}{\lambda(\lambda+1)}\left[K_{\lambda,\omega}(\theta_1,\theta_2) - (\lambda+1)E_{\theta_1}\left(f_\omega(X)\right) + \lambda E_{\theta_2}\left(f_\omega(X)\right)\right], \tag{13}$$

for $\lambda \neq 0, -1$, with

$$K_{\lambda,\omega}(\theta_1,\theta_2) = \int\limits_{\mathcal{X}} f_\omega(x)\frac{f_C^{\lambda+1}(x,\theta_1)}{f_C^\lambda(x,\theta_2)}d\mu(x), \tag{14}$$

$$E_{\theta_i}\left(f_\omega(X)\right) = \int\limits_{\mathcal{X}} f_\omega(x) f_C(x,\theta_i) d\mu(x) \tag{15}$$

and $\omega, \theta_i \in \Theta \subseteq R^k$, $i = 1, 2$.

The next proposition presents the analytic expression for $D^\omega_{\phi_\lambda}(\theta_1,\theta_2)$ when the kernel density f_ω is defined on $\mathcal{X}$ and it does not necessarily belong to the class of densities (12).

Proposition 2 *Let the kernel density f_ω be defined on $\mathcal{X}$ and consider two members $f_C(x,\theta_1)$ and $f_C(x,\theta_2)$ of* (12)*. If $(\lambda+1)\theta_1 - \lambda\theta_2 \in \Theta$, for $\lambda \neq 0, -1$, then the Cressie and Read local power divergence between $f_C(x,\theta_1)$ and $f_C(x,\theta_2)$, driven by the density f_ω, is given in view of* (13) *by*

$$\begin{aligned} D^\omega_{\phi_\lambda}(\theta_1,\theta_2) = \frac{1}{\lambda(\lambda+1)}&\left\{\left(\exp\left[M^{(1)}_{C,\lambda}(\theta_1,\theta_2)\right]\right)E_{(\lambda+1)\theta_1-\lambda\theta_2}\left(f_\omega(X)\right)\right. \\ &\left.-(\lambda+1)E_{\theta_1}\left(f_\omega(X)\right) + \lambda E_{\theta_2}\left(f_\omega(X)\right)\right\}, \end{aligned} \tag{16}$$

with

$$M^{(1)}_{C,\lambda}(\theta_1,\theta_2) = \lambda C(\theta_2) - (\lambda+1)C(\theta_1) + C((\lambda+1)\theta_1 - \lambda\theta_2) \tag{17}$$

and $E_{(\lambda+1)\theta_1-\lambda\theta_2}\left(f_\omega(X)\right)$, $E_{\theta_i}\left(f_\omega(X)\right)$, $i = 1, 2$, are defined by (15).

Proof Based on (14), straightforward calculations give

$$K_{\lambda,\omega}(\theta_1,\theta_2) = \int\limits_{\mathcal{X}} f_\omega(x)\exp\left([(\lambda+1)\theta_1^t - \lambda\theta_2^t]T(x)\right)\exp\left(\lambda C(\theta_2) - (\lambda+1)C(\theta_1)\right.$$
$$\left. + h(x)\right) d\mu(x).$$

Hence,

$$K_{\lambda,\omega}(\theta_1,\theta_2) = \int\limits_{\mathcal{X}} f_\omega(x)\exp\left([(\lambda+1)\theta_1^t - \lambda\theta_2^t]T(x) - C\left((\lambda+1)\theta_1 - \lambda\theta_2\right) + h(x)\right)$$
$$\times \exp\left(\lambda C(\theta_2) - (\lambda+1)C(\theta_1)\right) \times \exp\left(C\left((\lambda+1)\theta_1 - \lambda\theta_2\right)\right) d\mu(x),$$

which leads to the desired result, in view of (13) and (17). □

The proposition that follows states the analytic expression for $D_{\phi_\lambda}^{\omega}(\theta_1,\theta_2)$ when the kernel density f_ω belongs to the class of densities (12). The proof is given in "Appendix 2".

Proposition 3 *Consider two members $f_C(x,\theta_1)$ and $f_C(x,\theta_2)$ of (12) and consider the kernel density $f_\omega(x) = f_C(x,\omega)$ as a member of (12). Then, subject to the assumption $\theta_i + \omega \in \Theta$, $i = 1, 2$ and $(\lambda+1)\theta_1 - \lambda\theta_2 + \omega \in \Theta$, for $\lambda \neq 0, -1$, the Cressie and Read local power divergence between $f_C(x,\theta_1)$ and $f_C(x,\theta_2)$, driven by f_ω, is given by*

$$D_{\phi_\lambda}^{\omega}(\theta_1,\theta_2) = \frac{1}{\lambda(\lambda+1)}\Big\{\Big(\exp\Big[M_{C,\lambda}^{(2)}(\theta_1,\theta_2,\omega)\Big]\Big) E_{(\lambda+1)\theta_1-\lambda\theta_2+\omega}\left(\exp\left(h(X)\right)\right)$$
$$-(\lambda+1)\exp[C(\theta_1+\omega) - C(\theta_1) - C(\omega)] \times E_{\theta_1+\omega}\left(\exp\left(h(X)\right)\right)$$
$$+\lambda\exp[C(\theta_2+\omega) - C(\theta_2) - C(\omega)] \times E_{\theta_2+\omega}\left(\exp\left(h(X)\right)\right)\Big\}, \quad (18)$$

with

$$M_{C,\lambda}^{(2)}(\theta_1,\theta_2,\omega) = \lambda C(\theta_2) - (\lambda+1)C(\theta_1) - C(\omega) + C((\lambda+1)\theta_1 - \lambda\theta_2 + \omega) \quad (19)$$

and

$$E_{\theta_i+\omega}\left(\exp\left(h(X)\right)\right) = \int\limits_{\mathcal{X}} \{\exp\left(h(X)\right)\} f_C(x,\theta_i+\omega) d\mu(x), \quad i = 1, 2, \quad (20)$$

$$E_{(\lambda+1)\theta_1-\lambda\theta_2+\omega}\left(\exp\left(h(X)\right)\right) = \int\limits_{\mathcal{X}} \{\exp\left(h(X)\right)\} f_C(x,(\lambda+1)\theta_1 - \lambda\theta_2 + \omega) d\mu(x). \quad (21)$$

The multivariate normal model is widely used in statistics and related fields and it belongs to the exponential family model (12). The next proposition provides the explicit form of the local Cressie and Read power divergence, defined by (13), between two k-variate normal distributions, as it is driven by another k-variate normal distribution. Let the kernel density $f_{N(\mu,\Sigma)}$, be the multivariate normal distribution $N(\mu,\Sigma)$ with mean vector $\mu \in R^k$ and covariance matrix Σ. Consider also two densities $f_{N(\mu_1,\Sigma_1)}$ and $f_{N(\mu_2,\Sigma_2)}$ on $\mathcal{X} = R^k$, that follow k-variate normal distributions $N_k(\mu_1,\Sigma_1)$ and $N_k(\mu_2,\Sigma_2)$, with parameters (μ_1,Σ_1) and (μ_2,Σ_2).

The density function of the k-variate normal models with mean vectors $\mu_i \in R^k$ and covariance matrices Σ_i, $i = 1, 2$, are given by,

$$(2\pi)^{-k/2}|\Sigma_i|^{-1/2}\exp\left(-\frac{1}{2}(x-\mu_i)^t\Sigma_i^{-1}(x-\mu_i)\right),\ i=1,2.$$

It can be easily seen that the above k-variate normal distributions are included in the exponential family of distributions (12) with

$$\begin{aligned}
\theta_i &= (\theta_{i1},\theta_{i2}) = \left(\Sigma_i^{-1}\mu_i, -\frac{1}{2}\Sigma_i^{-1}\right), T(x) = (T_1(x),T_2(x)) = \left(x, xx^t\right),\\
C(\theta_i) &= \log\left((2\pi)^{k/2}|\Sigma_i|^{1/2}\right) + \frac{1}{2}\mu_i^t\Sigma_i^{-1}\mu_i = \log(2\pi)^{k/2} - \frac{1}{2}\log\left(|-2\theta_{i2}|\right)\\
&\quad -\frac{1}{4}\theta_{i1}^t\theta_{i2}^{-1}\theta_{i1}, \qquad (22)\\
h(x) &= 0,
\end{aligned}$$

where $|\ |$ is used to denote the determinant of the respective matrix. It should be noted that the inner product of $\alpha = (u, M)$ and $\beta = (v, N)$ which consist of two parts, a vectorial part u and v and a matrix part M and N, is defined by $\alpha^t\beta = u^tv + trace(M^tN)$ (cf. Nielsen and Nock 2011, p. 6).

Proposition 4 *The Cressie and Read local power divergence, defined by* (13)*, between two k-variate normal distributions $N_k(\mu_1,\Sigma_1)$ and $N_k(\mu_2,\Sigma_2)$, driven by a k-variate normal distributions $N_k(\mu,\Sigma)$, is given by*

$$\begin{aligned}
D_{\phi_\lambda}^{(\mu,\Sigma)}((\mu_1,\Sigma_1),(\mu_2,\Sigma_2)) = \frac{1}{\lambda(\lambda+1)}&\Big\{(2\pi)^{-\frac{k}{2}}|\Sigma|^{-\frac{1}{2}}|\Sigma_1|^{-\frac{\lambda+1}{2}}|\Sigma_2|^{\frac{\lambda}{2}}\Big|(\lambda+1)\Sigma_1^{-1}\\
&-\lambda\Sigma_2^{-1}+\Sigma^{-1}\Big|^{-\frac{1}{2}}\\
&\times\exp\left(-\frac{1}{2}\Big(\mu^t\Sigma^{-1}\mu+(\lambda+1)\mu_1^t\Sigma_1^{-1}\mu_1\right.\\
&\left.-\lambda\mu_2^t\Sigma_2^{-1}\mu_2 - B_1^tB_2B_1\Big)\right)\\
&-(\lambda+1)(2\pi)^{-k/2}|\Sigma|^{-1/2}|\Sigma_1|^{-1/2}\left|\Sigma_1^{-1}+\Sigma^{-1}\right|^{-1/2}\\
&\times\exp\left(-\frac{1}{2}(\mu-\mu_1)^t(\Sigma+\Sigma_1)^{-1}(\mu-\mu_1)\right)
\end{aligned}$$

$$+ \lambda(2\pi)^{-k/2}|\Sigma|^{-1/2}|\Sigma_2|^{-1/2}\left|\Sigma_2^{-1}+\Sigma^{-1}\right|^{-1/2}$$
$$\times \exp\left(-\frac{1}{2}(\mu-\mu_2)^t(\Sigma+\Sigma_2)^{-1}(\mu-\mu_2)\right)\Bigg\},$$

with

$$B_1 = (\lambda+1)\Sigma_1^{-1}\mu_1 - \lambda\Sigma_2^{-1}\mu_2 + \Sigma^{-1}\mu,$$
$$B_2 = \left((\lambda+1)\Sigma_1^{-1} - \lambda\Sigma_2^{-1} + \Sigma^{-1}\right)^{-1},$$

provided that $(\lambda+1)\Sigma_1^{-1} - \lambda\Sigma_2^{-1} + \Sigma^{-1} > 0$, *for* $\lambda \neq 0, -1$.

The proof of the proposition is given in "Appendix 3".

Remark 2 Explicit expressions for Cressie and Read local power divergence between univariate normal distributions can be derived by a direct application of the above proposition. The respective formulas are presented in Eq. (5) of the numerical Example 1.

The Kullback–Leibler local divergence is obtained from (3) or (4) for $\phi(u) = u\log u - u + 1$. It is defined by

$$\begin{aligned} D_0^R(P,Q) &= \int_{\mathcal{X}} \frac{dR}{d\mu}\frac{dP}{dQ}\log\left(\frac{dP}{dQ}\right)dQ - \int_{\mathcal{X}}\frac{dR}{d\mu}dP + \int_{\mathcal{X}}\frac{dR}{d\mu}dQ \\ &= \int_{\mathcal{X}} r(x)p(x)\log\left(\frac{p(x)}{q(x)}\right)d\mu(x) - \int_{\mathcal{X}} r(x)p(x)d\mu(x) \\ &\quad + \int_{\mathcal{X}} r(x)q(x)d\mu(x). \end{aligned} \tag{23}$$

It should be noted that Kullback–Leibler classic divergence is obtained from (1) for $\phi(u) = u\log u$ or $\phi(u) = u\log u - u + 1$. Both choices of the convex function ϕ lead to the same quantity. This is not the case for Kullback–Leibler local divergence. It is defined by (23), as a particular case of (3) or (4) for $\phi(u) = u\log u - u + 1$.

The next Proposition provides the explicit forms of Kullback–Leibler local divergence between two members of the exponential family and between two multivariate normal distributions, as well.

Proposition 5 *(a) The Kullback–Leibler local divergence* (23) *between two members* $f_C(x,\theta_1)$ *and* $f_C(x,\theta_2)$ *of the exponential family* (12)*, driven by the kernel density* $f_C(x,\omega)$ *in* (12)*, is given by*

$$\begin{aligned} D_0^{\omega}(\theta_1,\theta_2) = {} & (\exp(C(\theta_1+\omega)-C(\theta_1)-C(\omega)))\left\{(C(\theta_2)-C(\theta_1))\,E_{\theta_1+\omega}(\exp(h(X)))\right. \\ & \left. + (\theta_1-\theta_2)^t E_{\theta_1+\omega}(T(X)\exp(h(X)))\right\} \\ & - (\exp(C(\theta_1+\omega)-C(\theta_1)-C(\omega)))\,E_{\theta_1+\omega}(\exp(h(X))) \\ & + (\exp(C(\theta_2+\omega)-C(\theta_2)-C(\omega)))\,E_{\theta_2+\omega}(\exp(h(X))), \end{aligned}$$

and $E_{\theta_i+\omega}\left(\exp\left(h(X)\right)\right)$*,* $i=1,2$ *, are defined by* (20).
(b) The Kullback–Leibler local divergence (23) *between two multivariate normal distributions* $f_{N(\mu_1,\Sigma_1)}$ *and* $f_{N(\mu_2,\Sigma_2)}$*, on* $\mathcal{X}=R^k$*, driven by the multivariate normal density* $f_{N(\mu,\Sigma)}$*, is given by*

$$\begin{aligned}D_0^{(\mu,\Sigma)}((\mu_1,\Sigma_1),(\mu_2,\Sigma_2)) = {} & \frac{1}{2}(2\pi)^{-\frac{k}{2}}|\Sigma|^{-\frac{1}{2}}|\Sigma_1|^{-\frac{1}{2}}\left|\Sigma^{-1}+\Sigma_1^{-1}\right|^{-\frac{1}{2}} \\ & \times\exp\left(-\frac{1}{2}(\mu-\mu_1)^t(\Sigma+\Sigma_1)^{-1}(\mu-\mu_1)\right) \\ & \times\left\{\log\frac{|\Sigma_2|}{|\Sigma_1|}-trace\left((\Sigma^{-1}+\Sigma_1^{-1})^{-1}(\Sigma_1^{-1}-\Sigma_2^{-1})\right)\right. \\ & -(\mu^*-\mu_1)^t\Sigma_1^{-1}(\mu^*-\mu_1) \\ & \left.+(\mu^*-\mu_2)^t\Sigma_2^{-1}(\mu^*-\mu_2)\right\} \\ & -E_{(\mu_1,\Sigma_1)}\left(f_{N(\mu,\Sigma)}(X)\right)+E_{(\mu_2,\Sigma_2)}\left(f_{N(\mu,\Sigma)}(X)\right),\end{aligned}$$

where

$$\begin{aligned}E_{(\mu_i,\Sigma_i)}\left(f_{N(\mu,\Sigma)}(X)\right) = {} & (2\pi)^{-\frac{k}{2}}|\Sigma|^{-\frac{1}{2}}|\Sigma_i|^{-\frac{1}{2}}\left|\Sigma^{-1}+\Sigma_i^{-1}\right|^{-\frac{1}{2}} \\ & \times\exp\left\{-\frac{1}{2}(\mu-\mu_i)^t(\Sigma+\Sigma_i)^{-1}(\mu-\mu_i)\right\},\ i=1,2,\end{aligned}$$

and

$$\mu^*=(\Sigma^{-1}+\Sigma_1^{-1})^{-1}(\Sigma^{-1}\mu+\Sigma_1^{-1}\mu_1).$$

The proof of the previous proposition is given in “Appendix 4”.

Remark 3 (a) Explicit expressions for Kullback–Leibler local divergence between univariate normal distributions can be derived by a direct application of the part (b) of above proposition. The respective formulas, are given by

$$\begin{aligned}D_0^{(\mu,\sigma^2)}((\mu_1,\sigma_1^2),(\mu_2,\sigma_2^2)) = {} & \frac{1}{2}(2\pi(\sigma^2+\sigma_1^2))^{-\frac{1}{2}}\exp\left(-\frac{(\mu-\mu_1)^2}{2(\sigma^2+\sigma_1^2)}\right) \\ & \times\left(\log\frac{\sigma_2^2}{\sigma_1^2}-\frac{\sigma^2(\sigma_2^2-\sigma_1^2)}{\sigma_2^2(\sigma^2+\sigma_1^2)}-\frac{(\mu^*-\mu_1)^2}{\sigma_1^2}+\frac{(\mu^*-\mu_2)^2}{\sigma_2^2}\right) \\ & -E_{(\mu_1,\sigma_1^2)}\left(f_{N(\mu,\sigma^2)}(X)\right)+E_{(\mu_2,\sigma_2^2)}\left(f_{N(\mu,\sigma^2)}(X)\right),\end{aligned}$$

where

$$E_{(\mu_i,\sigma_i^2)}\left(f_{N(\mu,\sigma^2)}(X)\right)=\left(2\pi(\sigma^2+\sigma_i^2)\right)^{-1/2}\exp\left\{-\frac{(\mu-\mu_i)^2}{2(\sigma^2+\sigma_i^2)}\right\},\ i=1,2,$$

Table 3 Normality tests and descriptive statistics for the three populations of GPA scores

Population	Sample size	Normality tests (p values)		Sample mean and variance
	n	Shapiro-Wilk	Kolmogorov–Smirnov	
Π_1	31	0.86	0.2 (lower bound)	$(\overline{X}_1 = 3.40, S_1^2 = 0.04)$
Π_2	28	0.85	0.2 (lower bound)	$(\overline{X}_2 = 2.48, S_2^2 = 0.03)$
Π_3	26	0.11	0.2 (lower bound)	$(\overline{X}_3 = 2.99, S_3^2 = 0.03)$

Notice that normality is a reasonable assumption for all three populations

and

$$\mu^* = \frac{\mu\sigma_1^2 + \mu_1\sigma^2}{\sigma^2 + \sigma_1^2}.$$

(b) Proposition 5 (b) can be used in order to obtain the explicit expression for the Kullback–Leibler local divergence between two multivariate normal distributions with common covariance matrix $\Sigma_1 = \Sigma_2 = \Sigma_*$. It is easily obtained by a straightforward application of Proposition 5 (b) for $\Sigma_1 = \Sigma_2 = \Sigma_*$.

4 Application

We now illustrate the behaviour of local measures in real life situations. Consider grade point average (GPA) scores for students seeking admission in a business school (Johnson and Wichern 1992, p. 532, Example 11.11). There are three groups of applicants who have been categorized as Π_1: admit, Π_2: don't admit, and Π_3: borderline, depending on their GPA scores. Note that the support of the three distributions is the same. We are interested in exemplifying any differences between the three populations of students, either globally, i.e., over the whole domain of the distributions describing each population, or locally, by focusing on a specific area of the domain of observation where two populations might differ. The latter is accomplished by considering the center of the kernel distribution to be a convex combination of the means of the two populations under consideration. In this way, the kernel acts as a window that can move across the domain of observation and focus on a small region each time, that depends on the variability or spreadness of the kernel.

In Table 3 we display normality tests along with basic descriptive statistics for the three populations, including sample means and variances. Notice that the normality assumption is reasonable for the three groups of students, and hence we adapt the three univariate normal distributions in order to describe the data. Under normality, knowing the mean and variance completely determines the behavior of the distributions. Figure 3 illustrates the three densities for $\Pi_1 - \Pi_3$ using the estimated means and variances from Table 3.

Using the notation of Example 1, we utilize the Cressie–Read λ-power divergence in order to compare populations Π_1 with Π_2, Π_1 with Π_3, and Π_2 with Π_3, in Tables 4, 5 and 6, respectively. In all tables, we present the local Cressie–Read diver-

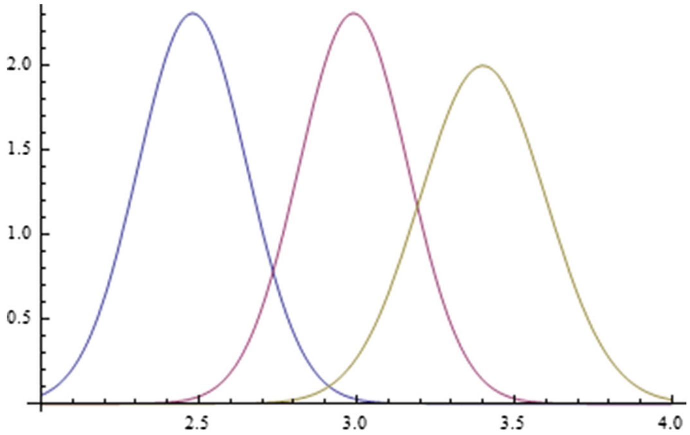

Fig. 3 Plot for three densities of $\Pi_1 - \Pi_3$ using the estimated means and variances

gence $D^{\omega}_{\phi_\lambda}(\widehat{\theta_1}, \widehat{\theta_2})$ for different values of λ, namely, $\lambda = -2, -0.5, \frac{2}{3}, 1$ or 2. The bottom rows show the values of the global measure $D_{\phi_\lambda}(\widehat{\theta_1}, \widehat{\theta_2})$. The kernel and population models are univariate normal distributions, with estimated parameters for Π_1, Π_2 and Π_3 given by $\widehat{\theta_1} = (\widehat{\mu}_1, \widehat{\sigma}_1^2) = (3.40, 0.04)$, $\widehat{\theta_2} = (\widehat{\mu}_2, \widehat{\sigma}_2^2) = (2.48, 0.03)$, and $\widehat{\theta_3} = (\widehat{\mu}_3, \widehat{\sigma}_3^2) = (2.99, 0.03)$, respectively. Using these estimators, we obtain convex combinations of the means for different values of k, and treat the result as the mean of the kernel. The kernel parameters are displayed in the second column of Tables 4, 5, and 6, i.e., the parameters of the kernel are $\theta = (\mu, \sigma^2) = (k\widehat{\mu}_1 + (1-k)\widehat{\mu}_2, 0.1)$. Notice that the variance of the kernel is 0.1 in all cases, a small value that puts more weight on values about μ, thus highlighting the differences of the two populations at a region near the mean of the kernel. The values of k considered are $k = 0, 0.1, 0.3, 0.5, 0.7, 0.9$ and 1, and lead to a window in the domain of observation that moves from one population mean towards the other. When two populations are close to each other in a certain window, we expect the local measure to take smaller values, unless the two populations are completely different. This assertion is supported by the results in Tables 4, 5 and 6, with all values away from zero, indicating that all populations are different from each other. For example, when comparing Π_2 with Π_3, using $\lambda = \frac{2}{3}$, the global measure is $D_{\phi_{2/3}}(\widehat{\theta_2}, \widehat{\theta_3}) = 110.3$, while a value of $k = 0$, yields a local measure with value $D^{\omega}_{\phi_{2/3}}(\widehat{\theta_2}, \widehat{\theta_3}) = 7.7$. The region we focus in this case is described by the kernel with mean being the same as the mean of Π_3, but the kernel variance ($\sigma^2 = 0.1$) is much larger than that of $\Pi_3(\widehat{\sigma}_3^2 = 0.03)$.

We investigate the behavior of the local Kullback–Leibler measure in Table 7, with similar results to the Cressie–Read divergence. All tables suggest that the three populations are clearly different globally, although some values of λ indicate that populations Π_2 and Π_3 are not as different locally, when the kernel focuses attention near the mean of Π_3.

Table 4 Displaying the local Cressie–Read divergence $D^{\omega}_{\phi_\lambda}(\widehat{\theta}_1, \widehat{\theta}_2)$ for different values of λ, in order to compare populations Π_1 and Π_2

k	Π_1-Π_2					
	$Kernel(\mu, \sigma^2)$	$\lambda = -2$	$\lambda = -0.5$	$\lambda = \frac{2}{3}$	$\lambda = 1$	$\lambda = 2$
0	(2.48, 0.1)	3.84881×10^{6}	2.20	1435.39	5.81244×10^{7}	7.44435×10^{41}
0.1	(2.57, 0.1)	2.49542×10^{6}	2.18	4509.07	2.67675×10^{8}	2.16225×10^{43}
0.3	(2.76, 0.1)	806689.19	1.93	42412.55	5.69638×10^{9}	2.35027×10^{46}
0.5	(2.94, 0.1)	211568.98	1.77	284582.94	8.38124×10^{10}	1.51982×10^{49}
0.7	(3.13, 0.1)	38797.18	1.91	1.68285×10^{6}	1.14943×10^{12}	1.20013×10^{52}
0.9	(3.31, 0.1)	5947.27	2.12	7.27367×10^{6}	1.11536×10^{13}	5.73363×10^{54}
1	(3.40, 0.1)	2111.43	2.14	1.39560×10^{7}	3.22035×10^{13}	1.18589×10^{56}
Global	$\phi_\lambda - divergence$	1.16070×10^{7}	3.81	5.23885×10^{8}	1.27025×10^{18}	4.08814×10^{109}

The bottom row shows the value of the global measure $D_{\phi_\lambda}(\widehat{\theta}_1, \widehat{\theta}_2)$. The kernel and population models are univariate normal distributions, with estimated parameters for Π_1 and Π_2 given by $\widehat{\theta}_1 = (\widehat{\mu}_1, \widehat{\sigma}_1^2) = (3.40, 0.04)$ and $\widehat{\theta}_2 = (\widehat{\mu}_2, \widehat{\sigma}_2^2) = (2.48, 0.03)$, respectively. Using these estimators, we obtain convex combinations of the means for different values of k, and treat the result as the mean of the kernel. The kernel parameters are displayed in the second column, i.e., $\theta = (\mu, \sigma^2) = (k\widehat{\mu}_1 + (1-k)\widehat{\mu}_2, 0.1)$. All values indicate divergence between Π_1 and Π_2

Table 5 Displaying the local Cressie–Read divergence $D^{\omega}_{\phi_\lambda}(\widehat{\theta}_1, \widehat{\theta}_3)$ for different values of λ, in order to compare populations Π_1 and Π_3

k	$\Pi_1 - \Pi_3$					
	$Kernel\,(\mu, \sigma^2)$	$\lambda = -2$	$\lambda = -0.5$	$\lambda = \frac{2}{3}$	$\lambda = 1$	$\lambda = 2$
0	(2.99, 0.1)	12.41	1.26	3.85	20.88	5.33459×10^7
0.1	(3.03, 0.1)	11.36	1.29	4.78	28.19	1.03967×10^8
0.3	(3.12, 0.1)	8.91	1.33	7.59	53.60	4.54339×10^8
0.5	(3.20, 0.1)	6.83	1.35	11.06	91.09	1.63409×10^9
0.7	(3.28, 0.1)	5.00	1.36	15.53	148.82	5.70873×10^9
0.9	(3.36, 0.1)	3.53	1.36	20.95	233.62	1.93717×10^{10}
1	(3.40, 0.1)	2.93	1.35	23.95	288.35	3.52976×10^{10}
Global	$\phi_\lambda - divergence$	14.40	1.82	49.84	2369.73	1.7256×10^{21}

The bottom row shows the value of the global measure $D_{\phi_\lambda}(\widehat{\theta}_1, \widehat{\theta}_3)$. The kernel and population models are univariate normal distributions, with estimated parameters for Π_1 and Π_3 given by $\widehat{\theta}_1 = (\widehat{\mu}_1, \widehat{\sigma}_1^2) = (3.40, 0.04)$ and $\widehat{\theta}_3 = (\widehat{\mu}_3, \widehat{\sigma}_3^2) = (2.99, 0.03)$, respectively. Using these estimators, we obtain convex combinations of the means for different values of k, and treat the result as the mean of the kernel. The kernel parameters are displayed in the second column, i.e., $\theta = (\mu, \sigma^2) = (k\widehat{\mu}_1 + (1-k)\widehat{\mu}_3, 0.1)$. All values indicate divergence between Π_1 and Π_3

Table 6 Displaying the local Cressie–Read divergence $D^{\omega}_{\phi_\lambda}(\widehat{\theta}_2, \widehat{\theta}_3)$ for different values of λ, in order to compare populations Π_2 and Π_3

k	$\Pi_2 - \Pi_3$					
	$Kernel\,(\mu, \sigma^2)$	$\lambda = -2$	$\lambda = -0.5$	$\lambda = \frac{2}{3}$	$\lambda = 1$	$\lambda = 2$
0	(2.99, 0.1)	1184.27	1.86	7.70	59.08	4.48366×10^6
0.1	(2.94, 0.1)	963.90	1.90	10.42	86.47	7.99876×10^6
0.3	(2.84, 0.1)	602.76	1.94	18.29	175.20	2.40296×10^7
0.5	(2.74, 0.1)	349.05	1.95	30.05	329.10	6.68441×10^7
0.7	(2.64, 0.1)	187.24	1.94	45.95	572.69	1.72176×10^8
0.9	(2.53, 0.1)	86.47	1.90	67.19	963.90	4.46054×10^8
1	(2.48, 0.1)	59.08	1.86	77.46	1184.27	6.66715×10^8
Global	$\phi_\lambda - divergence$	2912.25	2.65	110.30	2912.25	3.29495×10^{10}

The bottom row shows the value of the global measure $D_{\phi_\lambda}(\widehat{\theta}_2, \widehat{\theta}_3)$. The kernel and population models are univariate normal distributions, with estimated parameters for Π_2 and Π_3 given by $\widehat{\theta}_2 = (\widehat{\mu}_2, \widehat{\sigma}_2^2) = (2.48, 0.03)$ and $\widehat{\theta}_3 = (\widehat{\mu}_3, \widehat{\sigma}_3^2) = (2.99, 0.03)$, respectively. Using these estimators, we obtain convex combinations of the means for different values of k, and treat the result as the mean of the kernel. The kernel parameters are displayed in the second column, i.e., $\theta = (\mu, \sigma^2) = (k\widehat{\mu}_2 + (1-k)\widehat{\mu}_3, 0.1)$. All values indicate divergence between Π_2 and Π_3

Table 7 Displaying the local Kullback–Leibler divergence $D_0^{\omega}(.,.)$, in order to compare all populations

k	$\Pi_1-\Pi_2$		$\Pi_1-\Pi_3$		$\Pi_2-\Pi_3$	
	(μ,σ^2)	$D_0^{\omega}(\theta_1,\theta_2)$	(μ,σ^2)	$D_0^{\omega}(\theta_1,\theta_3)$	(μ,σ^2)	$D_0^{\omega}(\theta_2,\theta_3)$
0	(2.48, 0.1)	1.38	(2.99, 0.1)	1.24	(2.99, 0.1)	1.65
0.1	(2.57, 0.1)	1.62	(3.03, 0.1)	1.35	(2.94, 0.1)	1.85
0.3	(2.76, 0.1)	2.70	(3.12, 0.1)	1.61	(2.84, 0.1)	2.31
0.5	(2.94, 0.1)	5.06	(3.20, 0.1)	1.87	(2.74, 0.1)	2.85
0.7	(3.13, 0.1)	9.04	(3.28, 0.1)	2.13	(2.64, 0.1)	3.41
0.9	(3.31, 0.1)	12.81	(3.36, 0.1)	2.37	(2.53, 0.1)	3.93
1	(3.40, 0.1)	13.99	(3.40, 0.1)	2.47	(2.48, 0.1)	4.10
Global	$D_0(\theta_1,\theta_2)=14.13$		$D_0(\theta_1,\theta_3)=2.82$		$D_0(\theta_2,\theta_3)=4.34$	

The bottom row shows the value of the global measure $D_0(.,.)$. The kernel and population models are univariate normal distributions, with estimated parameters for $\Pi_{1,}\Pi_2$ and Π_3 given by $\widehat{\theta}_1=(\widehat{\mu}_1,\widehat{\sigma}_1^2)=(3.40, 0.04)$,$\widehat{\theta}_2=(\widehat{\mu}_2,\widehat{\sigma}_2^2)=(2.48, 0.03)$ and $\widehat{\theta}_3=(\widehat{\mu}_3,\widehat{\sigma}_3^2)=(2.99, 0.03)$, respectively. Using these estimators, we obtain convex combinations of the means for different values of k, and treat the result as the mean of the kernel. All values indicate divergence between $\Pi_1-\Pi_2$, $\Pi_1-\Pi_3$ and $\Pi_2-\Pi_3$

5 Conclusions

This paper introduces a broad class of divergence measures between two probability measures or between the respective probability distributions. The proposed measure has its origins on Csiszár classic ϕ -divergence, a measure with numerous applications not only in probability and statistics but in many areas of science and engineering. It provides us with a tool to locally quantify the pseudo-distance between two distributions on a specific area of their common domain that might be of particular interest from a theoretical or applied point of view. The range of values of the introduced class of local divergences has been derived and the measures attain their minimum value if and only if the underlined probability measures or the respective probability distributions coincide. Explicit expressions of the proposed local divergences have been derived when the underlined distributions are members of the exponential family of distributions or they are described by multivariate normal models.

Our simulations illustrated the robust behavior of the local against the global measure, in the sense that differences between two populations that cannot be captured or are otherwise obscured globally, are exemplified by using the appropriate kernel locally. Moreover, important aspects of the two models under comparison can be asserted more efficiently at the local level using the right kernel, including tail behavior, central tendency and local variability (see Example 2).

There are several extensions to this work that we will consider. Firstly, the theoretical framework laid down in this paper will be extended to study other important properties of the local divergence including sufficiency and robustness with respect to choice of models and kernel. Secondly, we will explore the use of the local measure in the creation of local tests for the difference in means and variances between the two models. Finally, the local measure will be illustrated as a tool for local goodness-of-fit tests. These are subjects of future research and will be explored elsewhere.

Acknowledgments The authors are grateful to a Reviewer for valuable comments and suggestions which have improved the presentation of the paper.

Appendix 1

This appendix provides a detailed proof of Theorem 1.

Proof of Theorem 1 (a) It is clear, from (9) and (10), that $0 \leq \widetilde{D}_{\phi}^{R}(P, Q)$. We proceed with the upper bound of $\widetilde{D}_{\phi}^{R}(P, Q)$. Given that $\overline{\phi}(1) = 0$ and motivated by a similar proof in Stummer and Vajda (2010, p. 174), we can write

$$\widetilde{D}_{\phi}^{R}(P, Q) = \int\limits_{\{p<q\}} r(x)q(x)\overline{\phi}\left(\frac{p(x)}{q(x)}\right) d\mu(x) + \int\limits_{\{q<p\}} r(x)q(x)\overline{\phi}\left(\frac{p(x)}{q(x)}\right) d\mu(x). \tag{24}$$

Define the function

$$\overline{\phi^{*}}(u) = \phi^{*}(u) + \phi_{+}'(1)(u-1). \tag{25}$$

Then, for $u = \frac{q(x)}{p(x)} = \frac{q}{p}, x \in \mathcal{X}$,

$$rp\overline{\phi^{*}}\left(\frac{q}{p}\right) = rp\phi^{*}\left(\frac{q}{p}\right) + rp\phi_{+}'(1)\left(\frac{q}{p} - 1\right). \tag{26}$$

Hence,

$$\int\limits_{\{q<p\}} r(x)p(x)\overline{\phi^{*}}\left(\frac{q(x)}{p(x)}\right) d\mu(x) = \int\limits_{\{q<p\}} r(x)p(x)\phi^{*}\left(\frac{q(x)}{p(x)}\right) d\mu(x) + \phi_{+}'(1) \times \int\limits_{\{q<p\}} r(x)\left(q(x) - p(x)\right) d\mu(x)$$

and taking into account that $\phi^{*}(u) = u\phi\left(1/u\right), u > 0$,

$$\begin{aligned} \int\limits_{\{q<p\}} r(x)p(x)\overline{\phi^{*}}\left(\frac{q(x)}{p(x)}\right) d\mu(x) &= \int\limits_{\{q<p\}} r(x)q(x)\phi\left(\frac{p(x)}{q(x)}\right) d\mu(x) - \phi_{+}'(1) \\ &\quad \times \int\limits_{\{q<p\}} r(x)\left(p(x) - q(x)\right) d\mu(x) \\ &= \int\limits_{\{q<p\}} r(x)q(x)\overline{\phi}\left(\frac{p(x)}{q(x)}\right) d\mu(x). \end{aligned} \tag{27}$$

Therefore, from Eqs. (24) and (27) we conclude

$$\widetilde{D}_{\phi}^{R}(P,Q)=\int\limits_{\{p<q\}} r(x)q(x)\overline{\phi}\left(\frac{p(x)}{q(x)}\right)d\mu(x)+\int\limits_{\{q<p\}} r(x)p(x)\overline{\phi^{*}}\left(\frac{q(x)}{p(x)}\right)d\mu(x). \tag{28}$$

On the other hand, based on Stummer and Vajda (2010, p. 174), for a convex function $\phi\in\Phi^{*}$ which is strictly convex at 1, with $\phi_{+}'(1)=0$, it is true that

$$0=\phi(1)\leq\phi(t_2)\leq\phi(t_1)\leq\phi(0),\ \text{ for any } 0\leq t_1\leq t_2\leq 1. \tag{29}$$

Applying inequality (29) to $\phi=\overline{\phi}$, it is clear that on the subset $\{x\in\mathcal{X}:\,p(x)<q(x)\}$ of $\mathcal{X}$, it holds that $0\leq\overline{\phi}\left(\frac{p}{q}\right)\leq\overline{\phi}(0)$, and therefore

$$0\leq\int\limits_{\{p<q\}} r(x)q(x)\overline{\phi}\left(\frac{p(x)}{q(x)}\right)d\mu(x)\leq\overline{\phi}(0)\int\limits_{\{p<q\}} r(x)q(x)d\mu(x).$$

Moreover, based on the non-negativity of r and q on any subset of $\mathcal{X}$, it is true that

$$0\leq\int\limits_{\{p<q\}} r(x)q(x)d\mu(x)\leq\int\limits_{\mathcal{X}} r(x)q(x)d\mu(x)=\xi_0.$$

So, the last two inequalities lead to

$$0\leq\int\limits_{\{p<q\}} r(x)q(x)\overline{\phi}\left(\frac{p(x)}{q(x)}\right)d\mu(x)\leq\overline{\phi}(0)\xi_0. \tag{30}$$

In a manner quite similar to the above, applying inequality (29) to $\phi=\overline{\phi^{*}}$, it is clear that on the subset $\{x\in\mathcal{X}:q(x)<p(x)\}$ of $\mathcal{X}$ it holds that $0\leq\overline{\phi^{*}}\left(\frac{q}{p}\right)\leq\overline{\phi^{*}}(0)$, and therefore

$$0\leq\int\limits_{\{q<p\}} r(x)p(x)\overline{\phi^{*}}\left(\frac{q(x)}{p(x)}\right)d\mu(x)\leq\overline{\phi^{*}}(0)\int\limits_{\{q<p\}} r(x)p(x)d\mu(x).$$

Hence,

$$0\leq\int\limits_{\{q<p\}} r(x)p(x)\overline{\phi^{*}}\left(\frac{q(x)}{p(x)}\right)d\mu(x)\leq\overline{\phi^{*}}(0)\int\limits_{\{q<p\}} r(x)p(x)d\mu(x)\leq\overline{\phi^{*}}(0)\xi_1. \tag{31}$$

A combination of (28), (30) and (31) gives

$$\widetilde{D}_{\phi}^{R}(P,Q)\leq\overline{\phi}(0)\xi_0+\overline{\phi^{*}}(0)\xi_1.$$

Based on (7) and (25), $\overline{\phi}(0) = \phi(0) + \phi'_+(1)$ and $\overline{\phi^*}(0) = \phi^*(0) - \phi'_+(1)$. These identities along with the previous inequality complete the proof of part (a) of the theorem.

(b) To proceed with the proof of part (b) of the theorem, suppose first that $P = Q$. Then, it is clear from (10) that $\widetilde{D}^R_\phi(P, P) = D^R_\phi(P, P) = \phi(1) = 0$, because $\phi \in \Phi^*$.

Conversely, let $\widetilde{D}^R_\phi(P, Q) = 0$. Taking into account (9) and the fact that $\phi(1) = 0$,

$$\phi\left(\frac{p(x)}{q(x)}\right) = \phi(1) + \phi'_+(1)\left(\frac{p(x)}{q(x)} - 1\right), \tag{32}$$

a.e. with respect to measure μ, for Radon-Nikodym derivative r positive on $\mathcal{X}$. On the other hand, based on Vajda (1989, p. 58)

$$\phi(x) > \phi(1) + \phi'_+(1)(x - 1), \text{ for every } x \neq 1,$$

because ϕ is strictly convex at 1. Therefore, the only way equality (32) to be valid, taking into account the above inequality, is when $\frac{p(x)}{q(x)} = 1$ or $P = Q$, which completes the proof of part (b) of the theorem.

(c) Suppose that $P \perp Q$. Then, following Vajda (1972, p. 227), $u = \frac{dQ}{dP+dQ} = 0\ [P]$ and $u = \frac{dQ}{dP+dQ} = 1\ [Q]$. Taking into account that $u = \frac{q}{p+q}$, we conclude that if $P \perp Q$, then $q(x) = 0$, $a.e.\ x \in \mathcal{X}\,[P]$ and $p(x) = 0$, $a.e.\ x \in \mathcal{X}\ [Q]$. Equation (28) is refined as follows,

$$\widetilde{D}^R_\phi(P, Q) = \int\limits_{\{p<q\}} r(x)\overline{\phi}\left(\frac{p(x)}{q(x)}\right) dQ(x) + \int\limits_{\{q<p\}} r(x)\overline{\phi^*}\left(\frac{q(x)}{p(x)}\right) dP(x),$$

and subject to the condition $P \perp Q$,

$$\begin{aligned}\widetilde{D}^R_\phi(P, Q) &= \int\limits_{\{p<q\}} r(x)\overline{\phi}\left(\frac{p(x)}{q(x)}\right) dQ(x) + \int\limits_{\{q<p\}} r(x)\overline{\phi^*}\left(\frac{q(x)}{p(x)}\right) dP(x)\\ &= \overline{\phi}\,(0) \int\limits_{\{p<q\}} r(x)dQ(x) + \overline{\phi^*}\,(0) \int\limits_{\{q<p\}} r(x)dP(x).\end{aligned} \tag{33}$$

On the other hand, because of $p(x) = 0$, $a.e.\ x \in \mathcal{X}\ [Q]$, it is clear that

$$Q(\{p \geq q\}) = Q(\{p > q\}) + Q(\{p = q\}) = Q(\{p > q\}) = Q(\{q < 0\}) = Q(\varnothing) = 0$$

since $Q(\{q = p\}) = 0$, by taking into account that $P \perp Q$. This last equality leads to $\int_{\{p \geq q\}} r(x) dQ(x) = 0$, and therefore

$$\xi_0 = \int\limits_{\mathcal{X}} r(x)q(x)d\mu(x) = \int\limits_{\mathcal{X}} r(x)dQ(x) = \int\limits_{\{p<q\}} r(x)dQ(x) + \int\limits_{\{p\geq q\}} r(x)dQ(x)$$

$$= \int\limits_{\{p<q\}} r(x)dQ(x). \tag{34}$$

Similarly, it can be shown that $P(\{q \geq p\}) = 0$ and hence

$$\xi_1 = \int\limits_{\mathcal{X}} r(x)q(x)d\mu(x) = \int\limits_{\{q<p\}} r(x)dP(x). \tag{35}$$

Equations (33), (34) and (35) give that if $P \perp Q$ then $\widetilde{D}_{\phi}^{R}(P, Q) = \overline{\phi}\,(0)\,\xi_0 + \overline{\phi^*}\,(0)\,\xi_1$ which completes the proof of this part of the theorem in view of the equations $\overline{\phi}(0) = \phi(0) + \phi'_{+}(1)$ and $\overline{\phi^*}(0) = \phi^*(0) - \phi'_{+}(1)$.

It remains to prove that if $\phi(0) + \phi^*(0) < \infty$ and $\widetilde{D}_{\phi}^{R}(P, Q) = \overline{\phi}\,(0)\,\xi_0 + \overline{\phi^*}\,(0)\,\xi_1$, then $P \perp Q$. Relationships (24) and (29) immediately lead to

$$\widetilde{D}_{\phi}^{R}(P, Q) \leq \overline{\phi}\,(0) \int\limits_{\{p<q\}} r(x)dQ(x) + \overline{\phi^*}\,(0) \int\limits_{\{q<p\}} r(x)dP(x).$$

This last inequality with the assumption $\widetilde{D}_{\phi}^{R}(P, Q) = \overline{\phi}\,(0)\,\xi_0 + \overline{\phi^*}\,(0)\,\xi_1$ and $\phi(0) + \phi^*(0) < \infty$ lead to

$$\int\limits_{\{p<q\}} r(x)dQ(x) = \xi_0 \text{ and } \int\limits_{\{q<p\}} r(x)dP(x) = \xi_1,$$

and therefore

$$\int\limits_{\{p\geq q\}} r(x)dQ(x) = \int\limits_{\{q\geq p\}} r(x)dP(x) = 0.$$

This last equation lead to

$$Q(\{p \geq q\}) = 0 \text{ and } P(\{q \geq p\}) = 0,$$

or

$$Q(\{p \geq q\}) = 0 \text{ and } P(\{q > p\}) = 0$$

for Radon-Nikodym derivative r positive on $\mathcal{X}$. This last conclusion proves that $P \perp Q$ and the proof of the theorem is completed. □

Appendix 2

This appendix provides a detailed proof of Proposition 3.

Proof of Proposition 3. Based on (14), straightforward calculations give

$$\begin{aligned} K_{\lambda,\omega}(\theta_1,\theta_2) &= \exp\{\lambda C(\theta_2)-(\lambda+1)C(\theta_1)-C(\omega)+C((\lambda+1)\theta_1-\lambda\theta_2+\omega)\} \\ &\quad \times \int\limits_{\mathcal{X}} \exp\{h(x)\}\exp\left\{\left(\sum_{i=1}^{k}[(\lambda+1)\theta_{1i}-\lambda\theta_{2i}+\omega_i]T_i(x)\right)\right. \\ &\quad \left. -C((\lambda+1)\theta_1-\lambda\theta_2+\omega)+h(x)\right\} d\mu(x). \end{aligned}$$

Taking into account (19) and (21),

$$K_{\lambda,\omega}(\theta_1,\theta_2) = \exp\left\{M_{C,\lambda}^{(2)}(\theta_1,\theta_2,\omega)\right\} E_{(\lambda+1)\theta_1-\lambda\theta_2+\omega}\{\exp(h(x))\}. \tag{36}$$

On the other hand, it can be easily shown that $E_{\theta_j}(f_\omega(X)) = \int_{\mathcal{X}} f_\omega(x) f_C(x,\theta_j)d\mu(x)$, $j=1,2$, defined by (15), is given by

$$\begin{aligned} E_{\theta_j}(f_\omega(X)) &= \exp\{-C(\theta_j)-C(\omega)+C(\theta_j+\omega)\} \\ &\quad \times \int\limits_{\mathcal{X}} \exp(h(x))\exp\left\{\left(\sum_{i=1}^{k}(\omega_i+\theta_{ji})T_i(x)-C(\theta_j+\omega)+h(x)\right)\right\} \\ &\quad d\mu(x),\ \ j=1,2, \end{aligned}$$

and therefore

$$E_{\theta_j}(f_\omega(X)) = \exp\{-C(\theta_j)-C(\omega)+C(\theta_j+\omega)\}E_{\theta_j+\omega}\left(\exp\left(h(X)\right)\right),\ \ j=1,2. \tag{37}$$

The result (18) follows as an application of (13), (36) and (37). □

Appendix 3

This appendix provides a detailed proof of Proposition 4.

Proof of Proposition 4. Based on Proposition 3,

$$\begin{aligned} &D_{\phi_\lambda}^{(\mu,\Sigma)}((\mu_1,\Sigma_1),(\mu_2,\Sigma_2)) \\ &= \frac{1}{\lambda(\lambda+1)}\left\{\left(\exp\left[M_{C,\lambda}^{(2)}(\theta_1,\theta_2,\omega)\right]\right) E_{(\lambda+1)\theta_1-\lambda\theta_2+\omega}\left(\exp\left(h(X)\right)\right)\right. \\ &\quad -(\lambda+1)\exp[C(\theta_1+\omega)-C(\theta_1)-C(\omega)]\times E_{\theta_1+\omega}\left(\exp\left(h(X)\right)\right) \\ &\quad \left.+\lambda\exp[C(\theta_2+\omega)-C(\theta_2)-C(\omega)]\times E_{\theta_2+\omega}\left(\exp\left(h(X)\right)\right)\right\}, \end{aligned} \tag{38}$$

with $\theta_1 = (\theta_{11}, \theta_{12}) = (\Sigma_1^{-1}\mu_1, -\frac{1}{2}\Sigma_1^{-1})$, $\theta_2 = (\theta_{21}, \theta_{22}) = (\Sigma_2^{-1}\mu_2, -\frac{1}{2}\Sigma_2^{-1})$, $\omega = (\omega_1, \omega_2) = (\Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1})$ and

$$M_{C,\lambda}^{(2)}(\theta_1, \theta_2, \omega) = \lambda C(\theta_2) - (\lambda+1)C(\theta_1) - C(\omega) + C((\lambda+1)\theta_1 - \lambda\theta_2 + \omega). \quad (39)$$

Based on (22),

$$\begin{aligned} C(\theta_i) &= \log\left((2\pi)^{k/2}|\Sigma_i|^{1/2}\right) + \frac{1}{2}\mu_i^t\Sigma_i^{-1}\mu_i, i = 1, 2 \\ C(\omega) &= \log\left((2\pi)^{k/2}|\Sigma|^{1/2}\right) + \frac{1}{2}\mu^t\Sigma^{-1}\mu. \end{aligned} \quad (40)$$

On the other hand,

$$\theta_1 + \omega = \left(\Sigma_1^{-1}\mu_1 + \Sigma^{-1}\mu, -\frac{1}{2}(\Sigma_1^{-1} + \Sigma^{-1})\right),$$

and it is immediate to see, by means of (22), that

$$\begin{aligned} C(\theta_1 + \omega) &= \log\left((2\pi)^{k/2}|\Sigma^{-1} + \Sigma_1^{-1}|^{-1/2}\right) \\ &\quad + \frac{1}{2}\left(\Sigma^{-1}\mu + \Sigma_1^{-1}\mu_1\right)^t\left(\Sigma^{-1} + \Sigma_1^{-1}\right)^{-1}\left(\Sigma^{-1}\mu + \Sigma_1^{-1}\mu_1\right). \end{aligned} \quad (41)$$

Taking into account the identity (cf. Pardo 2006, p. 49)

$$\begin{aligned} &\left(\Sigma^{-1}\mu + \Sigma_i^{-1}\mu_i\right)^t\left(\Sigma^{-1} + \Sigma_i^{-1}\right)^{-1}\left(\Sigma^{-1}\mu + \Sigma_i^{-1}\mu_i\right) - \mu^t\Sigma^{-1}\mu - \mu_i^t\Sigma_i^{-1}\mu_i \\ &\quad = -(\mu - \mu_i)^t(\Sigma + \Sigma_i)^{-1}(\mu - \mu_i), \quad i = 1, 2, \end{aligned}$$

straightforward algebra entails that

$$\begin{aligned} C(\theta_i + \omega) - C(\theta_i) - C(\omega) &= \log\left((2\pi)^{-k/2}|\Sigma^{-1} + \Sigma_i^{-1}|^{-1/2}|\Sigma|^{-1/2}|\Sigma_i|^{-1/2}\right) \\ &\quad - \frac{1}{2}(\mu - \mu_i)^t(\Sigma + \Sigma_i)^{-1}(\mu - \mu_i), \quad i = 1, 2. \end{aligned} \quad (42)$$

It remains to evaluate $M_{C,\lambda}^{(2)}(\theta_1, \theta_2, \omega)$, given by (39). It is easy to see that

$$\begin{aligned} (\lambda+1)\theta_1 - \lambda\theta_2 + \omega = \Big(&(\lambda+1)\Sigma_1^{-1}\mu_1 - \lambda\Sigma_2^{-1}\mu_2 + \Sigma^{-1}\mu, \\ &(-1/2)\left((\lambda+1)\Sigma_1^{-1} - \lambda\Sigma_2^{-1} + \Sigma^{-1}\right)\Big), \end{aligned}$$

and therefore

$$\begin{aligned}C((\lambda+1)\theta_1-\lambda\theta_2+\omega)=\log\left((2\pi)^{k/2}|(\lambda+1)\Sigma_1^{-1}-\lambda\Sigma_2^{-1}+\Sigma^{-1}|^{-1/2}\right)\\ +\frac{1}{2}\left((\lambda+1)\Sigma_1^{-1}\mu_1-\lambda\Sigma_2^{-1}\mu_2+\Sigma^{-1}\mu\right)^t\\ \times\left((\lambda+1)\Sigma_1^{-1}-\lambda\Sigma_2^{-1}+\Sigma^{-1}\right)^{-1}\\ \times\left((\lambda+1)\Sigma_1^{-1}\mu_1-\lambda\Sigma_2^{-1}\mu_2+\Sigma^{-1}\mu\right),\end{aligned}\tag{43}$$

with $(\lambda+1)\Sigma_1^{-1}-\lambda\Sigma_2^{-1}+\Sigma^{-1}>0$, for $\lambda\neq 0,-1$. Based now on (39), (40) and (43),

$$\begin{aligned}M_{C,\lambda}^{(2)}(\theta_1,\theta_2,\omega)=\log\left((2\pi)^{-k/2}\right)|\Sigma|^{-\frac{1}{2}}|\Sigma_1|^{-\frac{\lambda+1}{2}}|\Sigma_2|^{\frac{\lambda}{2}}\\ \times\left|(\lambda+1)\Sigma_1^{-1}-\lambda\Sigma_2^{-1}+\Sigma^{-1}\right|^{-\frac{1}{2}}\\ -\frac{1}{2}\left(\mu^t\Sigma^{-1}\mu+(\lambda+1)\mu_1^t\Sigma_1^{-1}\mu_1-\lambda\mu_2^t\Sigma_2^{-1}\mu_2-B_1^tB_2B_1\right),\end{aligned}\tag{44}$$

with

$$\begin{aligned}B_1&=(\lambda+1)\Sigma_1^{-1}\mu_1-\lambda\Sigma_2^{-1}\mu_2+\Sigma^{-1}\mu,\\ B_2&=\left((\lambda+1)\Sigma_1^{-1}-\lambda\Sigma_2^{-1}+\Sigma^{-1}\right)^{-1}.\end{aligned}$$

Taking into account that $h(X)=0$ (cf. Eq. (22)), the result follows as an application of (38), (42) and (44). □

Appendix 4

This appendix provides a detailed proof of Proposition 5.

Proof of Proposition 5. (a) Based on (23) and taking into account (12), straightforward algebra leads to the desired result.
(b) Based on part (a) and on Eq. (22),

$$\begin{aligned}D_0^{(\mu,\Sigma)}((\mu_1,\Sigma_1),(\mu_2,\Sigma_2))=\exp\{C(\theta_1+\omega)-C(\theta_1)-C(\omega)\}\\ \times\left(C(\theta_2)-C(\theta_1)+(\theta_1-\theta_2)^tE_{\theta_1+\omega}(T(X))\right)\\ -\exp\{C(\theta_1+\omega)-C(\theta_1)-C(\omega)\}\\ +\exp\{C(\theta_2+\omega)-C(\theta_2)-C(\omega)\},\end{aligned}\tag{45}$$

with

$$\theta_i=\left(\Sigma_i^{-1}\mu_i,-\frac{1}{2}\Sigma_i^{-1}\right),i=1,2,\ \omega=\left(\Sigma^{-1}\mu,-\frac{1}{2}\Sigma^{-1}\right)\text{ and }T(X)=\left(X,XX^t\right). \tag{46}$$

Simple algebraic manipulations lead to,

$$C(\theta_2)-C(\theta_1)=\frac{1}{2}\left(\log\frac{|\Sigma_2|}{|\Sigma_1|}+\mu_2^t\Sigma_2^{-1}\mu_2-\mu_1^t\Sigma_1^{-1}\mu_1\right). \tag{47}$$

On the other hand, taking into account that

$$E_{\theta_1+\omega}(X)=\int_{\mathcal{X}}xf_C(x,\theta_1+\omega)d\mu(x),$$

Eq. (46) entails,

$$E_{\theta_1+\omega}(X)=\left(\Sigma^{-1}+\Sigma_1^{-1}\right)^{-1}\left(\Sigma^{-1}\mu+\Sigma_1^{-1}\mu_1\right). \tag{48}$$

Then,

$$\begin{aligned}E_{\theta_1+\omega}\left(XX^t\right)&=Var_{\theta_1+\omega}(X)+\left(E_{\theta_1+\omega}(X)\right)\left(E_{\theta_1+\omega}(X)\right)^t\\&=(\Sigma^{-1}+\Sigma_1^{-1})^{-1}+\left\{(\Sigma^{-1}+\Sigma_1^{-1})^{-1}\left(\Sigma^{-1}\mu+\Sigma_1^{-1}\mu_1\right)\right.\\&\quad\left.\times\left(\Sigma^{-1}\mu+\Sigma_1^{-1}\mu_1\right)^t(\Sigma^{-1}+\Sigma_1^{-1})^{-1}\right\}\end{aligned} \tag{49}$$

and

$$\theta_1-\theta_2=\left(\Sigma_1^{-1}\mu_1-\Sigma_2^{-1}\mu_2,-\frac{1}{2}\left(\Sigma_1^{-1}-\Sigma_2^{-1}\right)\right). \tag{50}$$

Based on (46) and (50), algebraic manipulations entail that,

$$\begin{aligned}(\theta_1-\theta_2)^tE_{\theta_1+\omega}\left(T(X)\right)&=\left(\Sigma_1^{-1}\mu_1-\Sigma_2^{-1}\mu_2\right)^tE_{\theta_1+\omega}\left(X\right)\\&\quad+trace\left\{-\frac{1}{2}\left(\Sigma_1^{-1}-\Sigma_2^{-1}\right)^tE_{\theta_1+\omega}\left(XX^t\right)\right\},\end{aligned} \tag{51}$$

Hence, taking into account (48), (49) and (51)

$$\begin{aligned}(\theta_1-\theta_2)^tE_{\theta_1+\omega}\left(T(X)\right)&=\left(\Sigma_1^{-1}\mu_1-\Sigma_2^{-1}\mu_2\right)^t\mu^*\\&\quad-\frac{1}{2}\left\{trace\left(\left(\Sigma_1^{-1}-\Sigma_2^{-1}\right)\left(\Sigma^{-1}+\Sigma_1^{-1}\right)^{-1}\right)\right.\\&\quad\left.+(\mu^*)^t\left(\Sigma_1^{-1}-\Sigma_2^{-1}\right)\mu^*\right\},\end{aligned} \tag{52}$$

with $\mu^* = E_{\theta_1+\omega}(X)$. Based on the fact that, for $i = 1, 2$,

$$\begin{aligned}\exp\left(C(\theta_i+\omega) - C(\theta_i) - C(\omega)\right) &= (2\pi)^{-\frac{k}{2}} |\Sigma|^{-\frac{1}{2}} |\Sigma_i|^{-\frac{1}{2}} \left|\Sigma^{-1} + \Sigma_i^{-1}\right|^{-\frac{1}{2}} \\ &\quad \times \exp\left\{-\frac{1}{2}(\mu-\mu_i)^t (\Sigma+\Sigma_i)^{-1}(\mu-\mu_i)\right\} \\ &= E_{(\mu_i,\Sigma_i)}\left(f_{N(\mu,\Sigma)}(X)\right),\end{aligned}$$

Eqs. (42), (45), (47) and (52) complete the proof of part (b) of the proposition. □

References

Ali SM, Silvey SD (1966) A general class of coefficients of divergence of one distribution from another. J R Stat Soc Ser B 28:131–142

Basu A, Shioya H, Park C (2011) Statistical inference, the minimum distance approach. Chapman & Hall/CRC, Boca Raton

Cressie N, Read TRC (1984) Multinomial goodness-of-fit tests. J R Stati Soc Ser B 46:440–464

Csiszár I (1963) Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitat von Markoffschen Ketten. Magyar Tud Akad Mat Kutato Int Kozl 8:85–108

Csiszár I (1967) Information-type measures of difference of probability distributions and indirect observations. Stud Sci Math Hung 2:299–318

Csiszár I, Korner J (1981) Information theory. Coding theorems for discrete memoryless systems. Akademiai Kiado, Budapest

Ebrahimi N, Soofi S, Soyer R (2010) Information measures in perspective. Int Stat Rev 78:383–412

Johnson RA, Wichern DW (1992) Applied multivariate statistical analysis, 3rd edn. Prentice Hall International Editions, Englewood Cliffs

Kagan AM (1963) On the theory of Fisher's information quantity. Dokl Akad Nauk SSSR 151:277–278

Kullback S (1959) Information theory and statistics. Wiley, New York

Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22:79–86

Landaburu E, Morales D, Pardo L (2005) Divergence-based estimation and testing with misclassified data. Stat Pap 46:397–409

Landaburu E, Pardo L (2000) Goodness of fit tests with weights in the classes based on (h, ϕ)-divergences. Kybernetika 36:589–602

Landaburu E, Pardo L (2003) Minimum (h, ϕ) -divergences estimators with weights. Appl Math Comput 140:15–28

Liese F, Vajda I (1987) Convex statistical distances. Teubner Texts in Mathematics, Leipzig

Liese F, Vajda I (2006) On divergences and informations in statistics and information theory. IEEE Trans Inf Theory 52:4394–4412

McElroy T, Holan S (2009) A local spectral approach for assessing time series model misspecification. J Multivar Anal 100:604–621

Morales D, Pardo L, Vajda I (2000) Rényi statistics in directed families of exponential experiments. Statistics 34:151–174

Morales D, Pardo L, Pardo MC, Vajda I (2004) Ré nyi statistics for testing composite hypotheses in general exponential models. Statistics 38:133–147

Nielsen F, Nock R (2011) On Rényi and Tsallis entropies and divergences for exponential families. arXiv:1105.3259v1 [cs.IT] 17 May 2011

Papaioannou T (1986) Measures of information. In: Kotz S, Johnson NL (eds) Encyclopedia of statistical sciences, vol 5. Wiley, New York, pp 391–397

Papaioannou T (2001) On distances and measures of information: a case of diversity. In: Charalambides CA, Koutras MV, Balakrishnan N (eds) Probability and statistical models with applications. Chapman & Hall/CRC, Boca Raton, pp 503–515

Pardo L (2006) Statistical inference based on divergence measures. Chapman & Hall/CRC, Boca Raton

Rényi A (1960) On measures of entropy and information. In: Proceedings of the 4th Berkeley symposium on mathematical statistics and probability, vol 1, Berkeley, pp 547–561

Soofi E (2000) Principal information theoretic approaches. J Am Stat Assoc 95:1349–1353

Soofi E, Retzer JJ (2002) Information indices: unification and applications. Information and entropy econometrics. J Econom 107:17–40

Stummer W, Vajda I (2010) On divergences of finite measures and their applicability in statistics and information theory. Statistics 44:169–187

Ullah A (1996) Entropy, divergence and distance measures with econometric applications. J Stat Plan Inference 49:137–162

Vajda I (1972) On the f-divergence and singularity of probability measures. Period Math Hung 2(1–4):223–234

Vajda I (1973) χ^{α}-divergence and generalized Fisher's information. Transactions of the sixth Prague conference on information theory. Statistical decision functions, random processes, pp 873–886

Vajda I (1989) Theory of statistical inference and information. Kluwer Academic Publishers, Dordrecht

Vajda I (1995) Information theoretic methods in tatistics. Research report no, 1834, Academy of Sciences of the Czech Republic. Institute of Information Theory and Automation, Prague

Zografos K (2008) On Mardia's and Song's measures of kurtosis in elliptical distributions. J Multivar Anal 99:858–879

Zografos K, Nadarajah S (2005) Expressions for Rényi and Shannon entropies for multivariate distributions. Stat Probab Lett 71:71–84