Minimum divergence estimators, a proof of the duality formula

Michel Broniatowski

Minimum divergence estimators, a proof of the duality formula

Michel Broniatowski
LSTA Université Pierre et Marie Curie
5 Place Jussieu, 75005 Paris, France
e-mail: michel.broniatowski@upmc.fr

Abstract

This note provides a simple proof of the dual representation of the divergence between two distributions in a parametric model in the vein of Liese and Vajda (2006) and Broniatowski and Kéziou (2006). Resulting estimators for the divergence as for the parameter are derived. These estimators do not make use of any grouping nor smoothing.

1 Introduction

1.1 Context and scope of this note

This note presents a short proof of the duality formula for $\varphi-$ divergences in parametric models and discusses some unexpected phenomenon in the context of exponential families. First versions of this formula appear in [1] in the context of the Kullback-Leibler divergence and in [7] in a general form. The paper [3] introduces this form in the context of minimal χ^2- estimation; a global approach to this formulation is presented in Broniatowski and Kéziou (2006)[2]. Independently Liese and Vajda (2006)[9] have obtained a similar expression based on a much simpler argument as presented in all the above mentioned papers (formula (118) in their paper); however the proof of their result is merely sketched and we have found it useful to present a complete treatment of this interesting result. The context of this note is restricted to the parametric setting, in contrast with the aforementioned approaches.

The main interest of the resulting expression is that it leads immediately to a wide variety of estimators, by a plug in method of the empirical measure evaluated on the current data set; so, for any type of sampling its estimators and inference procedures, for any $\varphi-$divergence criterion.

In the case of the simple i.i.d. sampling resulting properties of those estimators and subsequent inferential procedures are studied in [4].

This note results from joint cooperation with late Igor Vajda.

1.2 Notation

Let $\mathcal{P} := \{P_\theta, \theta \in \Theta\}$ an identifiable parametric model on $\mathbb{R}^d$ where Θ is a subset of $\mathbb{R}^s$. All measures in $\mathcal{P}$ will be assumed to be measure equivalent sharing therefore the same support. The parameter space Θ need not be open in the present setting. It may even happen that the model includes measures which would not be probability distributions; cases of interest cover models including mixtures of probability distributions; see [4]. Let φ be a proper closed convex function from $]-\infty, +\infty[$ to $[0, +\infty]$ with $\varphi(1) = 0$ and such that its domain $\text{dom}\varphi := \{x \in \mathbb{R} \text{ such that } \varphi(x) < \infty\}$ is an interval with endpoints $a_\varphi < 1 < b_\varphi$ (which may be finite or infinite). For two measures P_α and P_θ in $\mathcal{P}$ the φ-divergence between Q and P is defined by

$$\phi(\alpha, \theta) := \int_{\mathcal{X}} \varphi\left(\frac{dP_\alpha}{dP_\theta}(x)\right) \, dP_\theta(x).$$

In a broader context, the φ-divergences were introduced by [5] as "f-divergences". The basic property of $\varphi-$ divergences states that when φ is strictly convex on a neighborhood of $x = 1$, then

$$\phi(\alpha, \theta) = 0 \text{ if and only if } \alpha = \theta.$$

We refer to [8] chapter 1 for a complete study of those properties. Let us simply quote that in general $\phi(\alpha, \theta)$ and $\phi(, \theta, \alpha)$are not equal. Hence, φ-divergences usually are not distances, but they merely measure some difference between two measures. A main feature of divergences between distributions of random variables X and Y is the invariance property with respect to common smooth change of variables.

1.3 Examples of φ-divergences

The Kullback-Leibler (KL), modified Kullback-Leibler (KL_m), χ^2, modified χ^2 (χ^2_m), Hellinger (H), and L_1 divergences are respectively associated to the convex functions $\varphi(x) = x \log x - x + 1$, $\varphi(x) = -\log x + x - 1$, $\varphi(x) = \frac{1}{2}(x-1)^2$, $\varphi(x) = \frac{1}{2}(x-1)^2/x$, $\varphi(x) = 2(\sqrt{x}-1)^2$ and $\varphi(x) = |x-1|$. All these divergences except the L_1 one, belong to the class of the so called "power divergences" introduced in [6] (see also [8] chapter 2). They are defined through the class of convex functions

$$x \in]0, +\infty[\mapsto \varphi_\gamma(x) := \frac{x^\gamma - \gamma x + \gamma - 1}{\gamma(\gamma - 1)} \tag{1}$$

if $\gamma \in \mathbb{R} \setminus \{0,1\}$, $\varphi_0(x) := -\log x + x - 1$ and $\varphi_1(x) := x \log x - x + 1$. So, the KL-divergence is associated to φ_1, the KL_m to φ_0, the χ^2 to φ_2, the χ^2_m to φ_{-1} and the Hellinger distance to $\varphi_{1/2}$.
Power divergences with index γ in $(0,1)$ are related to the Rényi divergences

$$\begin{aligned} R_\gamma(Q,P) &:= \log H_\gamma(Q,P) \text{ when defined} \\ &= +\infty \text{ otherwise} \end{aligned}$$

where $H_\gamma(Q,P) := \int \left(\frac{dQ}{dP}\right)^\gamma dP$ is the Hellinger integral of order γ.
It may be convenient to extend the definition of the power divergences in such a way that $\phi(\alpha,\theta)$ may be defined (possibly infinite) even when P_α or P_θ is not a probability measure. This is achieved setting

$$x \in]-\infty,+\infty[\mapsto \begin{cases} \varphi_\gamma(x) & \text{if } x \in [0,+\infty[, \\ +\infty & \text{if } x \in]-\infty,0[. \end{cases} \tag{2}$$

when $\mathrm{dom}\varphi = \mathbb{R}^+ / \{0\}$. Note that for the χ^2-divergence, the corresponding φ function $\phi_2(x) := \frac{1}{2}(x-1)^2$ is defined and convex on whole $\mathbb{R}$.

Besides convexity the divergence function φ is assumed to satisfy

(**RC**) There exists a positive δ such that for all c in $[1-\delta, 1+\delta]$, we can find numbers c_1, c_2, c_3 such that $\varphi(cx) \le c_1\varphi(x) + c_2|x| + c_3$, for all real x.

Condition (**RC**) holds for all power divergences including KL and KL_m divergences.

2 Dual form of the divergence in parametric models

We consider differentiable divergences φ which we assume to satisfy the regularity conditions **RC.**

By strict convexity, for all a and b the domain of φ it holds

$$\varphi(b) \ge \varphi(b) + \varphi'(a)(b-a) \tag{3}$$

with equality if and only if $a = b$.

Denote

$$\varphi^\#(x) := x\varphi'(x) - \varphi(x).$$

Select two values α, θ of the parameter and denote

$$a := \frac{dP_\theta}{dP_\alpha}(x)$$

and

$$b := \frac{dP_\theta}{dP_{\theta_T}}(x).$$

Inserting these values in (3) and integrating with respect to P_{θ_T} yields

$$\phi(\theta, \theta_T) \geq \int \left[\varphi' \left(\frac{dP_\theta}{dP_\alpha} \right) dP_\theta - \varphi^\# \left(\frac{dP_\theta}{dP_\alpha} \right) \right] dP_{\theta_T}.$$

Assume at present that this entails

$$\phi(\theta, \theta_T) \geq \int \varphi' \left(\frac{dP_\theta}{dP_\alpha} \right) dP_\theta - \int \varphi^\# \left(\frac{dP_\theta}{dP_\alpha} \right) dP_{\theta_T}. \tag{4}$$

for suitable α's in some set $\mathcal{F}_\theta$ included in Θ.

When $\alpha = \theta_T$ the inequality in (4) turns to equality, which yields

$$\phi(\theta, \theta_T) = \sup_{\alpha \in \mathcal{F}_\theta} \int \varphi' \left(\frac{dP_\theta}{dP_\alpha} \right) dP_\theta - \int \varphi^\# \left(\frac{dP_\theta}{dP_\alpha} \right) dP_{\theta_T} \tag{5}$$

Denote

$$h(\theta, \alpha, x) := \int \varphi' \left(\frac{dP_\theta}{dP_\alpha} \right) dP_\theta - \varphi^\# \left(\frac{dP_\theta}{dP_\alpha} \right) \tag{6}$$

from which

$$\phi(\theta, \theta_T) = \sup_{\alpha \in \mathcal{F}_\theta} \int h(\theta, \alpha, x) dP_{\theta_T}. \tag{7}$$

Furthermore by (4), for all suitable α

$$\begin{aligned} &\phi(\theta, \theta_T) - \int h(\theta, \alpha, x) dP_{\theta_T}. \\ &= \int h(\theta, \theta_T, x) dP_{\theta_T}. - \int h(\theta, \alpha, x) dP_{\theta_T}. \geq 0 \end{aligned}$$

and the function $x \to h(\theta, \theta_T, x) - h(\theta, \alpha, x)$ is non negative, due to (3). It follows that $\phi(\theta, \theta_T) - \int h(\theta, \alpha, x) dP_{\theta_T}$ is zero only if $h(\theta, \alpha, x) = h(\theta, \theta_T, x)$– P_{θ_T} a.e. Therefore for any x in the support of P_{θ_T}

$$\left[\int \varphi' \left(\frac{dP_\theta}{dP_{\theta_T}} \right) dP_\theta - \int \varphi' \left(\frac{dP_\theta}{dP_\alpha} \right) dP_\theta \right] - \varphi^\# \left(\frac{dP_\theta}{dP_\alpha}(x) \right) + \varphi^\# \left(\frac{dP_\theta}{dP_{\theta_T}}(x) \right) = 0$$

which cannot hold for all x when the functions $\varphi^\# \left(\frac{dP_\theta}{dP_\alpha}(x) \right), \varphi^\# \left(\frac{dP_\theta}{dP_{\theta_T}}(x) \right)$ and 1 are linearly independent, unless $\alpha = \theta_T$. We have proved that θ_T is the unique optimizer in (5).

We have skipped some sufficient conditions which ensure that (4) holds.

Assume that

$$\int \left|\varphi'\left(\frac{dP_\theta}{dP_\alpha}\right)\right| dP_\theta < \infty. \tag{8}$$

Assume further that $\phi(\theta, \theta_T)$ is finite. Since

$$\begin{aligned} -\int \varphi^{\#}\left(\frac{dP_\theta}{dP_\alpha}(x)\right) dP_{\theta_T} &\leq \phi(\theta,\theta_T) - \int \varphi'\left(\frac{dP_\theta}{dP_{\theta_T}}\right) dP_\theta \\ &\leq \phi(\theta,\theta_T) + \int \left|\varphi'\left(\frac{dP_\theta}{dP_{\theta_T}}\right)\right| dP_\theta < +\infty \end{aligned}$$

we obtain

$$\int \varphi^{\#}\left(\frac{dP_\theta}{dP_\alpha}\right) dP_{\theta_T} > -\infty$$

which entails (4). When $\int \varphi^{\#}\left(\frac{dP_\theta}{dP_\alpha}\right) dP_{\theta_T} = +\infty$ then clearly , under (8)

$$\phi(\theta,\theta_T) > \int \varphi'\left(\frac{dP_\theta}{dP_\alpha}\right) dP_\theta - \int \varphi^{\#}\left(\frac{dP_\theta}{dP_\alpha}\right) dP_{\theta_T} = -\infty.$$

We have proved that (5) holds when α satisfies (8).

Sufficient and simple conditions encompassing (8) can be assessed under standard requirements for nearly all divergences. We state the following Lemma (see Liese and Vajda (1987)[8]) and Broniatowski and Kéziou (2006) [2], Lemma 3.2).

Lemma 1 *Assume that **RC** holds and $\phi(\theta, \alpha)$ is finite. Then (8) holds.*

Summing up, we state

Theorem 2 *Let θ belong to Θ and let $\phi(\theta, \theta_T)$ be finite. Let $\mathcal{F}_\theta$ be the subset of all α's in Θ such that $\phi(\theta, \alpha)$ is finite . Then*

$$\phi(\theta,\theta_T) = \sup_{\alpha\in\mathcal{F}_\theta} \int \varphi'\left(\frac{dP_\theta}{dP_\alpha}\right) dP_\theta - \int \varphi^{\#}\left(\frac{dP_\theta}{dP_\alpha}\right) dP_{\theta_T}.$$

Furthermore the sup is reached at θ_T and uniqueness holds.

For the Cressie-Read family of divergences with $\gamma \neq 0, 1$ this representation writes

$$\phi_\gamma(\theta,\theta_T) = \sup_{\alpha\in\mathcal{F}_\theta} \left\{ \frac{1}{\gamma-1}\int \left(\frac{dP_\theta}{dP_\alpha}\right)^{\gamma-1} dP_\theta - \frac{1}{\gamma}\int \left(\frac{dP_\theta}{dP_\alpha}\right)^{\gamma} dP_{\theta_T} - \frac{1}{\gamma(\gamma-1)} \right\}.$$

The set $\mathcal{F}_\theta$ may depend on the choice of the parameter θ. Such is the case for the χ^2 divergence i.e. $\varphi(x) = (x-1)^2/2$, when $p_\theta(x) = \theta \exp(-\theta x) 1_{[0,\infty)}(x)$. In most cases the difficulty of dealing with a specific set $\mathcal{F}_\theta$ depending on θ can be encompassed when

$$\begin{array}{l} \text{There exists a neighborhood } \mathcal{U} \text{ of } \theta_T \text{ for which} \\ \phi(\theta, \theta') \text{ is finite whatever } \theta \text{ and } \theta' \text{ in } \mathcal{U} \end{array} \qquad ((\mathrm{A}))$$

which for example holds in the above case for any θ_T. This simplication deserves to be stated in the next result

Theorem 3 *When (A) holds then*

$$\phi(\theta, \theta_T) = \sup_{\alpha \in \mathcal{U}} \int \varphi' \left(\frac{dP_\theta}{dP_\alpha} \right) dP_\theta - \int \varphi^{\#} \left(\frac{dP_\theta}{dP_\alpha} \right) dP_{\theta_T}.$$

Furthermore the sup is reached at θ_T and uniqueness holds.

Remark 4 *Identifying $\mathcal{F}_\theta$ might be cumbersome. This difficulty also appears in the classical MLE case, a special case of the above statement with divergence function φ_0 ,which assumes that*

$$\int \log p_\theta(x) p_{\theta_T}(x) d\lambda(x) \text{ is finite}$$

for θ in a neighborhood of θ_T.

Under the above notation and hypotheses define

$$T_\theta \left(P_{\theta_T} \right) := \arg \sup_{\alpha \in \mathcal{F}_\theta} \int h(\theta, \alpha, x) dP_{\theta_T}. \tag{9}$$

It then holds

$$T_\theta \left(P_{\theta_T} \right) = \theta_T$$

for all θ_T in Θ. Also let

$$S \left(P_{\theta_T} \right) := \arg \inf_{\theta \in \Theta} \sup_{\alpha \in \mathcal{F}_\theta} \int h(\theta, \alpha, x) dP_{\theta_T}. \tag{10}$$

which also satisfies

$$S \left(P_{\theta_T} \right) = \theta_T$$

for all θ_T in Θ. We thus state

Theorem 5 *When $\phi(\theta, \theta_T)$ is finite for all θ in Θ both functionals $T_\theta \left(P_{\theta_T} \right)$ and $S \left(P_{\theta_T} \right)$ are Fisher consistent for all θ_T in Θ.*

3 Plug in estimators

From (7) simple estimators for θ_T can be defined, plugging any convergent empirical measure in place of P_{θ_T} and taking the infimum in θ in the resulting estimator of $\phi(\theta, \theta_T)$.

In the context of simple i.i.d. sampling, introducing the empirical measure

$$P_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

where the X_i's are i.i.d. r.v's with common unknown distribution P_{θ_T} in $\mathcal{P}$, the natural estimator of $\phi(\theta, \theta_T)$ is

$$\begin{aligned} \phi_n(\theta, \theta_T) &:= \sup_{\alpha \in \mathcal{F}_\theta} \left\{ \int h(\theta, \alpha, x) \, dP_n(x) \right\} \\ &= \sup_{\alpha \in \mathcal{F}_\theta} \int \varphi' \left(\frac{dP_\theta}{dP_\alpha} \right) dP_\theta - \frac{1}{n} \sum_{i=1}^{n} \varphi^{\#} \left(\frac{dP_\theta}{dP_\alpha} (X_i) \right). \end{aligned} \tag{11}$$

Since

$$\inf_{\theta \in \Theta} \phi(\theta, \theta_T) = \phi(\theta_T, \theta_T) = 0$$

the resulting estimator of $\phi(\theta_T, \theta_T)$ is

$$\phi_n(\theta_T, \theta_T) := \inf_{\theta \in \Theta} \phi_n(\theta, \theta_T) = \inf_{\theta \in \Theta} \sup_{\alpha \in \mathcal{F}_\theta} \left\{ \int h(\theta, \alpha, x) \, dP_n(x) \right\}. \tag{12}$$

Also the estimator of θ_T is obtained as

$$\widehat{\theta} := \arg \inf_{\theta \in \Theta} \sup_{\alpha \in \mathcal{F}_\theta} \left\{ \int h(\theta, \alpha, x) \, dP_n(x) \right\}. \tag{13}$$

When (A) holds then $\mathcal{F}_\theta$ may be substituted by $\mathcal{U}$ in the above definitions.

The resulting minimum dual divergence estimators (12) and (13) do not require any smoothing or grouping, in contrast with the classical approach which involves quantization. The paper [4] provides a complete study of those estimates and subsequent inference tools for the usual i.i.d. sample scheme. For all divergences considered here, these estimators are asymptotically efficient in the sense that they achieve the Cramer-Rao bound asymptotically. The case when $\varphi = \varphi_0$ leads to $\widehat{\theta}$ defined as the celebrated Maximum Likelihood Estimator (MLE), in the context of the simple sampling.

4

References

[1] Broniatowski, M. Estimation of the Kullback-Leibler divergence. Math. Methods Statist. 12 (2003), no. 4, 391–409 .

[2] Broniatowski, M.; Keziou, A. Minimization of ■-divergences on sets of signed measures. Studia Sci. Math. Hungar. 43 (2006), no. 4, 403–442.

[3] Broniatowski, M.; Leorato, S. An estimation method for the Neyman chi-square divergence with application to test of hypotheses. J. Multivariate Anal. 97 (2006), no. 6, 1409–1436.

[4] Broniatowski, M. Keziou, A. Parametric estimation and tests through divergences and the duality technique. J. Multivariate Anal. 100 (2009), no. 1, 16–36.

[5] Csiszár, I. Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. (German) Magyar Tud. Akad. Mat. Kutató Int. Közl. 8 1963 85–108.

[6] Read, T. R. C., Cressie, N. A. C. Goodness-of-fit statistics for discrete multivariate data. Springer Series in Statistics. Springer-Verlag, New York, 1988. xii+211 pp. ISBN: 0-387-96682-X

[7] Keziou, A. Dual representation of ■-divergences and applications. C. R. Math. Acad. Sci. Paris 336 (2003), no. 10, 857–862

[8] Liese, F., Vajda, I. Convex statistical distances. Teubner-Texte zur Mathematik [Teubner Texts in Mathematics], 95. BSB B. G. Teubner Verlagsgesellschaft, Leipzig, 1987, ISBN: 3-322-00428-7 .

[9] Liese, F., Vajda, I. On divergences and informations in statistics and information theory. IEEE Trans. Inform. Theory 52 (2006), no. 10, 4394–4412