Statistics and Probability Letters 93 (2014) 27–33

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Minimum divergence estimators, maximum likelihood and exponential families

Michel Broniatowski

LSTA, Université Pierre et Marie Curie, Paris, France

ARTICLE INFO

Article history:
Received 9 April 2014
Received in revised form 15 June 2014
Accepted 15 June 2014
Available online 21 June 2014

Keywords:
Statistical divergence
Minimum divergence estimator
Maximum likelihood
Exponential family

ABSTRACT

The dual representation formula of the divergence between two distributions in a parametric model is presented. Resulting estimators do not make use of any grouping or smoothing. For smooth divergences they all coincide with the MLE on any regular exponential family.

1. Introduction

1.1. Context and scope of this note

This note presents a short proof of the duality formula for φ-divergences defined through differentiable convex functions φ in parametric models and discusses some unexpected phenomena in the context of exponential families. First versions of this formula appear in Liese and Vajda (1987, p. 33), in Broniatowski (2003) in the context of the Kullback–Leibler divergence and in Keziou (2003) in a general form. The paper (Broniatowski and Leorato, 2006) introduces this form in the context of minimal χ^2-estimation; a global approach to this formulation is presented in Broniatowski and Keziou (2006). Independently Liese and Vajda (2006) 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 in the parametric setting, in contrast with the aforementioned approaches.

The main interest of the resulting expression is that it leads 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 φ-divergence criterion. In the case of the simple i.i.d. sampling resulting properties of those estimators and subsequent inferential procedures are studied in Broniatowski and Keziou (2009).

A striking fact is that all minimum divergence estimators defined through this dual formula coincide with the MLE in exponential families. They henceforth enjoy strong optimality under the standard exponential models, leading to estimators different from the MLE under different models than the exponential one. Also this result proves that MLE's of parameters of exponential families are strongly motivated by being generated by the whole continuum of φ-divergences. It also has the consequence that robustness properties (or any property not shared by the MLE) cannot be expected by minimum divergence estimators, but for models different from exponential ones.

This note results from joint cooperation with late Igor Vajda.

E-mail address: michel.broniatowski@upmc.fr.

http://dx.doi.org/10.1016/j.spl.2014.06.014
0167-7152/

1.2. Notation

Let $\mathcal{P} := \{P_\theta, \theta \in \Theta\}$ be an identifiable parametric model on $\mathbb{R}^s$ where Θ is a subset of $\mathbb{R}^d$. All measures in $\mathcal{P}$ will be assumed to be equivalent measures 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 Broniatowski and Keziou (2009). Let φ be a proper closed convex function from $]-\infty, +\infty[$ to $[0, +\infty]$ with $\varphi(1) = 0$ and such that its domain $\operatorname{dom}\varphi := \{x \in \mathbb{R}$ 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

$$\phi(\alpha, \theta) := \int_{\mathbb{R}^s} \varphi\left(\frac{dP_\alpha}{dP_\theta}(x)\right) dP_\theta(x).$$

In a broader context, the φ-divergences were introduced by Csiszár (1963) as "f-divergences". The basic property of φ-divergences states that when φ is strictly convex on a neighborhood of $x = 1$, then

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

We refer to Liese and Vajda (1987, 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 differences 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 Read and Cressie (1988) (see also Liese and Vajda, 1987, Chapter 2), a class which takes its origin from Rényi (1961). 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 $\varphi_{1/2}$.

1.4. Hypotheses

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 $\operatorname{dom}\varphi = \mathbb{R}^+/\{0\}$. Note that for the χ^2-divergence, the corresponding φ function $\varphi_2(x) := \frac{1}{2}(x-1)^2$ is defined and convex on whole $\mathbb{R}$; an example when $\mathcal{P}$ may contain signed finite measures and not be restricted to probability measures is considered in Broniatowski and Leorato (2006) in relation with a two components mixture model $\mathcal{P} := \{P = \theta P_1 + (1-\theta)P_2\}$ with P_1 and P_2 two known probability measures and where θ is allowed to assume values in an open neighborhood Θ of 1, in order to provide a test for $H0 := \{\theta = 1\}$ vs $H1 := \{0 \le \theta < 1\}$, with θ an interior point in Θ.

Considering any φ-divergence with φ a differentiable function but the likelihood divergence defined through the divergence function φ_0, when θ_T in $\operatorname{int}\Theta$ is defined as the true parameter of the distribution of the i.i.d. sample $(X_1, \ldots, X_n)$ it is convenient to assume that

$$\text{There exists a neighborhood } \mathcal{U} \text{ of } \theta_T \text{ for which } \phi(\theta, \theta') \text{ is finite whatever } \theta \text{ and } \theta' \text{ in } \mathcal{U}. \tag{A}$$

In the case when $\varphi = \varphi_0$, which yields the MLE, then the condition (A) is reduced to

$$\text{There exists a neighborhood } \mathcal{U} \text{ of } \theta_T \text{ for which } \int (\log p_\theta)\, p_{\theta_T} d\lambda \text{ is finite whatever } \theta \text{ in } \mathcal{U} \tag{A'}$$

where λ dominates $\mathcal{P}$, which is the classical assumption for the existence and consistency of the MLE under simple sampling.

We will only consider divergences defined through differentiable functions φ, which we assume to satisfy

$$(\mathbf{RC}) \quad \begin{array}{l} \text{There exists a positive } \delta \text{ such that for all } c \text{ in } [1-\delta, 1+\delta], \\ \text{we can find numbers } c_1, c_2, c_3 \text{ such that} \\ \varphi(cx) \le c_1 \varphi(x) + c_2 |x| + c_3, \text{ for all real } x. \end{array}$$

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