

Minimum divergence in inference and testing

*Alcuni aspetti delle tecniche delle divergenze minime
nell' inferenza statistica e nella teoria dei test*

Michel Broniatowski¹²

Dipartimento LSTA

Università Paris VI e Università di Reims

mbr@ccr.jussieu.fr

Riassunto: In questo testo si presentano alcuni approcci nuovi per l'inferenza ed i test sfruttando le proprietà di convessità delle divergenze fra leggi di probabilità. Viene proposto un' approccio unitario per lo studio di vari criteri statistici. Vengono anche illustrati alcuni esempi in ambito parametrico e semi parametrico.

Keywords: divergences, statistical criteria.

1. Introduction and notation

Let $(\mathcal{X}, \mathcal{B})$ be a measurable space. Let φ be a non-negative convex function defined from $(0, +\infty)$ onto $[0, \infty]$ and satisfying $\varphi(1) = 0$. Also extend φ at 0 defining $\varphi(0) = \lim_{x \rightarrow 0^+} \varphi(x)$. Let P be a probability measure (p.m.) defined on $(\mathcal{X}, \mathcal{B})$. For any p.m. Q absolutely continuous (a.c.) w.r.t. P , the ϕ -divergence between Q and P is defined by

$$\phi(Q, P) := \int \varphi \left(\frac{dQ}{dP} \right) dP. \quad (1)$$

When Q is not a.c. w.r.t. P , we set $\phi(Q, P) := +\infty$. Therefore $Q \rightarrow \phi(Q, P)$ is defined on the whole class of all probability measures on $(\mathcal{X}, \mathcal{B})$.

For any p.m. P , the mapping $Q \rightarrow \phi(Q, P)$ is convex and non-negative. When $Q = P$, the ϕ -divergence between Q and P is zero. When the function $x \rightarrow \varphi(x)$ is a strictly convex function in $x = 1$, then the fundamental property holds

$$\phi(Q, P) = 0 \text{ if and only if } Q = P.$$

We refer to Csiszár (1967) and to Liese and Vajda (1987) chapter 1, for the proofs of these properties.

In general $\phi(Q, P)$ and $\phi(P, Q)$ are not equal. Indeed, they coincide if and only if there exists some real number c such that for any positive x , it holds $\varphi(x) - x\varphi(1/x) = c(x - 1)$ (see Liese and Vajda (1987) Theorem 1.13). Hence ϕ -divergences usually are not distances, but they merely measure some difference between two p.m.'s. Of course a main feature of divergences between distributions of r.v's X and Y is the invariance property with respect to any common change of variables.

¹Address for correspondence: LSTA, Université Paris VI, 175 rue du Chevaleret
75013 Paris - FRANCIA

²Work supported by Progetto dell'Ateneo di Padova, coordinator Giorgio Celant.

1.1. Examples of ϕ -divergences.

Some divergences are widely used in statistics and in Information theory. They are the Kullback-Leibler (KL) and the modified KL (KL_m) divergences, defined as follows

$$KL(Q, P) := \begin{cases} \int \log \left(\frac{dQ}{dP} \right) dQ & \text{if } Q \text{ is a.c. w.r.t. } P, \\ +\infty & \text{otherwise} \end{cases}$$

$$KL_m(Q, P) := \begin{cases} \int -\log \left(\frac{dQ}{dP} \right) dP & \text{if } Q \text{ is a.c. w.r.t. } P, \\ +\infty & \text{otherwise.} \end{cases}$$

The first one corresponds to $\varphi(x) = x \log(x) - x + 1$, and the second one to $\varphi(x) = -\log(x) + x - 1$. The above examples are peculiar cases of the so-called “power divergences”, introduced by Cressie and Read (1984) (see also Liese and Vajda (1987) chapter 2), which are defined by the class of functions

$$x \in \mathbb{R}_+ \rightarrow \varphi_\gamma(x) := \frac{x^\gamma - \gamma x + \gamma - 1}{\gamma(\gamma - 1)} \quad (2)$$

for γ in $\mathbb{R}$, $\gamma \neq 0, 1$ and $\varphi_0(x) := -\log x + x - 1$, $\varphi_1(x) := x \log x - x + 1$. The KL divergence is associated to φ_0 , the KL_m to φ_1 , the χ^2 to φ_{-1} , the modified χ^2 to φ_2 , and the Hellinger distance to $\varphi_{1/2}$.

We are interested in estimation and test using such divergences. An i.i.d. sample $X_1, \dots, X_n$ with common unknown distribution P is observed and some p.m. Q is given. We intend to estimate $\phi(Q, P)$ and, more generally, $\inf_{Q \in \Omega} \phi(Q, P)$ where Ω is some set of p.m.’s, as well as the p.m. Q^* achieving the infimum in Ω . In the parametric context, these problems can be well defined and lead to new results in estimation and tests, extending classical notions and improving their properties. In the field of semi parametric statistics it provides new tools for the two sample problem, among others. Also this approach extends the empirical likelihood paradigm and provides estimates and tests for models satisfying linear constraints with unknown parameters.

1.2. Statistical examples and motivation

1.2.1. Tests of fit

Let Q_0 and P be two p.m.’s with same finite discrete support S . It holds

$$\phi(Q_0, P) = \sum_{i \in S} \varphi(Q_0(i)/P(i))P(i)$$

which can then be estimated via plug-in, setting

$$\hat{\phi}_n(Q_0, P) := \phi(Q_0, P_n) = \sum_{i \in S} \varphi(Q_0(i)/P_n(i))P_n(i),$$

where P_n is the empirical measure pertaining to the sample $X_1, \dots, X_n$ with distribution P .

In this vein, goodness of fit tests have been proposed by Cressie and Read (1984), Landaburu and Pardo (2000) for fixed number of classes, and by Györfi and Vajda (2002) when the number of classes depends on the sample size.

1.2.2. Estimation and parametric tests

Let $\{P_\theta, \theta \in \Theta\}$ be some parametric model with Θ a subset of $\mathbb{R}^d$. Assume that all p.m.'s P_θ share the same finite support S and that the support S does not depend upon θ . We then have

$$\phi(P_\theta, P_{\theta_0}) = \sum_{i \in S} \varphi \left(\frac{P_\theta(i)}{P_{\theta_0}(i)} \right) P_{\theta_0}(i).$$

For such models, Lindsay (1994) and Morales *et al.* (1995) introduced the so-called “Minimum ϕ -divergences estimates” (M ϕ E's) (Minimum Disparity Estimators in Lindsay (1994)) of the unknown true value of the parameter, say θ_0 , defined by

$$\hat{\theta}_n := \arg \inf_{\theta \in \Theta} \phi(P_\theta, P_n), \quad (3)$$

where $\phi(P_\theta, P_n)$ is the plug-in estimate of $\phi(P_\theta, P_{\theta_0})$

$$\phi(P_\theta, P_n) = \sum_{i \in S} \varphi \left(\frac{P_\theta(i)}{P_n(i)} \right) P_n(i).$$

Various parametric tests can be performed based on the previous estimates of the ϕ -divergences; see Lindsay (1994) and Morales, Pardo and Vajda (1995). The class of estimates in (3) contains the maximum likelihood estimate. Indeed when $\varphi(x) = -\log(x) + x - 1$, we obtain $\hat{\theta}_n := \arg \inf_{\theta \in \Theta} KL_m(P_\theta, P_n) = \arg \inf_{\theta \in \Theta} \sum_{i \in S} -\log(P_\theta(i)) = \text{MLE}$. When interested in testing $H_0 : \theta_0 = \alpha_0$ against $H_1 : \theta_0 \neq \alpha_0$, we can use the statistics $\phi(P_{\alpha_0}, P_n)$, the plug-in estimate of the divergence between P_{α_0} and P_{θ_0} , rejecting H_0 for large values of the statistics; see Cressie and Read (1984). In the case when $\varphi(x) = -\log(x) + x - 1$, this test does not coincide with the maximum likelihood ratio test, which defined through the Wilks likelihood ratio statistic $ML_n := 2 \log \frac{\sup_{\theta \in \Theta} \prod_{i=1}^n p_\theta(X_i)}{\prod_{i=1}^n p_{\alpha_0}(X_i)}$. The new estimate $\widehat{KL}_m(P_{\alpha_0}, P_{\theta_0})$ of $KL_m(P_{\alpha_0}, P_{\theta_0})$, which is proposed here, leads to the maximum likelihood ratio test.

When the support S is continuous, the estimates in (3) are not defined; Basu and Lindsay (1994) investigate the so-called “minimum disparity estimators” (MDE's) for continuous models; they consider $p_{n,h}(x) := \frac{1}{h} \int K\left(\frac{x-t}{h}\right) dP_n(t)$ the kernel estimate of the density of the X_i 's, and the smoothed version of p_θ , i.e. $p_\theta^*(x) := \frac{1}{h} \int K\left(\frac{x-t}{h}\right) p_\theta(t) dt$. The MDE is then defined as $\arg \min_{\theta \in \Theta} \widehat{\phi}(p_\theta^*, p_{n,h})$ with $\widehat{\phi}(p_\theta^*, p_{n,h}) := \int \varphi\left(\frac{p_\theta^*(t)}{p_{n,h}(t)}\right) p_{n,h}(t) dt$. When $\varphi(x) = -\log(x) + x - 1$, this estimate clearly, due to smoothing, does not coincide with the MLE. Also the test based on $\widehat{KL}_m(p_\theta^*, p_{n,h})$ is different from the likelihood ratio test.

No direct plug-in estimate of $\phi(Q, P)$ can be performed by substitution of P by P_n when Q belongs to some class of p.m.'s a.c. w.r.t. the Lebesgue measure λ . In order to build tests pertaining to the density $p := \frac{dP}{d\lambda}$, Beran (1977) and Berlinet and Van der Meulen (1998) proposed to use the smoothed kernel estimate p_n of p . Beran (1977) handles the Hellinger distance, while Berlinet and Van der Meulen (1998) obtains the limiting distribution of the estimate for the Kullback-Leibler divergence. The extension of their results to other divergences remains an open problem; see Berlinet (1999), Györfi *et al.* (1998), and Berlinet and Van der Meulen (1998). Also it seems difficult to use such methods to obtain the limiting distribution of an estimate of $\inf_{Q \in \Omega} \phi(Q, P)$ when Ω is some

class of p.m.'s; this problem will be treated in the present paper when Ω is a parametric class, avoiding the smoothing method.

When S is infinite or continuous then the plug-in estimate $\phi(P_\theta, P_n)$ usually takes infinite value when no use is done of some partition-based approximation. In Broniatowski (2002), a new estimation procedure is proposed in order to estimate the KL-divergence between some set of p.m.'s Ω and some p.m. P , without making use of any partitioning nor smoothing, but merely making use of the well known dual representation of the KL-divergence as the Fenchel-Legendre transform of the moment generating function.

Extending the paper by Broniatowski (2002), we expose some general representation for ϕ -divergences. This is obtained through the following duality lemma, whose proof can be found for example in Lemma 4.5.8 chapter 4 in Dembo and Zeitouni (1998).

Lemma 1 *Let S be a locally convex Hausdorff topological linear space and let $g : S \rightarrow (-\infty, +\infty]$ be some convex lower semi continuous (l.s.c.) function. Define the Fenchel-Legendre transform of g by*

$$g^*(l) := \sup_{x \in S} \{l(x) - g(x)\}, \quad l \in S^*,$$

where S^* is the topological dual space of S . We then have

$$g(x) = \sup_{l \in S^*} \{l(x) - g^*(l)\},$$

which is to say that g is the Fenchel-Legendre transform of g^* .

When applied in the context of divergences between distributions in a parametrized family the above Lemma is the starting point for the definition of estimates of the parameter θ_0 , which we will call “minimum dual ϕ -divergences estimates” (MD ϕ E's). They are defined in parametric models $\{P_\theta, \theta \in \Theta\}$, where the p.m.'s P_θ do not necessarily have finite support; it can be discrete or continuous, bounded or not. Also the same Lemma will be applied in order to estimate $\phi(P_\alpha, P_{\theta_0})$ and $\inf_{\alpha \in \Theta_0} \phi(P_\alpha, P_{\theta_0})$ where Θ_0 is a subset of Θ , which leads to various simple and composite tests pertaining to θ_0 , the true unknown value of the parameter. When $\varphi(x) = -\log(x) + x - 1$, the MD ϕ E's coincide with the maximum likelihood estimates (see remark 3.2 below); since our approach includes also test procedures, with this peculiar choice for the function φ , we recover the classical likelihood ratio test for simple hypotheses and for composite hypotheses. In the context of parametric estimation it is well known that the maximum likelihood method may lead to non robust estimates, although efficient after bias first order correction. A possible extension of the results presented here leads to minimum divergence estimates which can both meet the requirements of efficiency and robustness. When applied in proportional risk models these techniques provide new solutions for the two sample problem. Also in the context of estimation and tests for models satisfying linear constraints with unknown parameters, as appears in regression analysis or in test of fit, they provide solutions that have common features with the empirical likelihood techniques, extending them to a wide class of criterions.

We sometimes write Pf for $\int f dP$ for any measure P and function f .

2. Duality and ϕ -divergences

Let $(\mathcal{X}, \mathcal{B})$ denote a measurable space on which all the r.v.'s will be defined. Let M be the space of all finite signed measures defined on $(\mathcal{X}, \mathcal{B})$. We also consider a class $\mathcal{F}$ of

measurable real valued functions f defined on $\mathcal{X}$, and we assume that $\mathcal{F}$ contains $\mathcal{M}_b$, the set of all bounded measurable functions defined on $\mathcal{X}$. We will denote $\langle \mathcal{F} \rangle$ the linear span of $\mathcal{F}$. Define

$$M_{\mathcal{F}} := \left\{ Q \in M \text{ such that } \int |f| d|Q| < \infty, \text{ for all } f \text{ in } \mathcal{F} \right\}.$$

We extend the definition in (1) on the whole space $M_{\mathcal{F}}$ by stating $\varphi(x) = +\infty$ for negative values of x whenever $\varphi(x)$ is not defined. Note that for the χ_m^2 -divergence, φ is defined on whole $\mathbb{R}$.

We equip the linear space $M_{\mathcal{F}}$ with the $\tau_{\mathcal{F}}$ -topology, which is the coarsest topology for which all mappings $Q \rightarrow \int f dQ$ are continuous for all f in $\langle \mathcal{F} \rangle$.

The following result provides a first important tool in order to derive a useful representation for the divergence.

Proposition 1 *Equip $M_{\mathcal{F}}$ with the $\tau_{\mathcal{F}}$ -topology. Then $M_{\mathcal{F}}$ is a Hausdorff locally convex topological linear space. Further the topological dual space of $M_{\mathcal{F}}$ is the set of all mappings $Q \rightarrow \int f dQ$ when f belongs to $\langle \mathcal{F} \rangle$.*

In view of the above result we identify the dual space of $M_{\mathcal{F}}$ with the linear span of $\mathcal{F}$, that is with $\langle \mathcal{F} \rangle$.

We state that with ϕ -divergence defined as in (1), the mapping $Q \rightarrow \phi(Q, P)$ defined on $M_{\mathcal{F}}$ satisfies the conditions in Lemma 1. We have

Proposition 2 *The divergence function $Q \rightarrow \phi(Q, P)$ from $(M_{\mathcal{F}}, \tau_{\mathcal{F}})$ onto $(-\infty, +\infty]$ is lower semi continuous.*

2.1. Application of the duality Lemma

We now apply Lemma 1 when the space $\mathcal{S}$ is replaced by $M_{\mathcal{F}}$ and when the function g is defined by

$$\begin{aligned} (M_{\mathcal{F}}, \tau_{\mathcal{F}}) &\rightarrow (-\infty, +\infty] \\ Q &\rightarrow g(Q) = \phi(Q, P), \end{aligned}$$

in which P is an arbitrary p.m. Note that the hypotheses in Lemma 1 hold and that the topological dual space of $M_{\mathcal{F}}$ is one to one with $\langle \mathcal{F} \rangle$, following Proposition 1 and Proposition 2. Also the Fenchel-Legendre transform of $Q \rightarrow \phi(Q, P)$ is defined for any f in $\langle \mathcal{F} \rangle$ by

$$T(f, P) := \sup_{Q \in M_{\mathcal{F}}} \left\{ \int f dQ - \phi(Q, P) \right\}.$$

We thus state

Proposition 3 *For any measure Q in $M_{\mathcal{F}}$ and for any p.m. P , it holds*

$$\phi(Q, P) = \sup_{f \in \langle \mathcal{F} \rangle} \left\{ \int f dQ - T(f, P) \right\}. \quad (4)$$

The following result provides an explicit form for the above formula and is the basis for estimation and test in various contexts.

Theorem 1 Assume that the function φ is strictly convex and is $\mathcal{C}^2$ on $(0, +\infty)$. Let Q and P be two p.m.'s with Q a.c. w.r.t. P and $\phi(Q, P) < \infty$. Let $\mathcal{F}$ be a class of functions such that (i) for all f in $\mathcal{F}$, $\int |f| dQ$ is finite, (ii) $\varphi'(dQ/dP)$ belongs to $\mathcal{F}$ and (iii) for any f in $\mathcal{F}$, $\text{Im} f$ is included in $\text{Im} \varphi'$. We then have

(1) The divergence $\phi(Q, P)$ admits the “dual representation”

$$\phi(Q, P) = \sup_{f \in \mathcal{F}} \left\{ \int f dQ - \int f \overleftarrow{\varphi}'(f) - \varphi \left(\overleftarrow{\varphi}'(f) \right) dP \right\} \quad (5)$$

(2) The supremum in (5) is unique ($P - a.s$) and is reached at $f = \varphi'(dQ/dP)$ ($P - a.s$).

Remark 1 The difference from (4) and (5) lays in the substitution of $f \in \langle \mathcal{F} \rangle$ by $f \in \mathcal{F}$ in the sup. This will prove to be an important feature for statistical applications.

3. Parametric estimation and tests through minimum ϕ -divergence approach

We assume that the function φ is strictly convex and is $\mathcal{C}^2$ on $(0, +\infty)$. We consider an identifiable parametric model $\{P_\theta : \theta \in \Theta\}$ defined on some measurable space $(\mathcal{X}, \mathcal{B})$ and Θ is some subset of $\mathbb{R}^d$. For notational clearness we write $\phi(\alpha, \theta)$ for $\phi(P_\alpha, P_\theta)$ for α and θ in Θ . We assume that for any θ in Θ , P_θ has density p_θ with respect to some dominating σ -finite measure λ , which can be either with countable support or not. Assume further that the support S of the measure P_θ does not depend upon θ . On the basis of an i.i.d. sample $X_1, \dots, X_n$ with distribution P_{θ_0} , we intend to estimate θ_0 the true value of the parameter. We assume that for any α in Θ , the following condition holds

(C.0)

$$\int \left| \varphi' \left(\frac{p_\alpha}{p_\theta} \right) \right| dP_\alpha(x) < \infty, \text{ for any } \theta \in \Theta.$$

This condition is fulfilled if

$$\phi(\alpha, \theta) = \int \varphi \left(\frac{p_\alpha}{p_\theta} \right) dP_\theta < \infty, \text{ for any } \theta \in \Theta, \quad (6)$$

and φ fulfills the condition of Lemma 8.7 in Liese and Vajda (1987); see Liese and Vajda (1987) Lemma 8.9.

Consider the class of functions $\mathcal{F}$ defined by

$$\mathcal{F} := \left\{ x \rightarrow \varphi' \left(\frac{p_\alpha(x)}{p_\theta(x)} \right), \theta \in \Theta \right\}.$$

By Theorem 1, when (C.0) holds, we obtain

$$\phi(\alpha, \theta_0) = \sup_{f \in \mathcal{F}} \left\{ \int f dP_\alpha - \int f \overleftarrow{\varphi}'(f) - \varphi \left(\overleftarrow{\varphi}'(f) \right) dP_{\theta_0} \right\},$$

i.e.

$$\phi(\alpha, \theta_0) = \sup_{\theta \in \Theta} P_{\theta_0} m(\theta, \alpha), \quad (7)$$

with

$$m(\theta, \alpha) : x \rightarrow m(\theta, \alpha, x)$$

and

$$m(\theta, \alpha, x) := \int \varphi' \left(\frac{p_\alpha}{p_\theta} \right) dP_\alpha - \left\{ \varphi' \left(\frac{p_\alpha}{p_\theta}(x) \right) \frac{p_\alpha(x)}{p_\theta} - \varphi \left(\frac{p_\alpha}{p_\theta}(x) \right) \right\}.$$

Remark 2 The function $\theta \rightarrow P_{\theta_0} m(\theta, \alpha)$ has a unique maximizer $\theta = \theta_0$; see Theorem 1 (2).

Let $X_1, \dots, X_n$ be an i.i.d. sample with p.m. P_{θ_0} . For all $\alpha \in \Theta$, define the class of estimates of θ_0 , which we call “dual ϕ -divergence estimates” (D ϕ E’s), by

$$\hat{\theta}_n(\alpha) := \arg \sup_{\theta \in \Theta} P_n m(\theta, \alpha). \quad (8)$$

For any α in Θ , the divergence $\phi(P_\alpha, P_{\theta_0})$ between P_α and P_{θ_0} can be estimated by

$$\hat{\phi}_n(\alpha, \theta_0) := P_n m(\hat{\theta}_n(\alpha), \alpha) = \sup_{\theta \in \Theta} P_n m(\theta, \alpha).$$

Further, we have

$$\inf_{\alpha \in \Theta} \phi(\alpha, \theta_0) = \phi(\theta_0, \theta_0) = 0.$$

The infimum in the above display is unique when φ is strictly convex on a neighborhood of 1, and it is achieved at $\alpha = \theta_0$. It follows that a natural definition of estimates of θ_0 , which we call “minimum dual ϕ -divergence estimates” (MD ϕ E’s), is

$$\hat{\alpha}_n := \arg \inf_{\alpha \in \Theta} \hat{\phi}_n(\alpha, \theta_0) = \arg \inf_{\alpha \in \Theta} \sup_{\theta \in \Theta} P_n m(\theta, \alpha). \quad (9)$$

Remark 3 (An other view at the MLE). The maximum likelihood estimate (MLE) belongs to this class of estimates. Indeed it is obtained when $\varphi(x) = -\log(x) + x - 1$, that is as the dual modified KL-divergence estimate or as the minimum dual modified KL-divergence estimate, i.e. $\text{MLE} = \text{DKL}_m E = \text{MDKL}_m E$. Indeed, we then have $P_n m(\theta, \alpha) = -\int \log \left(\frac{p_\alpha}{p_\theta} \right) dP_n$; by definitions (8) and (9), we get $\hat{\theta}_n(\alpha) = \arg \sup_{\theta \in \Theta} \int \log(p_\theta) dP_n$ independently upon α , and $\hat{\alpha}_n = \arg \inf_{\theta \in \Theta} -\int \log(p_\theta) dP_n$. So the MLE is the estimate of θ that minimizes the estimate of the divergence between the parametric model and the measure P .

This talk will expose a number of results in connection with those estimates. A special attention will be given to confidence areas with small size as can be obtained through a precise choice of the divergence function φ according to the model. Also we will discuss properties of the estimates, indicating new ways to handle in the same frame efficiency and robustness.

4. Case control and semi parametric two sample density ratio models

We aim to introduce a new method in order to give new answers for the following problems : two-sample test for comparing two populations and estimation of the parameters for semiparametric density ratio models. Here we will discuss an extension of the above considerations developed by Kéziou (Université Paris 6) and Leoni (Università di Padova).

We dispose of two samples, $X_1, \dots, X_{n_0}$ (i.i.d. in G) and $Y_1, \dots, Y_{n_1}$ (i.i.d. in H) from two unknown distributions noted G and H , respectively. A semiparametric density ratio model has the form

$$\frac{dH}{dG}(x) = m(\theta_0, x), \quad (10)$$

where θ_0 is the unknown parameter of interest which we suppose to be unique and which belongs to some open set $\Theta \subseteq \mathbb{R}^d$. The function $m(\cdot, \cdot)$ is known and nonnegative.

We now give some statistical examples and motivations for model (10).

4.1. Comparison of two populations

In applications, we often come across with the problem of comparing two samples. Let $X_1, \dots, X_{n_0}$ and $Y_1, \dots, Y_{n_1}$ be two samples from unknown distributions G and H , respectively.

The use of the well known t -test, requires to assume that both samples are normally distributed with unknown means and common known or unknown variance. The t -test enjoys several optimal properties, for example it is the uniformly most powerful unbiased test (see e.g. Lehmann (1986)).

If both H and G are normally distributed with equal variances

$$H = \mathcal{N}(\mu_1, \sigma^2) \text{ and } G = \mathcal{N}(\mu_2, \sigma^2),$$

then, the ratio $\frac{dH}{dG}$ takes the form

$$\frac{dH}{dG}(x) = \exp \{ \alpha + \beta x \}, \quad (11)$$

where

$$\alpha = \frac{\mu_1^2 - \mu_2^2}{2\sigma^2} \quad \text{and} \quad \beta = \frac{\mu_2 - \mu_1}{\sigma^2}.$$

It follows that testing the hypothesis $\mathcal{H}_0 : H = G$ is equivalent to testing the hypothesis $\mathcal{H}_0 : \beta = 0$. We underline that $\beta = 0$ implies $\alpha = 0$.

Kay and Little (1987) and Fokianos (2002) observed that there are cases in which the choice

$$\frac{dH}{dG}(x) = \exp \{ \alpha + \beta r(x) \}, \quad (12)$$

where $r(x)$ is an arbitrary but known function of x , is more appropriate.

Model (10) includes models (11) and (12) by taking $m(\theta, x) = \exp \{ \alpha + \beta x \}$ and $m(\theta, x) = \exp \{ \alpha + \beta r(x) \}$, respectively, and $\theta = (\alpha, \beta)^T$.

In the case when the semiparametric assumption (10) fails, the test commonly used is the Mann-Whitney-Wilcoxon test.

4.2. Logistic model and multiplicative-intercept risk model

Consider the logistic model which has been widely used in statistical applications for the analysis of binary data (see e.g. Hosmer and Lemeshow (2000)). Suppose that y is a binary response variable and that x is the associate covariate vector. The logistic model has the form

$$\Pr(y = 1|x) = \frac{\exp(\gamma + x^T \beta)}{1 + \exp(\gamma + x^T \beta)}, \quad \gamma \in \mathbb{R}, \quad \beta \in \mathbb{R}^{d-1}. \quad (13)$$

Note that the marginal density of x , noted $f(x)$, is left completely unspecified.

In a case-control study the binary outcome variable is fixed by stratification. In this type of study design, two random samples of sizes n_0 and n_1 are chosen from the two strata defined by the outcome variable, i.e., from the subsets of the population with $y = 0$ and $y = 1$, respectively. Assume that $x_1, \dots, x_{n_0}$ are the observed covariates from the control group and let $x_{n_0+1}, \dots, x_n$ ($n = n_0 + n_1$) be those from the case group. We aim to estimate the parameters γ and β using the two samples $X_1, \dots, X_{n_0}$ and $X_{n_0+1}, \dots, X_n$. We show that the logistic model (13) writes in the form of the model (10). So, let f denote the density function of the covariates x , and put

$$\pi = \Pr(y = 1) = \int \Pr(y = 1|x) f(x) dx,$$

and assume that

$$f_i(x) = f(x|y = i) = dF(x|y = i)/dx \quad i = 0, 1$$

exist and represent the conditional density function of x given $y = i$. It is not difficult to manipulate the case-control likelihood function to obtain a logistic regression model in which the dependent variable is the outcome variable of interest to the investigator. The key step in this development is an application of the Bayes Theorem, that yields, letting $\alpha := \gamma + \log\left(\frac{1-\pi}{\pi}\right)$

$$\frac{f_1(x)}{f_0(x)} =: \exp(\alpha + x^T \beta),$$

where $\alpha := \gamma + \log\left(\frac{1-\pi}{\pi}\right)$. Thus, model (13) is equivalent to the following two-sample semiparametric model

$$\begin{aligned} x_1, \dots, x_{n_0} &\sim f(x|y = 0) = f_0(x), \\ x_{n_0+1}, \dots, x_n &\sim f(x|y = 1) = f_1(x) = \exp(\alpha + x^T \beta) f_0(x). \end{aligned} \quad (14)$$

More generally, we can consider the multiplicative-intercept risk model, i.e.,

$$\Pr(y = 1|x) = \frac{\exp(\gamma + r(x, \beta))}{1 + \exp(\gamma + r(x, \beta))}, \quad \gamma \in \mathbb{R}, \quad \beta \in \mathbb{R}^{d-1},$$

where $r(x, \beta)$ is a given function of x and β . In this case, we obtain the following two-sample semiparametric model

$$\begin{aligned} x_1, \dots, x_{n_0} &\sim f_0(x) \\ x_{n_0+1}, \dots, x_n &\sim f_1(x) = \exp(\alpha + r(x, \beta)) \cdot f_0(x). \end{aligned} \quad (15)$$

Models (14) and (15) are particular cases of models (10) by taking $\frac{dG}{dx} = f_0$, $\frac{dH}{dx} = f_1$ and $\theta = (\alpha, \beta^T)^T$. In the context of model (10), estimate of γ and β is equivalent to estimate of α and β .

For models (10), when the samples $X_1, \dots, X_{n_0}$ and $Y_1, \dots, Y_{n_1}$ are independent, Qin (1998) presents an estimate of θ_0 based on the empirical likelihood approach, using the likelihood of the independent variables $X_1, \dots, X_{n_0}, Y_1, \dots, Y_{n_1}$.

An important special case of the case-control study is the matched (or paired) study.

The minimum divergence approach provides a new approach for estimation of the parameter θ_0 and tests of the hypothesis $\mathcal{H}_0 : H = G$ in two-sample semiparametric models of the form (10) with independent or paired samples $X_1, \dots, X_{n_0}$ and $Y_1, \dots, Y_{n_1}$.

In general, in order to compare the laws of two variables X and Y , we estimate some measures of difference (distances, pseudo-distances or divergences) between the two laws, using the empirical distributions of the two samples. Denote H and G the laws of Y and X , respectively, and $H_{n_1}^Y$ and $G_{n_0}^X$ the empirical measures associated to the samples $Y_1, \dots, Y_{n_1}$ and $X_1, \dots, X_{n_0}$, respectively, namely

$$H_{n_1}^Y := \frac{1}{n_1} \sum_{i=1}^{n_1} \delta_{Y_i} \quad \text{and} \quad G_{n_0}^X := \frac{1}{n_0} \sum_{i=1}^{n_0} \delta_{X_i},$$

in which δ_x is the Dirac measure at point x , for all x .

We consider divergences, noted $\phi(\cdot, \cdot)$, between probability measures. The choice of the class of ϕ -divergences is motivated by their invariance property w.r.t. change of variables, and by the fact that it covers some classical methods.

The direct “plug-in” estimates $\phi(H_{n_1}^Y, G_{n_0}^X)$ of ϕ -divergences $\phi(H, G)$ is not defined when H and G do not share the same discrete finite support; we will solve this problem using the dual-representation of ϕ -divergences.

4.3. Estimation through the dual representation of ϕ -divergences

In order to introduce our estimates, we consider the following notations

1. $k(\theta) : x \mapsto k(\theta, x) = \varphi'(m(\theta, x))$;
2. $l(\theta) : x \mapsto l(\theta, x) = \varphi'(m(\theta, x)) \cdot m(\theta, x) - \varphi(m(\theta, x))$.
3. We sometimes write Pf for $\int f dP$ for any measure P and any function f .

We consider also the following conditions

(C.0) $\int |\varphi'(m(\theta, x))| dH(x)$ is finite for all $\theta \in \Theta$;

(C.1) $\phi(H, G)$ is finite.

Under conditions (C.0) and (C.1), by application of the above Theorem, for the probability measures H and G , choosing the class of functions

$$\mathcal{F} = \{x \mapsto \varphi'(m(\theta, x)) / \theta \in \Theta\},$$

we obtain, for the divergences $\phi(H, G)$, the representations

$$\phi(H, G) = \sup_{\theta \in \Theta} \{Hk(\theta) - Gl(\theta)\}. \quad (16)$$

So, we propose to estimate the divergences $\phi(H, G)$ by

$$\hat{\phi}_{n_0, n_1}(H, G) := \sup_{\theta \in \Theta} \{H_{n_1}^Y k(\theta) - G_{n_0}^X l(\theta)\}. \quad (17)$$

Further, the supremum in (16) is unique and is reached at $\theta = \theta_0$, that is

$$\theta_0 = \arg \sup_{\theta \in \Theta} \{Hk(\theta) - Gl(\theta)\}. \quad (18)$$

So, we propose the following estimates of θ_0

$$\hat{\theta}_{n_0, n_1} := \arg \sup_{\theta \in \Theta} \hat{\phi}_{n_0, n_1}(H, G) = \arg \sup_{\theta \in \Theta} \{H_{n_1}^Y k(\theta) - G_{n_0}^X l(\theta)\}. \quad (19)$$

We underline that the use of the Dual representation of ϕ -divergences provides estimates for ϕ -divergences and also estimates for the parameter θ_0 , while the plug-in direct approach obviously provides only estimates of ϕ -divergences between H and G .

Simulation results show that the choice of a good divergence depends on the ratio n_1/n_0 . Hence, we will introduce a class of ϕ -divergences depending on this ratio. In particular, in Section 4, we will consider the divergence associated to the nonnegative convex function

$$\varphi_\rho^*(x) := x \log x - \frac{1 + \rho x}{\rho} \log(1 + \rho x) + \frac{1 + \rho}{\rho} \log(1 + \rho) - \left(\frac{\rho - 1}{\rho} - \log(1 + \rho) \right) (x - 1),$$

which in the multiplicative-intercept risk model (15) yields to an estimate of the parameter θ_0 corresponding to the semiparametric maximum likelihood's.

Denote

$$n := n_0 + n_1 \quad \text{and} \quad \rho := \lim_{n \rightarrow \infty} \rho_n := \lim_{n \rightarrow \infty} \frac{n_1}{n_0},$$

which we suppose positive and finite. We consider any convex function depending on ρ , noted

$$\begin{aligned} \varphi_\rho &: [0, +\infty] \rightarrow [0, +\infty] \\ x &\mapsto \varphi_\rho(x) \end{aligned}$$

satisfying $\varphi_\rho(1) = 0$. Denote $\phi_\rho(H, G)$ the divergence between the probability laws H and G associated to the convex function φ_ρ . Consider also the following notation

1. $k_\rho(\theta) : x \mapsto k_\rho(\theta, x) = \varphi'_\rho(m(\theta, x));$
2. $l_\rho(\theta) : x \mapsto l_\rho(\theta, x) = \varphi'_\rho(m(\theta, x)) \cdot m(\theta, x) - \varphi_\rho(m(\theta, x)).$

We introduce estimates for the divergences $\phi_\rho(H, G)$ and for the parameter θ_0 using the ratio $\rho_n := \frac{n_1}{n_0}$ as follows

$$\hat{\phi}_{\rho_n}(H, G) := \sup_{\theta \in \Theta} \{H_{n_1}^Y k_{\rho_n}(\theta) - G_{n_0}^X l_{\rho_n}(\theta)\}, \quad (20)$$

and

$$\hat{\theta}_{\rho_n} := \arg \sup_{\theta \in \Theta} \hat{\phi}_{\rho_n}(H, G) = \arg \sup_{\theta \in \Theta} \{H_{n_1}^Y k_{\rho_n}(\theta) - G_{n_0}^X l_{\rho_n}(\theta)\}, \quad (21)$$

which we call “two-sample semiparametric dual ϕ -divergences estimates” (SD ϕ DE's).

The talk will discuss examples and present a reasonable choice for the divergence function in the case of small samples.

References

- Basu A. and Lindsay B. (1994) Minimum disparity estimation for continuous models: Efficiency, distributions and robustness, *Ann. Inst. Stat. Math.*, 46, 683–705.
- Beran R. (1977) Minimum hellinger distance estimates for parametric models, *Annals of Statistics*, 5, 445–463.
- Berlinet A. (1999) How to get central limit theorems for global errors of estimates, *Appl. Math.*, 44, 2, 81–96.
- Berlinet A. Vajda I. and Van der Meulen E. (1998) About the asymptotic accuracy of barron density estimates, *IEEE Trans. Inf. Theory*, 44,3, 999–1009.
- Broniatowski M. (2002) Estimation through kullback-leibler divergence, *to appear in Mathematical Methods of Statistics*, 12, 3.
- Cressie N. and Read T. (1984) Multinomial goodness of fit tests, *J. Roy. Statist. Soc. Ser. B*, 46,3, 440–464.
- Csiszár I. (1967) On topology properties of f -divergences, *Studia Sci. Math. Hungar.*, 2, 329–339.
- Dembo A. and Zeitouni O. (1998) *Large Deviation Techniques and Applications*, John and Bartlett Publishers.
- Fokianos K. (2002) Box-cox transformation for semiparametric comparison of two samples, *to appear in, Shores Conference Proceedings*.
- Györfi L., Liese F., Vajda I. and Van der Meulen E. (1998) Distribution estimates consistent in χ^2 -divergences, *Statistics*, 32,11, 31–57.
- Györfi L. and Vajda I. (2002) Asymptotic distributions for goodness of fit statistics in a sequence of multinomial models, *Statist. and Prob. Lett.*, 56,1, 57–67.
- Hosmer Jr. D.W. and Lemeshow S. (2000) *Applied logistic regression*, Wiley Series in Probability and Statistics: Texts and References Section, John Wiley & Sons Inc., New York.
- Kay R. and Little S. (1987) Transformations of the explanatory variables in the logistic regression model for binary data, *Biometrika*, 74, 3, 495–501.
- Landaburu E. and Pardo L. (2000) Goodness of fit tests with weights in the classes based on (h, ϕ) divergences, *Kybernetika*, 5, 589–602.
- Lehmann E.L. (1986) *Testing statistical hypotheses*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons Inc., New York, second edition.
- Liese F. and Vajda I. (1987) *Convex Statistical Distances*, BSB Teubner, Leipzig.
- Lindsay B.G. (1994) Efficiency versus robustness: the case for minimum Hellinger distance and related methods, *Ann. Statist.*, 22, 2, 1081–1114.
- Morales D., Pardo L. and Vajda I. (1995) Asymptotic divergence of estimates of discrete distributions, *Journal of Statistical Planning and Inference*, 38, 347–369.
- Qin J. (1998) Inferences for case-control and semiparametric two-sample density ratio models, *Biometrika*, 85, 3, 619–630.