Entropy **2014**, *16*, 2686-2698; doi:10.3390/e16052686

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Model Selection Criteria Using Divergences

Aida Toma [1,2]

[1] Department of Applied Mathematics, Bucharest Academy of Economic Studies, Piaţa Romană 6, Bucharest, 010374, Romania; E-Mail: aida_toma@yahoo.com; Tel.: +407-26093105

[2] "Gh. Mihoc-C. Iacob" Institute of Mathematical Statistics and Applied Mathematics, Romanian Academy, Calea 13 Septembrie 13, Bucharest, 050711, Romania

Received: 1 April 2014; in revised form: 12 May 2014 / Accepted: 13 May 2014 / Published: 14 May 2014

Abstract: In this note we introduce some divergence-based model selection criteria. These criteria are defined by estimators of the expected overall discrepancy between the true unknown model and the candidate model, using dual representations of divergences and associated minimum divergence estimators. It is shown that the proposed criteria are asymptotically unbiased. The influence functions of these criteria are also derived and some comments on robustness are provided.

Keywords: divergence measure; duality; model selection

1. Introduction

The minimum divergence approach is a useful technique in statistical inference. In recent years, the literature dedicated to the divergence-based statistical methods has grown substantially and the monographs of Pardo [1] and Basu *et al.* [2] are important references that present developments and applications in this field of research. Minimum divergence estimators and related methods have received considerable attention in statistical inference because of their ability to reconcile efficiency and robustness. Among others, Beran [3], Tamura and Boos [4], Simpson [5,6] and Toma [7] proposed families of parametric estimators minimizing the Hellinger distance between a nonparametric estimator of the observations density and the model. They showed that those estimators are both asymptotically efficient and robust. Generalizing earlier work based on the Hellinger distance, Lindsay [8] and Basu and Lindsay [9] have investigated minimum divergence estimators, for both discrete and continuous models. Some families of estimators based on approximate divergence criteria have also been considered; see Basu *et al.* [10]. Broniatowski and Keziou [11] have introduced a minimum divergence estimation method based on a dual representation of the divergence between probability measures. Their estimators, called minimum dual divergence estimators, are defined in a unified way for both continuous and discrete models. They do not require any prior smoothing and include the classical maximum likelihood estimators as a benchmark. Robustness properties of these estimators have been studied in [12,13].

In this paper we apply estimators of divergences in dual form and corresponding minimum dual divergence estimators, as presented by Broniatowski and Keziou [11], in the context of model selection.

Model selection is a method for selecting the best model among candidate models. A model selection criterion can be considered as an approximately unbiased estimator of the expected overall discrepancy, a nonnegative quantity that measures the distance between the true unknown model and a fitted approximating model. If the value of the criterion is small, then the approximated candidate model can be chosen.

Many model selection criteria have been proposed so far. Classical model selection criteria using least square error and log-likelihood include the C_p-criterion, cross-validation (CV), the Akaike information criterion (AIC) based on the well-known Kullback–Leibler divergence, Bayesian information criterion (BIC), a general class of criteria that also estimates the Kullback–Leibler divergence (GIC). These criteria have been proposed by Mallows [14], Stone [15], Akaike [16], Schwarz [17] and Konishi and Kitagawa [18], respectively. Robust versions of classical model selection criteria, which are not strongly affected by outliers, have been firstly proposed by Ronchetti [19], Ronchetti and Staudte [20]. Other references on this topic can be found in Maronna *et al.* [21]. Among the recent proposals for model selection we recall the criteria presented by Karagrigoriou *et al.* [22], the divergence information criteria (DIC) introduced by Mattheou *et al.* [23]. The DIC criteria use the density power divergences introduced by Basu *et al.* [10].

In the present paper, we apply the same methodology used for AIC, and also for DIC, to a general class of divergences including the Cressie–Read divergences [24] in order to obtain model selection criteria. These criteria also use dual forms of the divergences and minimum dual divergence estimators. We show that the criteria are asymptotically unbiased and compute the corresponding influence functions.

The paper is organized as follows. In Section 2 we recall the duality formula for divergences, as well as the definitions of associated dual divergence estimators and minimum dual divergence estimators, together with their asymptotic properties, all these being necessary in the next section where we define new criteria for model selection. In Section 3, we apply the same methodology used for AIC to the divergences in dual form in order to develop criteria for model selection. We define criteria based on estimators of the expected overall discrepancy and prove their asymptotic unbiasedness. The influence functions of the proposed criteria are also derived. In Section 4 we present some conclusions.

2. Minimum Dual Divergence Estimators

2.1. Examples of Divergences

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 $\varphi(0) = \lim_{x \downarrow 0} \varphi(x)$. Let $(\mathcal{X}, \mathcal{B})$ be a measurable space and P be a probability measure (p.m.) defined on $(\mathcal{X}, \mathcal{B})$. Following Rüschendorf [25], for any p.m. Q absolutely continuous (a.c.) w.r.t. P, the divergence between Q and P is defined by

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

When Q is not a.c. w.r.t. P, we set $D(Q, P) = \infty$. We refer to Liese and Vajda [26] for an overview on the origin of the concept of divergence in statistics.

A commonly used family of divergences is the so-called "power divergences" or Cressie–Read divergences. This family is defined by the class of functions

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

for $\gamma \in \mathbb{R} \setminus \{0, 1\}$ and $\varphi_0(x) := -\log x + x - 1$, $\varphi_1(x) := x \log x - x + 1$ with $\varphi_\gamma(0) = \lim_{x \downarrow 0} \varphi_\gamma(x)$, $\varphi_\gamma(\infty) = \lim_{x \to \infty} \varphi_\gamma(x)$, for any $\gamma \in \mathbb{R}$. The Kullback–Leibler divergence (KL) is associated with φ_1, the modified Kullback–Leibler (KL$_m$) to φ_0, the χ^2 divergence to φ_2, the modified χ^2 divergence (χ^2_m) to φ_{-1} and the Hellinger distance to $\varphi_{1/2}$. We refer to [11] for the modified versions of χ^2 and KL divergences.

Some applied models using divergence and entropy measures can be found in Toma and Leoni-Aubin [27], Kallberg *et al.* [28], Preda *et al.* [29] and Basu *et al.* [2], among others.

2.2. Dual Form of a Divergence and Minimum Divergence Estimators

Let $\{F_\theta, \theta \in \Theta\}$ be an identifiable parametric model, where Θ is a subset of $\mathbb{R}^p$. We assume that for any $\theta \in \Theta$, F_θ has density f_θ with respect to some dominating σ-finite measure λ. Consider the problem of estimating the unknown true value of the parameter θ_0 on the basis of an i.i.d. sample $X_1, \ldots, X_n$ with the law F_{θ_0}.

In the following, $D(f_\theta, f_{\theta_0})$ denotes the divergence between f_θ and f_{θ_0}, namely

$$D(f_\theta, f_{\theta_0}) := \int \varphi \left(\frac{f_\theta}{f_{\theta_0}} \right) f_{\theta_0} \mathrm{d}\lambda. \tag{3}$$

Using a Fenchel duality technique, Broniatowski and Keziou [11] have proved a dual representation of divergences. The main interest on this duality formula is that it leads to a wide variety of estimators, by a plug-in method of the empirical measure evaluated to the data set, without making use of any grouping, nor smoothing.

We consider divergences, defined through differentiable functions φ, that we assume to satisfy

(C.0) There exists $0 < \delta < 1$ such that for all $c \in [1 - \delta, 1 + \delta]$, there exist numbers c_1, c_2, c_3 such that

$$\varphi(cx) \leq c_1 \varphi(x) + c_2 |x| + c_3, \; \forall \, x \in \mathbb{R}. \tag{4}$$

Condition (C.0) holds for all power divergences, including KL and KL$_m$ divergences.

Assuming that $D(f_\theta, f_{\theta_0})$ is finite and that the function φ satisfies the condition (C.0), the dual representation holds

$$D(f_\theta, f_{\theta_0}) = \sup_{\alpha \in \Theta} \int m(\alpha, \theta, x) f_{\theta_0}(x) \mathrm{d}x, \tag{5}$$

with

$$m(\alpha, \theta, x) := \int \dot{\varphi}\left(\frac{f_\theta(z)}{f_\alpha(z)}\right) f_\theta(z)\mathrm{d}z - \left\{\dot{\varphi}\left(\frac{f_\theta(x)}{f_\alpha(x)}\right)\frac{f_\theta(x)}{f_\alpha(x)} - \varphi\left(\frac{f_\theta(x)}{f_\alpha(x)}\right)\right\}, \tag{6}$$

where $\dot{\varphi}$ is the notation for the derivative of φ, the supremum in Equation (5) being uniquely attained in $\alpha = \theta_0$, independently on θ.

We mention that the dual representation Equation (5) of divergences has been obtained independently by Liese and Vajda [30].

Naturally, for fixed θ, an estimator of the divergence $D(f_\theta, f_{\theta_0})$ is obtained by replacing Equation (5) by its sample analogue. This estimator is exactly

$$\widehat{D}(f_\theta, f_{\theta_0}) := \sup_{\alpha\in\Theta} \frac{1}{n}\sum_{i=1}^{n} m(\alpha, \theta, X_i), \tag{7}$$

the supremum being attained for

$$\widehat{\alpha}(\theta) := \arg\sup_{\alpha\in\Theta} \frac{1}{n}\sum_{i=1}^{n} m(\alpha, \theta, X_i). \tag{8}$$

Formula (8) defines a class of estimators of the parameter θ_0 called dual divergence estimators.

Further, since

$$\inf_{\theta\in\Theta} D(f_\theta, f_{\theta_0}) = D(f_{\theta_0}, f_{\theta_0}) = 0 \tag{9}$$

and since the infimum in the above display is unique, a natural definition of estimators of the parameter θ_0, called minimum dual divergence estimators, is provided by

$$\widehat{\theta} := \arg\inf_{\theta\in\Theta} \widehat{D}(f_\theta, f_{\theta_0}) = \arg\inf_{\theta\in\Theta}\sup_{\alpha\in\Theta} \frac{1}{n}\sum_{i=1}^{n} m(\alpha, \theta, X_i). \tag{10}$$

For more details on the dual representation of divergences and associated minimum dual divergence estimators, we refer to Broniatowski and Keziou [11].

2.3. Asymptotic Properties

Broniatowski and Keziou [11] have proved both the weak and the strong consistency, as well as the asymptotic normality for the classes of estimators $\widehat{\alpha}(\theta), \widehat{\alpha}(\widehat{\theta})$ and $\widehat{\theta}$. Here, we shortly recall those asymptotic results that will be used in the next sections. The following conditions are considered.

(C.1) The estimates $\widehat{\theta}$ and $\widehat{\alpha}(\widehat{\theta})$ exist.

(C.2) $\sup_{\alpha,\theta\in\Theta} |\frac{1}{n}\sum_{i=1}^{n} m(\alpha, \theta, X_i) - \int m(\alpha, \theta, x) f_{\theta_0}(x)\mathrm{d}x|$ tends to 0 in probability.

(a) for any positive ε, there exists some positive η such that for any $\alpha \in \Theta$ with $\|\alpha - \theta_0\| > \varepsilon$ and for all $\theta \in \Theta$ it holds that $\int m(\alpha, \theta, x) f_{\theta_0}(x)\mathrm{d}x < \int m(\theta_0, \theta, x) f_{\theta_0}(x)\mathrm{d}x - \eta$.

(b) there exists some neighborhood N_{θ_0} of θ_0 such that for any positive ε, there exists some positive η such that for all $\alpha \in N_{\theta_0}$ and all $\theta \in \Theta$ satisfying $\|\theta - \theta_0\| > \varepsilon$, it holds that $\int m(\alpha, \theta_0, x) f_{\theta_0}(x)\mathrm{d}x < \int m(\alpha, \theta, x) f_{\theta_0}(x)\mathrm{d}x - \eta$.

(C.3) There exists some neighborhood N_{θ_0} of θ_0 and a positive function H with $\int H(x) f_{\theta_0}(x)\mathrm{d}x$ finite, such that for all $\alpha \in N_{\theta_0}$, $\|m(\alpha, \theta_0, X)\| \leq H(X)$ in probability.

(C.4) There exists a neighborhood N_{θ_0} of θ_0 such that the first and the second order partial derivatives with respect to α and θ of $\dot{\varphi}\left(\frac{f_\theta(x)}{f_\alpha(x)}\right) f_\theta(x)$ are dominated on $N_{\theta_0} \times N_{\theta_0}$ by some λ-integrable functions. The third order partial derivatives with respect to α and θ of $m(\alpha, \theta, x)$ are dominated on $N_{\theta_0} \times N_{\theta_0}$ by some P_{θ_0}-integrable functions (where P_{θ_0} is the probability measure corresponding to the law F_{θ_0}).

(C.5) The integrals $\int \|\frac{\partial}{\partial\alpha} m(\theta_0, \theta_0, x)\|^2 f_{\theta_0}(x)\mathrm{d}x$, $\int \|\frac{\partial}{\partial\theta} m(\theta_0, \theta_0, x)\|^2 f_{\theta_0}(x)\mathrm{d}x$, $\int \|\frac{\partial^2}{\partial^2\alpha} m(\theta_0, \theta_0, x)\| f_{\theta_0}(x)\mathrm{d}x$, $\int \|\frac{\partial^2}{\partial^2\theta} m(\theta_0, \theta_0, x)\| f_{\theta_0}(x)\mathrm{d}x$, $\int \|\frac{\partial^2}{\partial\theta\partial\alpha} m(\theta_0, \theta_0, x)\| f_{\theta_0}(x)\mathrm{d}x$ are finite and the Fisher information matrix $I(\theta_0) := \int \frac{\dot{f}_{\theta_0}(z)\dot{f}^t_{\theta_0}(z)}{f_{\theta_0}(z)}\mathrm{d}z$ is nonsingular, t denoting the transpose.

Proposition 1. *Assume that conditions (C.1)–(C.3) hold. Then*

(a) $\sup_{\theta\in\Theta} \|\widehat{\alpha}(\theta) - \theta_0\|$ *tends to 0 in probability.*

(b) $\widehat{\theta}$ *converges to* θ_0 *in probability.*

If (C.1)–(C.5) are fulfilled, then

(c) $\sqrt{n}(\widehat{\theta} - \theta_0)$ *and* $\sqrt{n}(\widehat{\alpha}(\widehat{\theta}) - \theta_0)$ *converge in distribution to a centered* p*-variate normal random variable with covariance matrix* $I(\theta_0)^{-1}$.

For discussions and examples about the fulfillment of conditions (C.1)–(C.5), we refer to Broniatowski and Keziou [11].

3. Model Selection Criteria

In this section, we apply the same methodology used for AIC to the divergences in dual form in order to develop model selection criteria. Consider a random sample $X_1, \ldots, X_n$ from the distribution with density g (the true model) and a candidate model f_θ from a parametric family of models (f_θ) indexed by an unknown parameter $\theta \in \Theta$, where Θ is a subset of $\mathbb{R}^p$. We use divergences satisfying (C.0) and denote for simplicity the divergence $D(f_\theta, g)$ between f_θ and the true density g by W_θ.

3.1. The Expected Overall Discrepancy

The target theoretical quantity that will be approximated by an asymptotically unbiased estimator is given by

$$E[W_{\widehat{\theta}}] = E[W_\theta | \theta = \widehat{\theta}] \tag{11}$$

where $\widehat{\theta}$ is a minimum dual divergence estimator defined by Equation (10). The same divergence is used for both W_θ and $\widehat{\theta}$. The quantity $E[W_{\widehat{\theta}}]$ can be viewed as the average distance between g and (f_θ) and it is called the expected overall discrepancy between g and (f_θ).

The next Lemma gives the gradient vector and the Hessian matrix of W_θ and is useful for evaluating the expected overall discrepancy $E[W_{\widehat{\theta}}]$ through Taylor expansion. We denote by $\dot{f}_\theta$ and $\ddot{f}_\theta$ the first and the second order derivative of f_θ with respect to θ, respectively. We assume the following conditions allowing derivation under the integral sign.

(C.6) There exists a neighborhood N_θ of θ such that

$$\int \sup_{u\in N_\theta} \left\| \frac{\partial}{\partial u}\left[\varphi\left(\frac{f_u}{g}\right)\right]\right\| g\mathrm{d}\lambda < \infty. \tag{12}$$

(C.7) There exists a neighborhood N_θ of θ such that

$$\int \sup_{u\in N_\theta} \left\| \frac{\partial}{\partial u} \left[\dot{\varphi}\left(\frac{f_u}{g}\right) \dot{f}_u \right] \right\| \mathrm{d}\lambda < \infty. \tag{13}$$

Lemma 1. *Assume that conditions (C.6) and (C.7) hold. Then, the gradient vector* $\frac{\partial}{\partial\theta}W_\theta$ *of* W_θ *is given by*

$$\int \dot{\varphi}\left(\frac{f_\theta}{g}\right) \dot{f}_\theta \mathrm{d}\lambda \tag{14}$$

and the Hessian matrix $\frac{\partial^2}{\partial^2\theta}W_\theta$ *is given by*

$$\int \left[\ddot{\varphi}\left(\frac{f_\theta}{g}\right) \frac{\dot{f}_\theta \dot{f}_\theta^t}{g} + \dot{\varphi}\left(\frac{f_\theta}{g}\right) \ddot{f}_\theta \right] \mathrm{d}\lambda. \tag{15}$$

The proof of this Lemma is straightforward, therefore it is omitted.

Particularly, when using Cressie–Read divergences, the gradient vector $\frac{\partial}{\partial\theta}W_\theta$ of W_θ is given by

$$\begin{aligned}
&\frac{1}{\gamma-1}\int \left(\frac{f_\theta(z)}{g(z)}\right)^{\gamma-1} \dot{f}_\theta(z)\mathrm{d}z, && \text{if } \gamma\in\mathbb{R}\backslash\{0,1\} && (16)\\
&-\int \frac{g(z)}{f_\theta(z)}\dot{f}_\theta(z)\mathrm{d}z, && \text{if } \gamma = 0 && (17)\\
&\int \log\left(\frac{f_\theta(z)}{g(z)}\right) \dot{f}_\theta(z)\mathrm{d}z, && \text{if } \gamma = 1 && (18)
\end{aligned}$$

and the Hessian matrix $\frac{\partial^2}{\partial^2\theta}W_\theta$ is given by

$$\begin{aligned}
&\int \left(\frac{f_\theta(z)}{g(z)}\right)^{\gamma-1} \frac{\dot{f}_\theta(z)\dot{f}_\theta^t(z)}{f_\theta(z)}\mathrm{d}z + \frac{1}{\gamma-1}\int \left(\frac{f_\theta(z)}{g(z)}\right)^{\gamma-1} \ddot{f}_\theta(z)\mathrm{d}z, && \text{if } \gamma\in\mathbb{R}\backslash\{0,1\} && (19)\\
&\int \frac{g(z)}{f_\theta^2(z)}\dot{f}_\theta(z)\dot{f}_\theta^t(z)\mathrm{d}z - \int \frac{g(z)}{f_\theta(z)}\ddot{f}_\theta(z)\mathrm{d}z, && \text{if } \gamma = 0 && (20)\\
&\int \log\left(\frac{f_\theta(z)}{g(z)}\right) \ddot{f}_\theta(z)\mathrm{d}z + \int \frac{\dot{f}_\theta(z)\dot{f}_\theta^t(z)}{f_\theta(z)}\mathrm{d}z, && \text{if } \gamma = 1. && (21)
\end{aligned}$$

When the true model g belongs to the parametric model (f_θ), hence $g = f_{\theta_0}$, the gradient vector and the Hessian matrix of W_θ evaluated in $\theta = \theta_0$ simplify to

$$\left[\frac{\partial}{\partial\theta}W_\theta\right]_{\theta=\theta_0} = 0 \tag{22}$$

$$\left[\frac{\partial^2}{\partial^2\theta}W_\theta\right]_{\theta=\theta_0} = \ddot{\varphi}(1)I(\theta_0). \tag{23}$$

The hypothesis that the true model g belongs to the parametric family (f_θ) is the assumption made by Akaike [16]. Although this assumption is questionable in practice, it is useful because it provides the basis for the evaluation of the expected overall discrepancy (see also [23]).

Proposition 2. *When the true model* g *belongs to the parametric model* (f_θ)*, assuming that conditions (C.6) and (C.7) are fulfilled for* $g = f_{\theta_0}$ *and* $\theta = \theta_0$*, the expected overall discrepancy is given by*

$$E[W_{\widehat{\theta}}] = W_{\theta_0} + \frac{\ddot{\varphi}(1)}{2}E[(\widehat{\theta}-\theta_0)^t I(\theta_0)(\widehat{\theta}-\theta_0)] + E[R_n], \tag{24}$$

where $R_n = o(\|\widehat{\theta} - \theta_0\|^2)$ and θ_0 is the true value of the parameter.

Proof. By applying a Taylor expansion to W_θ around the true parameter θ_0 and taking $\theta = \widehat{\theta}$, on the basis of Equations (22) and (23), we obtain

$$W_{\widehat{\theta}} = W_{\theta_0} + \frac{\ddot{\varphi}(1)}{2}(\widehat{\theta} - \theta_0)^t I(\theta_0)(\widehat{\theta} - \theta_0) + o(\|\widehat{\theta} - \theta_0\|^2). \tag{25}$$

Then Equation (24) is proved. □

3.2. Estimation of the Expected Overall Discrepancy

In this section we construct an asymptotically unbiased estimator of the expected overall discrepancy, under the hypothesis that the true model g belongs to the parametric family (f_θ).

For a given $\theta \in \Theta$, a natural estimator of W_θ is

$$Q_\theta := \sup_{\alpha \in \Theta} \frac{1}{n} \sum_{i=1}^{n} m(\alpha, \theta, X_i) = \frac{1}{n} \sum_{i=1}^{n} m(\widehat{\alpha}(\theta), \theta, X_i), \tag{26}$$

where $m(\alpha, \theta, x)$ is given by formula (6), which can also be expressed as

$$Q_\theta = \int \dot{\varphi}\left(\frac{f_\theta(z)}{f_{\widehat{\alpha}(\theta)}(z)}\right) f_\theta(z) \mathrm{d}z - \frac{1}{n} \sum_{i=1}^{n} \left\{ \dot{\varphi}\left(\frac{f_\theta(X_i)}{f_{\widehat{\alpha}(\theta)}(X_i)}\right) \frac{f_\theta(X_i)}{f_{\widehat{\alpha}(\theta)}(X_i)} - \varphi\left(\frac{f_\theta(X_i)}{f_{\widehat{\alpha}(\theta)}(X_i)}\right) \right\} \tag{27}$$

using the sample analogue of the dual representation of the divergence.

The following conditions allow derivation under the integral sign for the integral term of Q_θ.

(C.8) There exists a neighborhood N_θ of θ such that

$$\int \sup_{u \in N_\theta} \left\| \frac{\partial}{\partial u} \left[\dot{\varphi}\left(\frac{f_u}{f_{\widehat{\alpha}(u)}}\right) f_u \right] \right\| \mathrm{d}\lambda < \infty. \tag{28}$$

(C.9) There exists a neighborhood N_θ of θ such that

$$\int \sup_{u \in N_\theta} \left\| \frac{\partial}{\partial u} \left[\ddot{\varphi}\left(\frac{f_u}{f_{\widehat{\alpha}(u)}}\right) \left\{ \frac{f_u}{f_{\widehat{\alpha}(u)}} \dot{f}_u - \left(\frac{f_u}{f_{\widehat{\alpha}(u)}}\right)^2 \cdot \frac{\partial}{\partial u}\widehat{\alpha}(u) \cdot \dot{f}_{\widehat{\alpha}(u)} \right\} + \dot{\varphi}\left(\frac{f_u}{f_{\widehat{\alpha}(u)}}\right) \dot{f}_u \right] \right\| \mathrm{d}\lambda < \infty. \tag{29}$$

Lemma 2. *Under (C.8) and (C.9), the gradient vector and the Hessian matrix of Q_θ are*

$$\frac{\partial}{\partial \theta} Q_\theta = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} m(\widehat{\alpha}(\theta), \theta, X_i) \tag{30}$$

$$\frac{\partial^2}{\partial^2 \theta} Q_\theta = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial^2 \theta} m(\widehat{\alpha}(\theta), \theta, X_i). \tag{31}$$

Proof. Since

$$Q_\theta = \frac{1}{n} \sum_{i=1}^{n} m(\widehat{\alpha}(\theta), \theta, X_i) \tag{32}$$

derivation yields

$$\frac{\partial}{\partial\theta}Q_\theta \;=\; \frac{\partial}{\partial\theta}\widehat{\alpha}(\theta)\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\alpha}m(\widehat{\alpha}(\theta),\theta,X_i)\right]+\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}m(\widehat{\alpha}(\theta),\theta,X_i). \tag{33}$$

Note that, by its very definition, $\widehat{\alpha}(\theta)$ is a solution of the equation

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\alpha}m(\alpha,\theta,X_i)=0 \tag{34}$$

taken with respect to α, therefore

$$\frac{\partial}{\partial\theta}Q_\theta=\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}m(\widehat{\alpha}(\theta),\theta,X_i). \tag{35}$$

On the other hand,

$$\frac{\partial^2}{\partial^2\theta}Q_\theta \;=\; \frac{\partial}{\partial\theta}\widehat{\alpha}(\theta)\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta\partial\alpha}m(\widehat{\alpha}(\theta),\theta,X_i)\right]+\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial^2\theta}m(\widehat{\alpha}(\theta),\theta,X_i) \tag{36}$$

$$\;=\; \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial^2\theta}m(\widehat{\alpha}(\theta),\theta,X_i). \quad \square \tag{37}$$

Proposition 3. *Under conditions (C.1)–(C.3) and (C.8)–(C.9) and assuming that the integrals* $\int\|\frac{\partial^2}{\partial^2\theta}m(\theta_0,\theta_0,x)\|f_{\theta_0}(x)\mathrm{d}x$, $\int\|\frac{\partial^3}{\partial^2\theta\partial\alpha}m(\theta_0,\theta_0,x)\|f_{\theta_0}(x)\mathrm{d}x$ *and* $\int\|\frac{\partial^3}{\partial^3\theta}m(\theta_0,\theta_0,x)\|f_{\theta_0}(x)\mathrm{d}x$ *are finite, the gradient vector and the Hessian matrix of* Q_θ *evaluated in* $\theta=\widehat{\theta}$ *satisfy*

$$\left[\frac{\partial}{\partial\theta}Q_\theta\right]_{\widehat{\theta}} \;=\; 0 \tag{38}$$

$$\left[\frac{\partial^2}{\partial^2\theta}Q_\theta\right]_{\widehat{\theta}} \;=\; \ddot{\varphi}(1)I(\theta_0)+o_P(1). \tag{39}$$

Proof. By the very definition of $\widehat{\theta}$, the equality (38) is verified. For the second relation, we take $\theta=\widehat{\theta}$ in Equation (31) and obtain

$$\left[\frac{\partial^2}{\partial^2\theta}Q_\theta\right]_{\widehat{\theta}}=\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial^2\theta}m(\widehat{\alpha}(\widehat{\theta}),\widehat{\theta},X_i). \tag{40}$$

A Taylor expansion of $\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial^2\theta}m(\alpha,\theta,X_i)$ as function of (α,θ) around to (θ_0,θ_0) yields

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial^2\theta}m(\widehat{\alpha}(\widehat{\theta}),\widehat{\theta},X_i)=\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial^2\theta}m(\theta_0,\theta_0,X_i)+\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial^2\theta\partial\alpha}m(\theta_0,\theta_0,X_i)\right]\cdot$$

$$\cdot(\widehat{\alpha}(\widehat{\theta})-\theta_0)+\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^3}{\partial^3\theta}m(\theta_0,\theta_0,X_i)\right](\widehat{\theta}-\theta_0)+o(\sqrt{\|\widehat{\alpha}(\widehat{\theta})-\theta_0\|^2+\|\widehat{\theta}-\theta_0\|^2}).$$

Using the fact that $\int\|\frac{\partial^2}{\partial^2\theta}m(\theta_0,\theta_0,x)\|f_{\theta_0}(x)\mathrm{d}x$ is finite, the weak law of large numbers leads to

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial^2\theta}m(\theta_0,\theta_0,X_i)\xrightarrow{P}\ddot{\varphi}(1)I(\theta_0). \tag{41}$$

Then, since $(\widehat{\alpha}(\widehat{\theta}) - \theta_0) = o_P(1)$ and $(\widehat{\theta} - \theta_0) = o_P(1)$, and taking into account that $\int \|\frac{\partial^3}{\partial^2\theta\partial\alpha} m(\theta_0, \theta_0, x)\| f_{\theta_0}(x)\mathrm{d}x$ and $\int \|\frac{\partial^3}{\partial^3\theta} m(\theta_0, \theta_0, x)\| f_{\theta_0}(x)\mathrm{d}x$ are finite, we deduce that

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial^2\theta} m(\widehat{\alpha}(\widehat{\theta}), \widehat{\theta}, X_i) = \ddot{\varphi}(1)I(\theta_0) + o_P(1). \tag{42}$$

Thus we obtain Equation (39). □

In the following, we suppose that conditions of Proposition 1, Proposition 2 and Proposition 3 are all satisfied. These conditions allow obtaining an asymptotically unbiased estimator of the expected overall discrepancy.

Proposition 4. *When the true model g belongs to the parametric model (f_θ), the expected overall discrepancy evaluated at $\widehat{\theta}$ is given by*

$$E[W_{\widehat{\theta}}] = E[Q_{\widehat{\theta}} + \ddot{\varphi}(1)(\widehat{\theta} - \theta_0)^t I(\theta_0)(\widehat{\theta} - \theta_0) + R_n], \tag{43}$$

where $R_n = o(\|\theta_0 - \widehat{\theta}\|^2)$.

Proof. A Taylor expansion of Q_θ around to $\widehat{\theta}$ yields

$$Q_\theta = Q_{\widehat{\theta}} + (\theta - \widehat{\theta})^t \left[\frac{\partial}{\partial\theta} Q_\theta\right]_{\widehat{\theta}} + \frac{1}{2}(\theta - \widehat{\theta})^t \left[\frac{\partial^2}{\partial^2\theta} Q_\theta\right]_{\widehat{\theta}} (\theta - \widehat{\theta}) + o(\|\theta - \widehat{\theta}\|^2) \tag{44}$$

and using Proposition 3, we have

$$Q_\theta = Q_{\widehat{\theta}} + \frac{1}{2}(\theta - \widehat{\theta})^t[\ddot{\varphi}(1)I(\theta_0) + o_P(1)](\theta - \widehat{\theta}) + o(\|\theta - \widehat{\theta}\|^2). \tag{45}$$

Taking $\theta = \theta_0$, for large n, it holds

$$Q_{\theta_0} = Q_{\widehat{\theta}} + \frac{\ddot{\varphi}(1)}{2}(\theta_0 - \widehat{\theta})^t I(\theta_0)(\theta_0 - \widehat{\theta}) + o(\|\theta_0 - \widehat{\theta}\|^2) \tag{46}$$

and consequently

$$E[Q_{\theta_0}] = E[Q_{\widehat{\theta}}] + \frac{\ddot{\varphi}(1)}{2}E[(\theta_0 - \widehat{\theta})^t I(\theta_0)(\theta_0 - \widehat{\theta})] + E[R_n], \tag{47}$$

where $R_n = o(\|\theta_0 - \widehat{\theta}\|^2)$.

According to Proposition 2 it holds

$$E[W_{\widehat{\theta}}] = W_{\theta_0} + \frac{\ddot{\varphi}(1)}{2}E[(\widehat{\theta} - \theta_0)^t I(\theta_0)(\widehat{\theta} - \theta_0)] + E[R_n]. \tag{48}$$

Note that

$$\begin{aligned} E[Q_\theta] &= E\left[\sup_{\alpha\in\Theta} \frac{1}{n}\sum_{i=1}^{n} m(\alpha, \theta, X_i)\right] = \sup_{\alpha\in\Theta} E\left[\frac{1}{n}\sum_{i=1}^{n} m(\alpha, \theta, X_i)\right] \\ &= \sup_{\alpha\in\Theta} E\left[m(\alpha, \theta, X_i)\right] = \sup_{\alpha\in\Theta} \int m(\alpha, \theta, x) f_{\theta_0}(x)\mathrm{d}x = W_\theta. \end{aligned} \tag{49}$$

Then, combining Equation (48) with Equations (49) and (47), we get

$$E[W_{\widehat{\theta}}] = E[Q_{\widehat{\theta}} + \ddot{\varphi}(1)(\widehat{\theta} - \theta_0)^t I(\theta_0)(\widehat{\theta} - \theta_0) + R_n]. \quad \square \tag{50}$$

Proposition 4 shows that an asymptotically unbiased estimator of the expected overall discrepancy is given by

$$Q_{\widehat{\theta}} + \ddot{\varphi}(1)(\widehat{\theta} - \theta_0)^t I(\theta_0)(\widehat{\theta} - \theta_0). \tag{51}$$

According to Proposition 1, $\sqrt{n}(\widehat{\theta} - \theta_0)$ is asymptotically distributed as $\mathcal{N}_p(0, I(\theta_0)^{-1})$. Consequently, $n(\widehat{\theta}-\theta_0)^t I(\theta_0)(\widehat{\theta}-\theta_0)$ has approximately a χ^2_p distribution. Then, taking into account that $no(\|\widehat{\theta}-\theta_0\|^2) = o_P(1)$, an asymptotically unbiased estimator of n-times the expected overall discrepancy evaluated at $\widehat{\theta}$ is provided by

$$nQ_{\widehat{\theta}} + \ddot{\varphi}(1)p. \tag{52}$$

3.3. Influence Functions

In the following, we compute the influence function of the statistics $Q_{\widehat{\theta}}$. As it is known, the influence function is a useful tool for describing the robustness of an estimator. Recall that a map T defined on a set of distribution functions and parameter space valued is a statistical functional corresponding to an estimator $\widehat{\theta}$ of the parameter θ, if $\widehat{\theta} = T(F_n)$, where F_n is the empirical distribution function associated to the sample. The influence function of T at F_θ is defined by

$$\mathrm{IF}(x; T, F_\theta) := \left.\frac{\partial T(\widetilde{F}_{\varepsilon x})}{\partial \varepsilon}\right|_{\varepsilon=0} \tag{53}$$

where $\widetilde{F}_{\varepsilon x} := (1-\varepsilon)F_\theta + \varepsilon\delta_x$, $\varepsilon > 0$, δ_x being the Dirac measure putting all mass at x. Whenever the influence function is bounded with respect to x, the corresponding estimator is called robust (see [31]).

Since

$$Q_{\widehat{\theta}} = \frac{1}{n}\sum_{i=1}^{n} m(\widehat{\alpha}(\widehat{\theta}), \widehat{\theta}, X_i), \tag{54}$$

the statistical functional corresponding to $Q_{\widehat{\theta}}$, which we denote by $U(\cdot)$, is defined by

$$U(F) := \int m(T_{V(F)}(F), V(F), y)\mathrm{d}F(y) \tag{55}$$

where $T_\theta(F)$ is the statistical functional associated to the estimator $\widehat{\alpha}(\theta)$ and $V(F)$ is the statistical functional associated to the estimator $\widehat{\theta}$.

Proposition 5. *The influence function of* $Q_{\widehat{\theta}}$ *is*

$$\mathrm{IF}(x; U, F_{\theta_0}) = \ddot{\varphi}(1)\frac{\dot{f}_{\theta_0}(x)}{f_{\theta_0}(x)}. \tag{56}$$

Proof. For the contaminated model $\widetilde{F}_{\varepsilon x} := (1-\varepsilon)F_{\theta_0} + \varepsilon\delta_x$, it holds

$$U(\widetilde{F}_{\varepsilon x}) = (1-\varepsilon)\int m(T_{V(\widetilde{F}_{\varepsilon x})}(\widetilde{F}_{\varepsilon x}), V(\widetilde{F}_{\varepsilon x}), y)\mathrm{d}F_{\theta_0}(y) + \varepsilon m(T_{V(\widetilde{F}_{\varepsilon x})}(\widetilde{F}_{\varepsilon x}), V(\widetilde{F}_{\varepsilon x}), x). \tag{57}$$

Derivation with respect to ε yields

$$\begin{aligned}\frac{\partial}{\partial\varepsilon}[U(\widetilde{F}_{\varepsilon x})]_{\varepsilon=0} &= -\int m(\theta_0,\theta_0,y)\mathrm{d}F_{\theta_0}(y) + \left[\int \frac{\partial}{\partial\alpha}m(\theta_0,\theta_0,y)\mathrm{d}F_{\theta_0}(y)\right]\frac{\partial}{\partial\varepsilon}[T_{V(\widetilde{F}_{\varepsilon x})}(\widetilde{F}_{\varepsilon x})]_{\varepsilon=0} + \\ &\quad + \left[\int \frac{\partial}{\partial\theta}m(\theta_0,\theta_0,y)\mathrm{d}F_{\theta_0}(y)\right]\mathrm{IF}(x; V, F_{\theta_0}) + m(\theta_0,\theta_0,x).\end{aligned}$$

Note that $m(\theta_0, \theta_0, y) = 0$ for any y and $\int \frac{\partial}{\partial \alpha} m(\theta_0, \theta_0, y) \mathrm{d}F_{\theta_0}(y) = 0$. Also, some straightforward calculations give

$$\int \frac{\partial}{\partial \theta} m(\theta_0, \theta_0, y) \mathrm{d}F_{\theta_0}(y) = \ddot{\varphi}(1) I(\theta_0). \tag{58}$$

On the other hand, according to the results presented in [12], the influence function of the minimum dual divergence estimator is

$$\mathrm{IF}(x; V, F_{\theta_0}) = I(\theta_0)^{-1} \frac{\dot{f}_{\theta_0}(x)}{f_{\theta_0}(x)}. \tag{59}$$

Consequently, we obtain Equation (60). □

Note that, for Cressie–Read divergences, it holds

$$\mathrm{IF}(x; U, F_{\theta_0}) = \frac{\dot{f}_{\theta_0}(x)}{f_{\theta_0}(x)} \tag{60}$$

irrespective of the used divergence, since $\ddot{\varphi}_\gamma(1) = 1$, for any γ.

Generally, $\mathrm{IF}(x; U, F_{\theta_0})$ is not bounded, therefore the robustness of the statistics $Q_{\widehat{\theta}}$, as measured by the influence function, does not hold.

4. Conclusions

The dual representation of divergences and corresponding minimum dual divergence estimators are useful tools in statistical inference. The presented theoretical results show that, in the context of model selection, these tools provide asymptotically unbiased criteria. These criteria are not robust in the sense of the bounded influence function, but this fact does not exclude the stability of the criteria with respect to other robustness measures. The computation of $Q_{\widehat{\theta}}$ could lead to serious difficulties, for example when considering various regression models to choose from. Such difficulties are implied by the double optimization in the criterion. Therefore, from the computation point of view, some other existing model selection criteria could be preferred. On the other hand, some performant computation techniques, involving such a double optimization, could arrive in the favor of using these new criteria also. These problems represent the topic of future research.

Acknowledgments

The author thanks the referees for a careful reading of the paper and for the suggestions leading to an improved version of the paper. This work was supported by a grant of the Romanian National Authority for Scientific Research, CNCS-UEFISCDI, project number PN-II-RU-TE-2012-3-0007.

Conflicts of Interest

The author declares no conflict of interest.

References

1. Pardo, L. *Statistical Inference Based on Divergence Measures*; Chapmann & Hall: Boca Raton, FL, USA, 2006.

2. Basu, A.; Shioya, H.; Park, C. *Statistical Inference: The Minimum Distance Approach*; Chapmann & Hall: Boca Raton, FL, USA, 2011.
3. Beran, R. Minimum Hellinger distance estimates for parametric models. *Ann. Stat.* **1977**, *5*, 445–463.
4. Tamura, R.N.; Boos, D.D. Minimum Hellinger distance estimation for multivariate location and covariance. *J. Am. Stat. Assoc.* **1986**, *81*, 223–229.
5. Simpson, D.G. Minimum Hellinger distance estimation for the analysis of count data. *J. Am. Stat. Assoc.* **1987**, *82*, 802–807.
6. Simpson, D.G. Hellinger deviance tests: Efficiency, breakdown points, and examples. *J. Am. Stat. Assoc.* **1989**, *84*, 104–113.
7. Toma, A. Minimum Hellinger distance estimators for multivariate distributions from Johnson system. *J. Stat. Plan. Inference.* **2008**, *183*, 803–816.
8. Lindsay, B.G. Efficiency versus robustness: The case of minimum Hellinger distance and related methods. *Ann. Stat.* **1994**, *22*, 1081–1114.
9. Basu, A.; Lindsay, B.G. Minimum disparity estimation for continuous models: Efficiency, distributions and robustness. *Ann. Inst. Stat. Math.* **1994**, *46*, 683–705.
10. Basu, A.; Harris, I.R.; Hjort, N.L.; Jones, M.C. Robust and efficient estimation by minimising a density power divergence. *Biometrika*. **1998**, *85*, 549–559.
11. Broniatowski, M.; Keziou, A. Parametric estimation and tests through divergences and duality technique. *J. Multivar. Anal.* **2009**, *100*, 16–36.
12. Toma, A.; Broniatowski, M. Dual divergence estimators and tests: Robustness results. *J. Multivar. Anal.* **2011**, *102*, 20–36.
13. Toma, A.; Leoni-Aubin, S. Robust tests based on dual divergence estimators and saddlepoint approximations. *J. Multivar. Anal.* **2010**, *101*, 1143–1155.
14. Mallows, C.L. Some comments on Cp. *Technometrics* **1973**, *15*, 661–675.
15. Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B* **1974**, *36*, 111–147.
16. Akaike, H. Information theory and an extension of the maximum likelihood principle. In Proceedings of the Second International Symposium on Information Theory, Akademiai Kaido, Budapest, 1973; Petrov, B.N., Csaki, I.F., Eds.; pp. 267–281.
17. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464.
18. Konishi, S.; Kitagawa, G. Generalised information criteria in model selection. *Biometrika* **1996**, *83*, 875–890.
19. Ronchetti, E. Robust model selection in regression. *Stat. Probab. Lett.* **1985**, *3*, 21–23.
20. Ronchetti, E.; Staudte, R.G. A robust version of Mallows' CP. *J. Am. Stat. Assoc.* **1994**, *89*, 550–559.
21. Maronna, R.A.; Martin, R.D.; Yohai, V.J. *Robust Statistics: Theory and Methods*; Wiley: New York, NY, USA, 2006.
22. Karagrigoriou, A.; Mattheou, K.; Vonta, F. On asymptotic properties of AIC variants with applications. *Open J. Stat.* **2011**, *1*, 105–109.

23. Mattheou, K.; Lee, S.; Karagrigoriou, A. A model selection criterion based on the BHHJ measure of divergence. *J. Stat. Plan. Inference* **2009**, *139*, 228–235.
24. Cressie, N.; Read, T.R.C. Multinomial goodness of fit tests. *J. R. Stat. Soc. Ser. B*. **1984**, *46*, 440–464.
25. Rüschendorf, L. On the minimum discrimination information theorem. *Stat. Decis.* **1984**, *1*, 163–283.
26. Liese, F.; Vajda, I. *Convex Statistical Distances*; BSB Teubner: Leipzig, Germany, 1987.
27. Toma, A.; Leoni-Aubin, S. Portfolio selection using minimum pseudodistance estimators. *Econ. Comput. Econ. Cybern. Stud. Res.* **2013**, *46*, 117–132.
28. Kallberg, D.; Leonenko, N.; Seleznjev, O. Statistical inference for Rényi entropy functionals. In *Conceptual Modelling and Its Theoretical Foundations*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7260, pp. 36–51.
29. Preda, V.; Dedu, S.; Sheraz, M. New measure selection for Hunt-Devolder semi-Markov regime switching interest rate models. *Physica A* **2014**, *407*, 350–359.
30. Liese, F.; Vajda, I. On divergences and informations in statistics and information theory. *IEEE Trans. Inf. Theory* **2006**, *52*, 4394–4412.
31. Hampel, F.R.; Ronchetti, E.; Rousseeuw, P.J.; Stahel, W. *Robust Statistics: The Approach Based on Influence Functions*; Wiley: New York, NY, USA, 1986.