On Measures of Divergence and the Divergence Information Selection Criterion

Kyriacos Mattheou and Alex Karagrigoriou
Department of Mathematics and Statistics, University of Cyprus, Cyprus

Abstract: The aim of this work is to develop a new model selection criterion using a general discrepancy based technique, by constructing an asymptotically unbiased estimator of the overall average discrepancy between the true and the fitted models. Furthermore, the lower bound for the mean squared error of prediction is established.

Keywords and phrases: DIC, power divergence, MSE of prediction

68.1 Introduction

The divergence measures are used as indices of similarity or dissimilarity between populations. They are also used either to measure mutual information concerning two variables or to construct model selection criteria. A model selection criterion can be constructed as an approximately unbiased estimator of an expected "overall discrepancy" (or divergence), a nonnegative quantity which measures the "distance" between the true model and a fitted approximating model. A well known discrepancy is Kullback-Leibler discrepancy that was used by Akaike (1973) to develop Akaike Information Criterion (AIC).

Measures of discrepancy or divergence between two probability distributions have a long history. A unified analysis was recently provided by Cressie and Read (1984) who introduced for both the continuous and the discrete case the so called power divergence family of statistics that depends on a parameter $\lambda$ and is used for multinomial goodness-of-fit tests. The additive and non-additive directed divergences of order $\alpha$ were introduced in the 60’s and the 70’s (Renyi, 1961 and Rathie and Kannappan, 1972). It should be noted that for $\lambda$ tending to 0 and for $\alpha$ tending to 1 the above measures become the Kullback-Leibler measure. Another family of measures is the $\Phi$-divergence known also as
Csiszar’s measure of information (Csiszar, 1963) the discrete form of which is given by \( I^c(P; Q) = \sum_{i=1}^{k} q_i \Phi(p_i/q_i) \), where \( \Phi \) is a real valued convex function on \([0, \infty]\) and \( P = (p_1, p_2, \ldots, p_k) \) and \( Q = (q_1, q_2, \ldots, q_k) \) are two discrete finite probability distributions. For various functions for \( \Phi \) the measure takes different forms. The Kullback-Leibler measure is obtained for \( \Phi(u) = u \log(u) \) while the additive directed divergence is obtained for \( \Phi(u) = \text{sgn}(\alpha - 1)u^\alpha \) and for the transformation \((\alpha - 1)^{-1}\log I^c\). For a comprehensive discussion on measures of divergence the reader is referred to Pardo (2006).

A new discrepancy measure was recently introduced by Basu et. al (1998). In this paper, we develop a new model selection criterion which is an approximately unbiased estimator of the expected overall power divergence that corresponds to Basu’s power divergence measure. Furthermore, we obtain a lower bound for the mean squared error (MSE) of prediction.

### 68.2 Basu’s Power Divergence Measure

One of the most recently proposed discrepancies is Basu’s Power Divergence [Basu et. al (1998)] which is defined as:

\[
d_a(g, f) = \int \left\{ f^{1+a}(z) - \left(1 + \frac{1}{a}\right) g(z) f^a(z) + \frac{1}{a} g^{1+a}(z) \right\} dz, \quad a > 0 (68.2.1)
\]

where \( g \) is the true model, \( f \) the fitted approximating model, and \( a \) a positive number. The discrete form of the measure is given by

\[
\sum_{i=1}^{k} \left\{ p_i^{1+a} - (1 + \frac{1}{a}) p_i^a q_i + \frac{1}{a} q_i^{1+a} \right\},
\]

where \( p_i \) and \( q_i \), \( i = 1, 2, \ldots, k \) are as in Section 1.1. Observe that the above measure takes the form \( \sum_{i=1}^{k} q_i^{1+a} \Phi(p_i/q_i) \) where \( \Phi(u) = u^{1+a} - (1 + a^{-1})u^a + a^{-1} \).

**Lemma 68.2.1** The limit of (68.2.1) when \( a \to 0 \) is the Kullback-Leibler divergence. Furthermore, the discrete form of the measure tends to the Kullback-Leibler measure for \( a \to 0 \) and for \( \Phi(u) = u \log(u) \).

It is easy to see that Basu’s measure satisfies the basic properties of measures, namely the properties of nonnegativity and the continuity. In particular, the value of measure is nonnegative while small changes in the distributions result in small changes in the measure. Finally, the value of the discrete measure is not affected by the simultaneous and equivalent reordering of the discrete masses which confirms the symmetry property of the Basu’s measure.

Consider a random sample \( X_1, \ldots, X_n \) from the distribution \( g \) and a candidate model \( f_t \) from a parametric family of models \( \{f_t\} \), indexed by an unknown
Measures of divergence and the DIC criterion

395

parameter \( t \in \Theta \). To construct the new criterion for goodness of fit we shall consider the quantity:

\[
W_t = \int \left\{ f_{t}^{1+a}(z) - \left(1 + \frac{1}{a}\right) g(z) f_{t}^a(z) \right\} dz, \quad a > 0.
\]

(68.2.2)

which is the same as (68.2.1) without the last term that remains constant irrespectively of the model \( f_t \) used. Observe that (68.2.2) can also be written as:

\[
W_t = \int f_{t}^{1+a}(z) dz - \left(1 + \frac{1}{a}\right) E_g(f_{t}^a(z)), \quad a > 0.
\]

(68.2.3)

Our target theoretical quantity that would be estimated by the new criterion is

\[
E\{W_t | t = \hat{\theta}\}
\]

(68.2.4)

which can be viewed as the average distance between \( g \) and \( f_t \) up to a constant and is known as the expected overall discrepancy between \( g \) and \( f_t \). In (68.2.4), \( \hat{\theta} \) is the estimator of \( t \) that minimizes an estimate of \( d_a(g, f_t) \) with respect to \( t \). Note that the estimator of \( t \) is obtained by minimizing (68.2.3) when the expectation is replaced by its sample analogue, namely \( n^{-1} \sum_{i=1}^{n} f_{t}^\alpha(X_i) \). In the theorem below, Basu et. al. (1998) provide the asymptotic properties of \( \hat{\theta} \).

Theorem 68.2.1 (Basu et. al. (1998)) Under regularity conditions, there exists estimator \( \hat{\theta} \) which is consistent and asymptotically normal with mean zero and variance \( J(\theta)^{-2} K(\theta) \), where under the assumption that the true distribution \( g \) belongs to the parametric family \( \{f_t\} \), \( \theta \) being the true value of the parameter and \( \xi = \int u_{\theta}(z) f_{\theta}^{1+a}(z) dz \),

\[
J(\theta) = \int [u_{\theta}(z)]^2 f_{\theta}^{1+a}(z) dz \quad \text{and} \quad K(\theta) = \int [u_{\theta}(z)]^2 f_{\theta}^{1+2\alpha}(z) dz - \xi^2.
\]

(68.2.5)

The Lemma below provides the derivatives of (68.2.3) in the case where \( g \) belongs to the family \( \{f_t\} \) (see Mattheou and Karagrigoriou (2006a)).

Lemma 68.2.2 For \( a > 0 \) and if the true distribution \( g \) belongs to the parametric family \( \{f_t\} \), the derivatives of (68.2.3) are:

\[
(a) \quad \frac{\partial W_t}{\partial t} = (a + 1) \left[ \int u_t(z) f_{t}^{1+a}(z) dz - E_g(u_t(z) f_{t}^a(z)) \right] = 0,
\]

\[
(b) \quad \frac{\partial^2 W_t}{\partial t^2} = (a + 1) \left\{ (a + 1) \int [u_t(z)]^2 f_{t}^{1+a}(z) dz - \int i_t f_{t}^{1+a} dz + \frac{E_g(i_t(z) f_{t}^a(z)) - E g\left(a [u_t(z)]^2 f_{t}^a(z)\right)}{} = (a + 1) J \right\}
\]

where \( u_t = \frac{\partial }{\partial t} (\log (f_t)) \), \( i_t = -\frac{\partial^2 }{\partial t^2} (\log (f_t)) \) and \( J = \int [u_t(z)]^2 f_{t}^{1+a}(z) dz \).
Theorem 68.2.2 Under the assumptions of Lemma (68.2.2) and for a p-dimensional parameter \( t \), the expected overall discrepancy at \( t = \hat{\theta} \) is given by

\[
E \left( W_t \mid t = \hat{\theta} \right) = W_\theta + \frac{(a+1)}{2} E \left( \left( \hat{\theta} - \theta \right) J \left( \hat{\theta} - \theta \right)' \right).
\] (68.2.6)

68.3 The divergence information criterion

In this section we introduce the new criterion and prove that it is an approximately unbiased estimator of (68.2.4). Due to the unknown true distribution \( g \), we estimate (68.2.3) by the empirical distribution function given by:

\[
Q_t = \int f_t^{1+a}(z) \, dz - \left( 1 + \frac{1}{a} \right) \frac{1}{n} \sum_{i=1}^{n} u_t(X_i) f_t^{a}(X_i).
\] (68.3.7)

Lemma 68.3.1 The derivatives of (68.3.7) are:

\[
(a) \quad \frac{\partial Q_t}{\partial t} = (a + 1) \left[ \int u_t(z) f_t^{1+a}(z) \, dz - \frac{1}{n} \sum_{i=1}^{n} u_t(X_i) f_t^{a}(X_i) \right], \quad a > 0,
\]

\[
(b) \quad \frac{\partial^2 Q_t}{\partial t^2} = (a + 1) \left\{ (a + 1) \int [u_t(z)]^2 f_t^{1+a}(z) \, dz - \int i_t f_t^{1+a}(z) \, dz 
+ \frac{1}{n} \sum_{i=1}^{n} i_t(z) f_t^{a}(z) - \frac{1}{n} \sum_{i=1}^{n} a [u_t(z)]^2 f_t^{a}(z) \right\}
\]

where \( u_t = \frac{\partial}{\partial t} \log (f_t) \) and \( i_t = - (u_t)' = - \frac{\partial^2}{\partial t^2} \log (f_t) \).

It is easy to see that by the weak law of large numbers, as \( n \to \infty \), we have:

\[
\left[ \frac{\partial Q_t}{\partial t} \right]_\theta \to_P \left[ \frac{\partial W_t}{\partial t} \right]_\theta \quad \text{and} \quad \left[ \frac{\partial^2 Q_t}{\partial t^2} \right]_\theta \to_P \left[ \frac{\partial^2 W_t}{\partial t^2} \right]_\theta.
\] (68.3.8)

The consistency of \( \hat{\theta} \) and (68.3.8) can be used to evaluate the expectation of the empirical estimator evaluated at the true point \( \theta \).

Theorem 68.3.1 Under the assumptions of Lemma (68.2.2), the expectation of \( Q_t \) evaluated at \( \theta \) is given by

\[
EQ_\theta \equiv E (Q_t \mid t = \theta) = EQ_\hat{\theta} + \frac{a+1}{2} E \left( \left( \theta - \hat{\theta} \right) J \left( \theta - \hat{\theta} \right)' \right).
\]

The asymptotically unbiased estimator of \( E \left( W_t \mid t = \hat{\theta} \right) \) is provided in the theorem below (see Mattheou and Karagrigoriou (2006b)).
Theorem 68.3.2 An asymptotically unbiased estimator of the expected overall discrepancy evaluated at $\hat{\theta}$ is given by

$$DIC = Q_{\hat{\theta}} + (a + 1)(2\pi)^{-\frac{a}{2}} \left( \frac{1 + a}{1 + 2a} \right)^{1 + \frac{2}{a}} p.$$  \hspace{1cm} (68.3.9)

68.4 Lower bound of the MSE of prediction

Let $X_j$ be the design matrix of the model $Y = X_j\beta + \varepsilon$ where $\beta = (\beta_0, \beta_1, \beta_2, \ldots)'$, $\varepsilon \sim N(0, \sigma^2 I)$ and $I$ is the infinite dimensional identity matrix.

Let $V(j) = \{ \beta(j), \text{ s.t. } \beta(j) = (\beta_0, 0, ..., \beta_{j1}, 0, ..., \beta_{j_k}, 0, ...) \}$ be the subspace that contains only the $k_j + 1$ parameters $\beta_{ji}$ involved in the model and let $\beta^{(n)}$ to be the projection of $\beta$ on $V(j)$.

The prediction $\hat{y}$ is given by $\hat{y} = X_j\hat{\beta}$, where the estimator of $\beta^{(n)}$ obtained through a set of observations $(X_{ij1}, X_{ij_k}, y_i)$, $i = 1, 2, \ldots, n$ is denoted by $\hat{\beta} = \left( \hat{\beta}_0, 0, ..., \hat{\beta}_{j1}, 0, ..., \hat{\beta}_{j_2}, 0, ..., \hat{\beta}_{j_k}, 0, ... \right)'$.

The mean squared error (MSE) of prediction and the average MSE of prediction are defined respectively by

$$Q_n(j) = \mathbb{E} \left[ (\hat{y}_{n+1} - y_{n+1})^2 | X \right] - n\sigma^2 \quad \text{and} \quad L_n(j) = \mathbb{E} (Q_n(j)).$$

Lemma 68.4.1 Under the notation and conditions of this section we have that

$$Q_n(j) = \| \hat{\beta} - \beta \|_{M_n(j)}^2 \quad \text{and} \quad L_n(j) = \mathbb{E} \| \hat{\beta} - \beta \|_{M_n(j)}^2,$$

where $M_n(j) = X_j'X_j$ and $\| A \|_R^2 = A'RA$.

The Lemma below provides a lower bound for the MSE of prediction. In particular, we show that $Q_n(j)$ is asymptotically never below the quantity $L_n(j^*) = \min_j L_n(j)$.

Lemma 68.4.2 Let $L_n(j^*) = \min_j L_n(j)$. Under certain regularity conditions, we have that for every $\delta > 0$

$$\lim_{n \to \infty} P \left[ \frac{Q_n(j)}{L_n(j^*)} > 1 - \delta \right] = 1.$$

68.5 Discussion

Note that the family of candidate models is indexed by a single parameter $a$. The value of $a$ dictates to what extent the estimating methods become more robust than the maximum likelihood methods. One should be aware of the fact
that the larger the value of $a$ the bigger the efficiency loss. As a result one should be interested in small values of $a \geq 0$, say between zero and one.

The proposed DIC criterion could be used in applications where outliers or contaminated observations are involved. The prior knowledge of contamination may be useful in identifying an appropriate value of $\alpha$. Preliminary simulations with a 10% contamination proportion show that DIC has a tendency of underestimation in contrast with AIC which overestimates the true model.

References


