Density Power Downweighting and Robust Inference: Some New Strategies

Saptarshi Roy, Kaustav Chakraborty, Somnath Bhadra and Ayanendranath Basu

Abstract: Preserving the robustness of the procedure has, at the present time, become almost a default requirement for statistical data analysis. Since efficiency at the model and robustness under misspecification of the model are often in conflict, it is important to choose such inference procedures which provide the best compromise between these two concepts. Some minimum Bregman divergence estimators and related tests of hypothesis seem to be able to do well in this respect, with the procedures based on the density power divergence providing the existing standard. In this paper we propose a new family of Bregman divergences which is a superfamily encompassing the density power divergence. This paper describes the inference procedures resulting from this new family of divergences and makes a strong case for the utility of this divergence family in statistical inference.

Keywords: Minimum Distance Inference, Density Power Divergence, Robustness, Optimal Tuning Parameter, Logarithmic $\phi$-DPD

Introduction

In statistical modeling, parameter estimation is an inevitable and formidable task. Accurate estimation of the model facilitates the characterization and the subsequent understanding of the mechanism that generates the observed data. Statistical distances can be useful tools for the estimation of the model parameters.

Statistical distances can be naturally applied to the case of parametric statistical inference. The most important idea in parametric minimum distance inference is the quantification of the degree of closeness between the sample data and parametric model as a function of an unknown set of parameters through a suitable distance-like measure. Thus the estimate of the parameter is obtained by minimizing this “distance” over the parameter space.

It is worthwhile to mention here that the class of distances which we will consider are not mathematical metrics in the strict sense of the term. They may not be symmetric in their arguments and may not satisfy the triangle inequality. The only properties that we require of these measures are that they should be nonnegative and should equal zero if and only if the arguments are identically equal. However, we will, somewhat loosely, continue to call them distances, or “statistical distances”. In a practical sense, the word “divergence” is a good descriptor of these measures. We will, in fact, use the “minimum distance” and the “minimum divergence” terminologies interchangeably.

Density-based divergences form a special class of statistical distances. Several minimum distance estimators in this family have high model efficiency. In particular, the Maximum Likelihood Estimator (MLE) also belongs to the class of density-based minimum distance estimators, being the minimizer of the likelihood disparity (Lindsay, 1994), which is a version of the Kullback-Leibler divergence. But one of the major drawbacks of the MLE is that it is notoriously nonrobust and even a small proportion of outlying observations can lead to meaningless inference. In fact it is the failure of the classical methods like maximum likelihood to deal with outliers and mild deviations from the model which had led to the emergence of the field of robustness; see, for example, Huber and Ronchetti (2009), Hampel et al. (1986), Maronna et al. (2019) and Basu et al. (2011). However, some of the other members of the class of minimum distance estimators have been observed to do
much better in the sense of combining strong robustness with high model efficiency. See, for example, Csiszar (1963), Ali and Silvey (1966), Lindsay (1994), Pardo (2005) and Basu et al. (2011) for a description of the φ-divergence class of minimum distance measures.

A more modern class of minimum distance estimators is based on the family of Bregman divergences. The Bregman divergence (Bregman, 1967) is a distance like measure between points and has been used in mathematics and information theory for some time. When the points are represented by probability distributions, the corresponding Bregman divergence is a statistical distance. See, for example, Jones and Byrne (1990), Csiszar (1991), Banerjee et al. (2005) and Stummer and Vajda (2012) for some examples of statistical and related applications of the Bregman divergence. The principal representatives of Bregman divergence estimators in the current statistical literature are the Minimum Density Power Divergence Estimators (MDPDEs), based on the Density Power Divergence (DPD) class of Basu et al. (1998). Over the last two decades, this class of divergences has provided a popular and frequently used method to balance the trade-off between robustness and efficiency in parameter estimation, hypothesis testing and related inference. The minimum divergence estimators based on the DPD have been shown to provide a high degree of stability under model misspecification, often with minimal loss in model efficiency. Our primary purpose in this paper is to refine the minimum distance procedure based on the DPD, so as to achieve even better compromise between efficiency and robustness.

The Bregman Divergence

Consider a parametric family of densities \( \mathcal{F} = \{ f_\theta : \theta \in \Theta \subseteq \mathbb{R}^p \} \). Let \( X_1, X_2, \ldots, X_n \) be i.i.d. observations from a distribution \( G \) having probability density function (pdf) \( g \). For the sake of a unified notation we will continue to use the term pdf irrespective of whether the distribution of \( G \) is continuous or discrete. Let the common support of \( g \) and \( f_0 \) be \( \mathcal{X} \subseteq \mathbb{R} \). The Bregman divergence between the density \( g \) and model density \( f_0 \) is given by:

\[
D_B(g, f_0) = \int \left( B(g(x)) - B(f_0(x)) - (g(x) - f_0(x)) B'(f_0(x)) \right) dx,
\]

where the index function \( B(\cdot) \) is strictly convex and \( B'(\cdot) \) represents its first derivative with respect to its argument. In practice, where \( f_0 \) is the pdf of the parametric family, \( g \) is the true density, the minimization of the above divergence over the parameter space \( \Theta \) will generate the corresponding minimum distance functional which can lead to meaningful inference, depending on the form of the function \( B(\cdot) \). The DPD, defined later in this section, is a special case of the Bregman divergence for \( B(y) = \frac{y^{1+\alpha} - y}{\alpha}, \alpha \neq 0 \).

When the model is differentiable, the general estimating equation under the divergence in Equation 1 is:

\[
\int u_\theta(x) B'(f_0(x)) f_0'(x) dx - \int u_\theta(x) B'(f_0(x)) f_\theta(x) g(x) dx = 0,
\]

or equivalently:

\[
\int u_\theta(x) B'(f_0(x)) f_0'(x) dx - \int u_\theta(x) B'(f_0(x)) f_\theta(x) dG(x) = 0.
\]

where, \( u_\theta(x) = \nabla \log f_\theta(x) \) is the score function of the model \( f_\theta(x) \), \( \nabla \) represents derivative with respect to \( \theta \) and \( B''(\cdot) \) represents the second derivative of \( B(\cdot) \) with respect to its argument. Since \( G \) is unknown, we construct an empirical version of the divergence in Equation 1, or the estimating equation given in Equation 3, by replacing \( G \) (the true data generating distribution) by its empirical counterpart \( G_n \). This leads to a class of unbiased (under the model) estimating equations:

\[
\sum_{i=1}^{n} u_\theta(x_i) B'(f_0(x_i)) f_0'(x_i) = 0.
\]

The root of the Equation 4 is defined to be the Minimum Bregman Divergence Estimator (MBDE). Here the robustness of the corresponding minimum distance estimator may be at least partially understood by observing the effect of the downweighting function \( B''(f_\theta(x)) f_\theta(x) \) on \( u_\theta(x) \) for less probable values of \( x \) under \( f_\theta \). For the DPD, this weight becomes \( (\alpha + 1) f_\theta''(x) \).

In this paper we attempt to find a refinement of the DPD downweighting scheme and, by reconstruction, a corresponding divergence. We will show that the corresponding minimum distance procedure provides a better compromise between robustness and efficiency in many cases compared to the minimum density power divergence estimator (MDPDE).

The Density Power Divergence

As mentioned earlier, the Density Power Divergence (DPD) is obtained by substituting \( B(y) = \frac{y^{1+\alpha} - y}{\alpha} \) in
Equation 1. The general form of this divergence, as a function of a nonnegative tuning parameter $\alpha$, is:

$$DPD_u(g, f_\theta) = \int \left( f_\theta^{1+\alpha} - \left(1 + \frac{1}{\alpha} \right) g f_\theta^\alpha + \frac{1}{\alpha} g^1 \right) dx.$$  \hfill (5)

For simplicity we have dropped the dummy variable in the above equation. One can define the minimum DPD functional $T_u(G)$ at $G$ through the relation:

$$DPD_u(g, f_\theta) = \inf_{\phi_u} DPD_u(g, f_\theta).$$  \hfill (6)

Under the estimation setup of this paper, the empirical objective function, ignoring the terms independent of $\theta$, becomes:

$$\int f_\theta^{1+\alpha} - \left(1 + \frac{1}{\alpha} \right) \frac{1}{n} \sum_{i=1}^{n} f_\theta (X_i).$$

and under differentiability of the model, the estimating equation becomes (by equating the negative of the derivative of the above objective function to 0):

$$\frac{1}{n} \sum_{i=1}^{n} u_{\phi} (X_i) f_\theta (X_i) - \int u_{\phi} (x) f_\theta^{1+\alpha} (x) dx = 0.$$  \hfill (7)

It is evident that as $\alpha \to 0^+$, Equation 7 converges to the maximum likelihood score equation:

$$\frac{1}{n} \sum_{i=1}^{n} u_{\phi} (X_i) = 0.$$  \hfill (8)

Note that in the part involving real data in Equation 7, a downweighting effect is exerted on the score function $u_{\phi} (x)$ by the factor $f_\theta^{1+\alpha} (x)$. This downweighting philosophy will be crucial for developing the new class of procedures. Note that there is no downweighting for the case $\alpha = 0$.

The asymptotic properties of the MDPDE have been well studied and are available, for example, in Basu et al. (2011), where the asymptotic distribution of the MDPDE has been explicitly derived. It is useful to note that the MDPDE solves an estimating equation of the form

$$\sum_{i=1}^{n} \psi (X_i, \theta) = 0,$$

where:

$$\psi (x, \theta) = u_{\phi} (x) f_\theta^{1+\alpha} (x) - \int u_{\phi} (x) f_\theta^{1+\alpha} (x) dx.$$  \hfill (9)

Hence it belongs the class of M-estimators. So, the asymptotic properties of the MDPDE also follow from M-estimator theory.

A New Divergence

Our key philosophy for constructing new divergences and estimation strategies involves manipulating the downweighting factor $B''(f_\theta(X_i)) f_\theta(X_i)$ in Equation 4. Here we are going to develop a stronger downweighting effect compared to the MDPD estimating equation. Our exploration will generate an estimation scheme with two tuning parameters and we will explore the possibility of coming up with specific candidates which might beat the MDPDEs both in terms of efficiency and robustness.

Choosing the $B$ Function

The downweighting effect on the score $u_{\phi}(x)$ applied by the MDPD estimating equation is $f_\theta^{1+\alpha} (x)$. As we want to impose a stronger downweighting in relation to this, we wish to choose the $B$ function (or rather, the $B''$ function) so that as $x \to 0^+$, $xB''(x)$ converges to zero faster than $x^\alpha$ for $\alpha > 0$ fixed. (Note, from Equation 4, the downweighting term for $u_{\phi}(x)$ in the general Bregman divergence is $f_\theta B''(f_\theta(x))$). In particular, we will assume the following conditions on $B''$:

- (P1) $B''(x) > 0 \forall x > 0$, so that $B$ is a strictly convex function over $\mathbb{R}^+$
- (P2) $xB''(x)$ is an increasing function over $x$ in $(0,\infty)$. Thus the less likely observations will be downweighted more
- (P3) For all $\beta \in (0,1)$, $\lim_{x \to 0^+} \frac{x B''(x)}{x^{\beta}} = 0$, i.e., the Bregman formulation attaches weights to the score function which go to zero at a rate faster than the corresponding weights in the MDPD estimating equation
- (P4) $B''(x) = x^\beta \phi(x, \gamma)$ ($0 < \beta \leq 1, 0 < \gamma \leq 1$). Where $\phi(x, \gamma)$ is a continuous and positive function over $\gamma \in (0,1]$ and $x > 0$. Furthermore, we demand $\lim_{x \to 0^+} \phi(x, \gamma) = \frac{1}{x}$.

To prove that such choice of $B(\cdot)$ satisfying (P1)-(P4) can help us generate divergences which have the desired properties and provide superior inference compared to the DPD, let us first demonstrate the general asymptotic properties of the minimum Bregman divergence estimators. For ease of representation, we refer to the divergence generated by the $B(\cdot)$ function satisfying (P1) to (P4) as $\phi$-DPD.

General Asymptotic Properties of the MBDE

We need some regularity assumptions to prove the asymptotic properties of the general MBDE, which we list below:
The pdfs $f_0$ of $X$ have common support, so that the set $\{x | f_0(x) > 0\}$ is independent of $\theta$. The distribution $G$ is also supported on $\mathcal{X}$, on which the corresponding density $g$ is greater than zero.

There is an open subset $\omega$ of the parameter space $\Theta$, containing the best fitting parameter $\theta^0$ such that for almost all $x \in \mathcal{X}$ and all $\theta \in \omega$ the density $f_0(x)$ is three times differentiable with respect to $\theta$ and the third partial derivatives are continuous with respect to $\theta$. (The best fitting parameter $\theta^0$ depends on the index function $B(\cdot)$ also, but we suppress that notation for brevity.)

The integrals $\int B''(f_0(x))f_0(x)dx$ and $\int B''(f_0(x))f_0(x)g(x)dx$ can be differentiated with respect to $\theta$ and the derivatives can be taken under the integral sign.

The $p \times p$ matrix $J_{\theta}(\theta)$ defined by:

$$J_{\theta}(\theta) = E_x\left[\nabla_\theta \left(\int \left[ B'(f_0(x))f_0(x) - B(f_0(x)) \right] dx \right) \right]$$

is positive definite where $E_x$ represents the expectation under the density $g$. Where $\nabla_\theta$ represents the partial derivative with respect to the indicated components of $\theta$.

There exist functions $M_\theta(X)$, $j, k, l = 1, \ldots, p$, such that:

$$|\nabla_\theta \left(\int \left[ B'(f_0(x))f_0(x) - B(f_0(x)) \right] dx - B'(f_0(x)) \right)| \leq M_\theta(X) ; \forall \theta \in \omega$$

where, $E_x[M_\theta(X)] < m_{jkl} < \infty \forall j, k, l$.

**Theorem 1**

Under the conditions (A1)-(A5), the following results hold:

(a) The MBDE estimating equation given in Equation 4 has a consistent sequence of roots $\hat{\theta}_n$.

(b) $\sqrt{n}(\hat{\theta}_n - \theta^0)$ has an asymptotic multivariate normal distribution with mean vector zero and covariance matrix $J_{\theta}^{-1}K_\theta J_{\theta}^{-1}$, where $J_{\theta} = J_{\theta}(\theta^0)$, $K_\theta = K_{\theta}(\theta^0)$, $K_{\theta}(\theta) = Var_x(\mu_0(X) f_0(X) B'(f_0(X)))$.

When $g = f_0$ for some $\theta \in \Theta$ then the above expressions simplify to:

$$J_{\theta} = \int u_0 \mu_0^* f_0^* B'(f_0). K_{\theta} = \int u_0 \mu_0^* f_0^* \left(B'(f_0)\right)^2 - \zeta_{n} \zeta_{n}^{\top}, \zeta_{n} = \int u_0 f_0^* B'(f_0). \ (10)$$

We are now going to establish that the DPD belongs to the class of $\phi$-DPD. We will also show that under certain conditions a judicial choice of $\phi(\cdot)$ yields estimators which may fit with our aims. Now our unbiased estimating equation for $\phi$-DPD is:

$$\frac{1}{n} \sum_{i=1}^{n} u_i f_i(x_i) f_i^* \phi(f_i(x_i), \gamma) - \int u_i(x) f_i^* \phi(f_i(x), \gamma) dx = 0,$$

and under assumptions similar to (A1)-(A5) and $g = f_0$, the expressions in Equation 10 simplify to:

$$J_{\theta} = \int u_0 \mu_0^* f_0^* \phi(f_0, \gamma), K_{\theta} = \int u_0 \mu_0^* f_0^* \phi(f_0, \gamma) - \zeta_{\theta} \zeta_{\theta}^{\top}, \zeta_{\theta} = \int u_0 f_0^* \phi(f_0, \gamma). \ (12)$$

A straightforward simplification of the expressions in part (b) of Theorem 1 under $\phi$-DPD leads to the general expressions:

$$K_{\theta} = \int u_0 \mu_0^* f_0^* \phi(f_0, \gamma) g - \zeta_{\theta} \zeta_{\theta}^{\top}, \zeta_{\theta} = \int u_0 f_0^* \phi(f_0, \gamma) g, \ (13)$$

and:

$$J_{\theta} = \int u_0 \mu_0^* f_0^* \phi(f_0, \gamma) + \int \left(\kappa_{\theta} - \beta \mu_0 \mu_0^* (g - f_0) f_0^* \phi(f_0, \gamma) \right). \ (14)$$

$$\phi(x, \gamma) = \frac{\partial \phi(x, \gamma)}{\partial x}. \ (15)$$

**Remark**

Notice that $1 + \frac{\phi'(x, \gamma)}{\phi(x, \gamma)} = \frac{1}{\phi(x, \gamma)} \frac{\partial}{\partial x} [\phi(x, \gamma)]$. If

$$\lim_{\gamma \to \gamma^*} \text{ and } \frac{\partial}{\partial x} \text{ are interchangeable for } \phi(\cdot) \text{ then by (P4) it can be concluded that } \kappa_{\theta} \text{ converges to } i\omega \text{ as } \gamma \to \gamma^*. \ (16)$$
Theorem 2

If \( u_0(x) f_0(x)^{1+\beta}, \) \( u_0(x) u_0(x)^x f_0(x)^{1+\beta}, \) \( u_0(x)u_0(x)^x f_0(x)^{1+\beta} \) are integrable and \( f_0(x)\delta(f_0(x), \gamma) \) is bounded by some universal constant then the following hold:

(a) The usual DPD defined in Equation 5 is a special limiting case of \( \phi \)-DPD

(b) If \( g = f_0 \) for some \( \theta \in \Theta \) and if for the DPD there exists \( \alpha, \beta \) such that the Asymptotic Relative Efficiency (ARE) of the estimator under tuning parameter \( \beta \) is greater than that of the estimator under tuning parameter \( \alpha \), then there exists \( \gamma \) such that \( \phi \)-DPD with tuning parameter \( (\beta, \gamma) \) generates an estimator with higher ARE than the MDPDE with tuning parameter \( \alpha \).

If \( \phi(x, \gamma) = \frac{1}{\gamma} \log \left( 1 + \frac{\gamma x}{y} \right) \), then for \( x > 0 \), \( \phi(x, \gamma) < 1 \) (as \( \log(1 + y) < y \) for \( y > 0 \)) and \( u_0(x) f_0(x)^{1+\beta}, \) \( u_0(x)u_0(x)^x f_0(x)^{1+\beta} \) are integrable under standard parametric models. So Theorem 2 holds for such a choice of \( \theta(x) \). Symbolically, the divergence generated by the \( B(\cdot) \) function obtained through this formulation will be referred to as the logarithmic \( \phi \)-DPD (or \( L_{\phi} \)-DPD). We will denote this divergence between the densities \( g \) and \( f \) corresponding to tuning parameters \( \beta \) and \( \alpha \) as \( L_{\phi} \)-DPD(\( \beta, \gamma \), \( g, f \)).

Our choice for \( B^\alpha(\cdot) \) in the \( L_{\phi} \)-DPD case is \( B^\alpha(x) = \frac{1}{\gamma} x^\alpha \log \left( 1 + \frac{\gamma x}{y} \right) \) if \( 0 < \beta \leq 1, \) \( 0 < \gamma \leq 1, \) \( x > 0 \). The corresponding \( B \) function may be expressed in the integral form as:

\[
B(x) = \frac{1}{\gamma} \int \int f^\alpha \log \left( 1 + \frac{\gamma x}{y} \right) dx dt.
\]

(16)

Obviously other choices are possible, but we have found the \( L_{\phi} \)-DPD to be a very useful divergence for our purpose and for the rest of the paper all our illustrations will be in relation to the \( L_{\phi} \)-DPD. We will refer to the corresponding minimum distance estimator as \( M_{\phi} \)-DE.

The Influence Function of \( M_{\phi} \)-DE

It is easy to see that the \( M_{\phi} \)-DE is also an M-estimator. Let the minimum \( L_{\phi} \)-DPD functional \( T_{\beta, \gamma}(G) \) be defined as:

\[
L_{\phi} \text{DPD}_{\beta, \gamma}(g, f; G) = \inf_{\phi \in \Theta} L_{\phi} \text{DPD}_{\beta, \gamma}(g, f; G).
\]

Under \( g = f_0 \) the influence function of this minimum distance estimator simplifies to:

\[
IF(y, \tau, G) = \left[ \frac{u_0(y) f_0^{1+\beta}}{\beta} \log \left( 1 + \frac{\gamma f_0}{f_0(y)} \right) \right]^{-1}
\]

for \( 0 < \beta \leq 1,0 < \gamma \leq 1 \). If \( \beta \rightarrow 0 \) and \( \gamma \rightarrow 0 \) then the expressions in Equation 13 are finite if \( \frac{u_0(x) f_0^{1+\beta}(y) \log \left( 1 + \frac{\gamma f_0}{f_0(y)} \right)}{\beta} \) is finite which is indeed the case for most parametric models suggesting the observed robustness of the \( M_{\phi} \)-DE under those parametric models.

In Fig. 1 it is clearly seen that the tuning parameter \( \beta \) has a significant impact on the robustness of the estimator and the influence functions redescend faster for larger values of \( \beta \). On the other hand, for fixed \( \beta \) the influence functions are somewhat closer for different \( \gamma \) as seen in Fig. 2. It suggests that \( \gamma \) has a less pronounced impact on robustness than \( \beta \), although the graphs in Fig. 2 indicate that larger \( \gamma \) lead to relatively stronger downweighting.

The Breakdown Point Under the Location Model

Now we will establish the breakdown point of the minimum \( L_{\phi} \)-DPD functional under the location family of densities \( \tau = \{ f_0(x) = f(x-\theta): \theta \in \Theta \} \). Let \( B(\cdot) \) be the function defined in Equation 16. Define the quantities:

\[
\tilde{B}(x, \theta) = \tilde{B}(x, \theta) dx = M_{1, \gamma}^{(1)},
\]

\[
\tilde{B}(f(x-\theta)) + (\varepsilon - 1)f(x-\theta)B'(f(x-\theta)) dx
\]

\[
= M_{1, \gamma}^{(2)},
\]

\[
\tilde{B}(f(x)) + (\varepsilon - 1)f(x)B'(f(x)) dx
\]

\[
= M_{1, \gamma}^{(3)}.
\]

Define \( d(g, f) = B(g) - B(f) \) and let \( D(g, f) = \tilde{d}(g, f) \). From Equation 16 we have \( d(g, 0) = \lim_{\varepsilon \rightarrow 0} d(g, f) = B(g) \).

Consider the contamination model \( H_{\varepsilon_0} = (1-\varepsilon)G + \varepsilon K_\varepsilon \), where \( K_\varepsilon \) is a sequence of contaminating distributions. Let \( h_{\varepsilon_0}, g \) and \( k_\varepsilon \) be the corresponding densities. We say that there is breakdown in the minimum \( L_{\phi} \)-DPD functional for \( \varepsilon \) level contamination if there exists a sequence \( K_\varepsilon \) such that \( \tilde{T}_{\beta, \gamma}(H_{\varepsilon_0}) \rightarrow \infty \) as \( n \rightarrow \infty \). We write below \( \varepsilon_0 = T_{\beta, \gamma}(H_{\varepsilon_0}) \) and assume that the true distribution belongs to the model family, i.e., \( g = f_0 \). We make the following assumptions:
• (BP1) \( \int \min \{ f_\theta(x), k_\theta(x) \} dx \to 0 \) as \( n \to \infty \) uniformly for \( |\theta| \leq c \) for any fixed \( c \), i.e., the contamination distribution is asymptotically singular to the true distribution and to specified models within the parametric family

• (BP2) \( \int \min \{ f_\theta(x), f_\theta(x) \} dx \to 0 \) as \( n \to \infty \) as \( n \to \infty \), i.e., large values of \( \theta \) give distributions which become asymptotically singular to the true distribution

• (BP3) The contaminating sequence \( \{ k_n \} \) is such that:

\[
D(\epsilon k_n, f_\theta) \geq D(\epsilon f_\theta, f_\theta) = M^{(0)}_{f,\theta} - M^{(0)}_{f,\theta}(c,\epsilon)
\]

for any \( \theta \in \Theta \) and \( 0 < \epsilon < 1 \) and

\[
\limsup_{n \to \infty} B(\epsilon k_n) \leq M^{(0)}_{f,\theta}.
\]

**Theorem 3**

Under the assumptions (BP1)-(BP3) above, the asymptotic breakdown point \( \varepsilon^* \) of the L\( \Phi \)DPD functional is at least 0.5 at the location model.
Simulation Study Under LφDPD and the Advantages of MLφDE

Description and Results

Here we have performed a simulation study to analyze the performance of the LφDPD and the associated minimum distance estimators under the N(µ, 1) model at a given level of contamination. In the following study data are generated from two normal mixtures, 0.9 N(0, 1) + 0.1 N(5, 1) and 0.8 N(0, 1) + 0.2 N(5, 1), where N(0, 1) represents the target distribution and the second component is the contamination. The sample size is 50. The empirical MSE for the location model has been calculated by replicating the process 1000 times, evaluating the estimate for each replication and taking average squared error loss against the target value, i.e., µ = 0. In Table 1 the theoretical asymptotic relative efficiency of minimum LφDPD estimator and MDPDE is shown for different values of (β, γ) while in Table 2 and 3 the simulated mean square errors are presented under contaminated normal data under two different contamination levels.

The LφDPD Versus the DPD

We briefly note our observations as may be evident from Table 1 and 2. The asymptotic efficiencies of the minimum divergence estimators decrease with increasing β and increasing γ. Note that given an αЄ(0, 1), it may be possible to choose βЄ(0, α) and γЄ(0, 1) so that, in relation to our numerical study, MLφDEβ, γ beats MDPDEα both in terms of asymptotic model efficiency and the empirical mean square error under contamination. As an illustration, consider MDPDE0.5 in the first contaminated model. The corresponding MSE and asymptotic relative efficiency are 0.0294 and 83.8% respectively. Now choose the LφDPD parameter (β, γ) = (0.3, 0.01). In this case, the corresponding MSE and efficiency of the MLφDE are 0.0281 and 89% respectively. Similarly MLφDE0.2,0.04 appears to dominate MDPDE0.4 both in terms of asymptotic efficiency and empirical mean square error. In fact, for practically all the MDPDEs that are considered in the Tables 1 and 2 (as also in Tables 1 and 3), there exists a better MLφDE, both in terms of asymptotic model efficiency and obtained mean square error under contamination. In most of these cases there are several (β, γ) combinations which provide the domination over a given MDPDE. Tables 2 and 3 also show that the robust minimum distance estimators hold out well against the outliers at both 10 and 20 percent contamination. Simulation results not presented here indicate that the same holds for higher levels of contamination smaller than 1/2, a consequence of the high breakdown point of the method under location models.

Table 1: Asymptotic relative efficiency of the MLφDE and DPDE (%) for different (β, γ) under N(0,1) location model. Here the γ = 0 column represents the MDPDE

| β | γ = 0 | γ = 0.01 | γ = 0.02 | γ = 0.03 | γ = 0.04 | γ = 0.05 | γ = 0.06 | γ = 0.07 | γ = 0.08 |
|---|---|---|---|---|---|---|---|---|---|
| 0.1 | 98.8 | 95.7 | 94.0 | 92.8 | 91.7 | 90.8 | 90.0 | 89.4 | 88.7 |
| 0.2 | 95.9 | 92.6 | 90.8 | 89.5 | 88.4 | 87.4 | 86.6 | 85.9 | 85.3 |
| 0.3 | 92.1 | 89.0 | 87.2 | 85.9 | 84.8 | 83.9 | 83.0 | 82.3 | 81.7 |
| 0.4 | 88.0 | 85.2 | 83.5 | 82.2 | 81.1 | 80.2 | 79.4 | 78.7 | 78.1 |
| 0.5 | 83.8 | 81.3 | 79.7 | 78.5 | 77.5 | 76.6 | 75.9 | 75.2 | 74.6 |
| 0.6 | 79.7 | 77.4 | 76.0 | 74.9 | 74.0 | 73.2 | 72.5 | 71.8 | 71.2 |
| 0.7 | 75.7 | 73.8 | 72.5 | 71.4 | 70.6 | 69.8 | 69.1 | 68.5 | 68.0 |
| 0.8 | 71.9 | 70.2 | 69.0 | 68.1 | 67.3 | 66.6 | 66.0 | 65.4 | 64.9 |
| 0.9 | 68.3 | 66.9 | 65.8 | 64.9 | 64.2 | 63.6 | 63.0 | 62.5 | 62.0 |
| 1 | 65.0 | 63.7 | 62.7 | 61.9 | 61.2 | 60.7 | 60.1 | 59.7 | 59.2 |

Table 2: Empirical MSE of the MLφDE and DPDE for different values of (β, γ) under 10% contaminated data for location model. Here they = 0 column represents the MDPDE

| β | γ = 0 | γ = 0.01 | γ = 0.02 | γ = 0.03 | γ = 0.04 | γ = 0.05 | γ = 0.06 | γ = 0.07 | γ = 0.08 |
|---|---|---|---|---|---|---|---|---|---|
| 0.1 | 0.1000 | 0.0293 | 0.0278 | 0.0290 | 0.0259 | 0.0271 | 0.0278 | 0.0282 | 0.0259 |
| 0.2 | 0.0560 | 0.0273 | 0.0277 | 0.0254 | 0.0252 | 0.0248 | 0.0268 | 0.0260 | 0.0266 |
| 0.3 | 0.0360 | 0.0281 | 0.0257 | 0.0267 | 0.0277 | 0.0264 | 0.0279 | 0.0268 | 0.0266 |
| 0.4 | 0.0268 | 0.0267 | 0.0261 | 0.0265 | 0.0270 | 0.0261 | 0.0289 | 0.0273 | 0.0265 |
| 0.5 | 0.0294 | 0.0277 | 0.0276 | 0.0276 | 0.0276 | 0.0307 | 0.0284 | 0.0296 | 0.0291 |
| 0.6 | 0.0275 | 0.0272 | 0.0298 | 0.0307 | 0.0293 | 0.0305 | 0.0295 | 0.0293 | 0.0296 |
| 0.7 | 0.0277 | 0.0276 | 0.0296 | 0.0294 | 0.0301 | 0.0320 | 0.0311 | 0.0311 | 0.0300 |
| 0.8 | 0.0292 | 0.0305 | 0.0327 | 0.0308 | 0.0342 | 0.0300 | 0.0288 | 0.0336 | 0.0315 |
| 0.9 | 0.0309 | 0.0299 | 0.0313 | 0.0345 | 0.0309 | 0.0295 | 0.0358 | 0.0326 | 0.0320 |
| 1 | 0.0313 | 0.0320 | 0.0350 | 0.0362 | 0.0369 | 0.0361 | 0.0368 | 0.0340 | 0.0335 |
Algorithm for Finding the Optimal \( (\beta, \gamma) \)

The L\( \beta \)DPD can generate many different kinds of estimators, starting from the most efficient estimator to highly robust estimators. For example, in the limit \( \gamma \rightarrow 0 \) and \( \beta \rightarrow 0 \), one gets the likelihood disparity which is minimized by the classical maximum likelihood estimator. On the other hand, relatively larger values of \( \beta \) and \( \gamma \) lead to estimators with extremely high outlier stability. In a given situation, therefore, it is imperative that one is able to choose the most suitable tuning parameters for that particular case. Here we consider a data driven algorithm for selecting the “optimal” tuning parameters \( (\beta, \gamma) \) which would provide best compromise for the given situation. For this purpose we modify an approach of Warwick (2002), pp. 78-82 and minimize an empirical version of the asymptotic summed mean square error. The optimization technique is a two stage process. Suppose that the data are generated by a contaminated version of a model distribution and let \( \theta_0 \) be the parameter for the model component. Although the data are generated by a contaminated version, the parameter \( \theta_0 \) of the model component is our target parameter. The spirit of such a set up is described in Warwick and Jones (2005). Let \( \hat{\theta}_{\beta, \gamma} = T_{\beta, \gamma}(G) \) be the corresponding minimum distance functional and \( \hat{\theta}_{\beta, \gamma} \) is the solution of the unbiased equation of L\( \beta \)DPD with tuning parameter \( (\beta, \gamma) \) based on the data. The summed mean square error of the minimum L\( \beta \)DPD estimator has the asymptotic formula:

\[
E\left[ (\hat{\theta}_{\beta, \gamma} - \theta)^T (\hat{\theta}_{\beta, \gamma} - \theta) \right] = (\hat{\theta}_{\beta, \gamma} - \theta)^T (\hat{\theta}_{\beta, \gamma} - \theta) + n^{-1} \text{tr} \left\{ \text{var} \left( \hat{\theta}_{\beta, \gamma} \right) \right\}.
\]  

(17)

Here \( \theta^* \) is the pilot estimator playing the role of \( \theta_0 \) and \( \text{tr} \{ \cdot \} \) represents the trace of matrix. The asymptotic covariance matrix of \( \sqrt{n}(\hat{\theta}_{\beta, \gamma} - \theta_{\beta, \gamma}) \) is \( J^{-1}KJ^{-1} \), where \( J \) and \( K \) are as in Equation 12 with \( \phi(x, g) = \frac{1}{\gamma} \log \left( 1 + \frac{y}{x} \right) \).

So the estimated asymptotic summed mean square of the ML\( \beta \)DPD is:

\[
(\hat{\theta}_{\beta, \gamma} - \theta)^T (\hat{\theta}_{\beta, \gamma} - \theta) + \frac{1}{n} J^{-1}KJ^{-1}.
\]

(18)

For the multiparameter case, the above quantity is a matrix. So trace of the matrix is used to provide a global measure of the summed mean square error for minimization. Thus when there are two parameters to be estimated (say \( (\mu, \sigma) \) for \( N(\mu, \sigma) \) model) then the expression to be minimized is:

\[
n^{-1} \text{tr} \left\{ J^{-1}(\hat{\theta}_{\beta, \gamma})K(\hat{\theta}_{\beta, \gamma})J^{-1}(\hat{\theta}_{\beta, \gamma}) \right\} + (\mu - \mu^*)^2 + (\sigma - \sigma^*)^2.
\]

(19)

The optimal value of \( (\beta, \gamma) \) is the minimizer of Equation 19 under certain conditions. One important note is that in the first stage of minimization our pilot estimate for \( \hat{\theta}^* \) is taken to be a good robust estimate based on the data as suggested in Warwick (2002). The empirical summed mean square error is then obtained by evaluating the expressions in Equation 18 or Equation 19 after substituting \( \hat{\theta}_{\beta, \gamma} \) for \( \theta_{\beta, \gamma} \) and the empirical distribution \( G_n \) in place of the true unknown distribution \( G \). Let us denote this empirical summed mean square error by AMSE in the following.

Algorithm

Given a dataset \( X_{n \times 1} \) we perform the following steps to obtain the estimate of \( \theta^* \):

1. Apply the method suggested in Warwick (2002) to get an optimal \( \alpha \) for MDPDE. Suppose this value is \( \alpha_0 \). This step is the 1st stage of optimization by assuming an initial pilot estimate of \( \theta^* \).
2. Consider the interval \((0, \alpha_n)\). Update the pilot estimate for \(\theta^* = \hat{\theta}_{\alpha_n}\), which is MDPDE of \(\theta\) with \(\alpha_n\) as the tuning parameter.

3. Perform a two dimensional optimization which selects the value of \((\beta, \gamma)\) for which the minimum:

\[
\min_{\beta \in (0,1), \gamma \in (0,1)} \left[ \min_{\alpha_n \in (0,1)} \text{AMSE}(\hat{\theta}_{\beta, \gamma}) \right]
\] (20)

is attained under the constraint \(\text{AMSE}(\hat{\theta}_{\beta, \gamma}) < \text{AMSE}(\hat{\theta}_{\alpha_n})\).

An alternative to this approach could be to perform an unrestricted minimization of \(\text{AMSE}(\hat{\theta}_{\beta, \gamma})\) with respect to \((\beta, \gamma)\) over the set \((0,1) \times (0,1)\).

**Real Data Examples**

Here we take some real data sets and use our algorithm to find the optimal tuning parameters to be used in estimating the parameters of the model. We worked with two data sets, Newcomb’s light speed data and Short’s parallax of the sun data, under normality assumptions. We have used the minimum L2 distance estimates as our pilot estimates of \((\mu, \sigma)\).

**Newcomb’s Data (Speed of Light)**

This example involves Newcomb’s light speed data (Stigler, 1977, Table 5). The data size is \(n = 66\). Under the normal model, the MLE of the mean and standard deviation for these data are found to be equal to 26.212 and 10.664, respectively. We employ our algorithm for tuning parameter selection and Table 4 reports the optimal tuning parameters for DPD and L\(\phi\)DPD, as well as the parameter estimates at these optimal values. The estimators are extremely close, but the estimated asymptotic summed mean square, for whatever it is worth, is lower in case of the ML\(\phi\)DE.

**Short’s Data (Parallax of the Sun)**

This example involves Short’s data for the determination of the parallax of the sun, the angle subtended by the earth’s radius as if viewed and measured from the surface of the sun. From this angle and available knowledge of the physical dimensions of the earth, the mean distance from earth to the sun can be easily determined. The raw observations are presented in Table 4 of Stigler (1977). The data size is \(n = 53\). Under the normal model, the MLE of the mean and standard deviation for these data are found to be equal to 8.378 and 0.846 respectively. We perform all the steps of the aforesaid tuning parameter selection algorithm and the results of the analysis are now listed in Table 5. Again, the empirical asymptotic MSE for the ML\(\phi\)DE is slightly better than that of the MDPDE.

From Fig. 3 and 4, it is evident that the normal fits coming from the MDPDE and ML\(\phi\)DE are in the same ballpark. However, if the empirical asymptotic summed mean square error is accepted as a reasonable criterion for discrimination, then the performance of the ML\(\phi\)DE is better than that of the MDPDE, although the order of improvement is small.

Fig. 3: Normal density fits for Newcomb’s data
The MLφDE for Independent Nonhomogeneous Observations

Here we generalize the above concept to the case of independent but not identically distributed observations. Ghosh and Basu (2013) explains the methodology for this problem in the case of DPD, but here we will extend it to the case of LφDPD.

Let us assume that the observed data $Y_1, ..., Y_n$ are independent but for each $i$, $Y_i \sim g_i$ where the densities $g_1, ..., g_n$ may not be same. We want to model $g_i$ by the family $\mathcal{G}_{i,\theta} \equiv \{f(\cdot; \theta) | \theta \in \Theta\}$ for all $i = 1, 2, ..., n$. We want to estimate $\theta$ by minimizing the LφDPD between the data and the model. However, the model density may not be same for each $Y_i$’s and hence we need to calculate the divergence between data and model separately for each data point. For this purpose, we minimize the average divergence between the data points and the models.

Therefore, we minimize:

$$
\frac{1}{n} \sum_{i=1}^{n} d(\hat{g}_i, f_i(\cdot; \theta))
$$

with respect to $\theta \in \Theta$, where $d(\hat{g}_i, f_i(\cdot; \theta))$ denotes the LφDPD between the density estimate corresponding to the $i$-th data point and the associated model density. In the presence of only one data point $Y_i$ from density $g_i$, the best possible density estimate of $g_i$ is the (degenerate) density which puts the entire mass on $Y_i$ so that we have:

$$
d(\hat{g}_i, f_i(\cdot; \theta)) = \frac{1}{n} \int \int f_i(\cdot; \theta) s^\theta \log \left(1 + \frac{Y}{s}\right) ds dt dy
$$

$$
= \frac{1}{n} \int_0^\infty \int_0^1 s^\theta \log \left(1 + \frac{Y}{s}\right) ds dt dy
$$

$$
= -\frac{1}{n} \int_0^\infty \int_{-\infty}^\infty s^\theta \log \left(1 + \frac{Y}{s}\right) ds dy ds dt dy + K.
$$

where, $K$ is a constant independent of $\theta$, the parameter of interest. Thus, for the purpose of estimation it suffices to minimize the objective function:

$$
H_\phi(\theta) = \frac{1}{n} \sum_{i=1}^{n} V_i(Y_i; \theta),
$$

(21)

where:
We also define, for each $i = 1, 2, \ldots$ the $p \times p$ matrix $J^0$ whose $(k, l)$-th entry is given by:

$$J^0_{kl} = E_x \left[ \nabla \Psi \left( Y_i ; \theta \right) \right].$$

where, $\nabla kl$ represents the partial derivative with respect to the indicated components of $\theta$. We further define the quantities:

$$\Psi_n = \frac{1}{n} \sum_{i=1}^{n} J^0_{i},$$

$$\Omega_n = \frac{1}{n} \sum_{i=1}^{n} \text{Var}_n \left[ \nabla H \left( Y_i ; \theta \right) \right].$$

A simple calculation shows that:

$$J^0 = \frac{1}{\gamma} \int f_i \left( \gamma ; \theta \right) \left( \gamma ; \theta \right) \left[ \frac{1}{f_i \left( \gamma ; \theta \right)} - u_i \left( \gamma ; \theta \right) \left( \frac{\gamma}{\gamma + 1} \right) \right] dy$$

and:

$$\Omega_n = \frac{1}{n} \sum_{i=1}^{n} \left\{ \int f_i \left( \gamma ; \theta \right) \left[ \frac{1}{f_i \left( \gamma ; \theta \right)} \left( \frac{\gamma}{\gamma + 1} \right) \right] \left( \gamma ; \theta \right) dy \right\}$$

where:

$$\xi_i = \frac{1}{\gamma} \int f_i \left( \gamma ; \theta \right) \left[ \frac{1}{f_i \left( \gamma ; \theta \right)} \left( \frac{\gamma}{\gamma + 1} \right) \right] g_i \left( \gamma ; \theta \right) dy.$$

We will make the following assumptions to establish the asymptotic properties of the ML DE:

- (G1) The support $\mathcal{S} = \left\{ y \mid f_i \left( y ; \theta \right) > 0 \right\}$ is independent of $i$ and $\theta$ for all $i$; the true distributions $G_i$ are also supported on $\mathcal{S}$ for all $i$. 



(G2) There is an open subset \( \Theta \) of the parameter space \( \Theta \), containing the best fitting parameter \( \theta^* \) such that for almost all \( y \in \mathcal{X} \) and all \( \theta \in \Theta \), all \( i = 1, 2, \ldots \), the density \( f_i(y;\theta) \) is thrice differentiable with respect to \( \theta \) and the third partial derivatives are continuous with respect to \( \theta \).

(G3) For each \( i = 1, 2, \ldots \), the three integrals
\[
\int f_i(y;\theta) \log \left( 1 + \frac{Z}{s} \right) ds \, dy,
\]
\[
\int f_i(y;\theta) \log \left( 1 + \frac{Z}{s} \right) ds \, dt \, dy
\]
and
\[
\int f_i(y;\theta) \log \left( 1 + \frac{Z}{s} \right) ds \, \left( g_i(y) \right) dy
\]
can be differentiated thrice with respect to \( \theta \) and the derivatives can be taken under the integral sign (the first indefinite integral).

(G4) For each \( i = 1, 2, \ldots \), the matrices \( J^{(i)} \) are positive definite and:
\[
\lambda_n = \inf_{n} \left[ \min \text{eigenvalue of } \Psi_n \right] > 0.
\]

(G5) There exists functions \( M^{(i)}(Y) \) such that:
\[
| \nabla \beta V_i(Y;\theta) | \leq M^{(i)}(Y) \forall \theta \in \Theta, \forall i \|
\]
with \( E_{\Theta | Y} \left[ M^{(i)}(Y) \right] < \infty \forall j, k, l \).

(G6) For all \( j, k \), we have:
\[
\lim_{N \to \infty} \sup_{\Theta} \left\{ \int E_{\Theta} \left[ \log \left( 1 + \frac{Z}{s} \right) \right] ds \, dt \, dy \right\} = 0, \tag{31}
\]
\[
\lim_{N \to \infty} \sup_{\Theta} \left\{ \int E_{\Theta} \left[ \log \left( 1 + \frac{Z}{s} \right) \right] ds \, \left( g_i(y) \right) dy \right\} = 0 \tag{32}
\]
Here \( I(\cdot) \) stands for indicator function.

(G7) For all \( \varepsilon > 0 \), we have:
\[
\lim_{N \to \infty} \sup_{\Theta} \left\{ \int E_{\Theta} \left[ \log \left( 1 + \frac{Z}{s} \right) \right] ds \, dt \, dy \right\} = 0 \tag{33}
\]

**Theorem 4**

Under assumptions (G1)-(G7), the following results hold:

\[ \text{(i) There exists a consistent sequence } \theta_n \text{ of roots to the minimum } \varphi \text{DPD estimator Equation 23.} \]

(ii) The asymptotic distribution of
\[
\Omega_n^2 \Psi_n \left[ \sqrt{n}(\hat{\theta}_n - \theta^*) \right] \text{ is } p\text{-dimensional normal with } (vector) \text{ mean 0 and covariance matrix } I_p, \text{ the } p\text{-dimensional identity matrix.}
\]

Note that, putting \( f_i = f \) for all \( i \), we get back the corresponding asymptotic properties of the minimum \( L_0 \)DPD estimator for the i.i.d. case. If \( f_i = f_i, i = 1, 2, \ldots \), we get \( J^{(i)} = J \) for all \( i \); thus \( \Psi_n = J \) and \( \Omega_n = K \). Here \( J \) and \( K \) are as defined previously. In this case assumptions (G1)-(G5) are exactly the same as the assumptions (A1)-(A5), while assumptions (G6) and (G7) are automatically satisfied by the dominated convergence theorem. Thus the result, which establishes the consistency and asymptotic normality of the minimum \( L_0 \)DPD estimator \( \hat{\theta} \) with \( n^{1/2} \left( \hat{\theta} - \theta^* \right) \)

Note that, putting \( f_i = f \) for all \( i \), we get back the corresponding asymptotic properties of the minimum \( L_0 \)DPD estimator for the i.i.d. case. If \( f_i = f_i, i = 1, 2, \ldots \), we get \( J^{(i)} = J \) for all \( i \); thus \( \Psi_n = J \) and \( \Omega_n = K \). Here \( J \) and \( K \) are as defined previously. In this case assumptions (G1)-(G5) are exactly the same as the assumptions (A1)-(A5), while assumptions (G6) and (G7) are automatically satisfied by the dominated convergence theorem. Thus the result, which establishes the consistency and asymptotic normality of the minimum \( L_0 \)DPD estimator \( \hat{\theta} \) with \( n^{1/2} \left( \hat{\theta} - \theta^* \right) \)

having the asymptotic covariance matrix \( \Psi_n^2 \Omega_n \Psi_n^{-1} = J^{-1} K J^{-1} \), emerges as a special case of Theorem 4.

**Normal Linear Regression**

A natural situation where the theory proposed above would be immediately applicable is the case of linear regression. We consider the linear regression model:
\[
y_i = x_i^T \beta + \varepsilon_i, \quad i = 1, \ldots, n, \tag{34}
\]
where the error \( \varepsilon_i \)'s are i.i.d. normal variables with mean zero and variance \( \sigma^2 \), \( x_i = (x_{i1}, \ldots, x_{ip}) \) is the vector of the independent variables corresponding to the \( i \)-th observation and \( \beta = (\beta_1, \ldots, \beta_p)^T \) represents the regression coefficients. We will assume that \( x_i \)'s are fixed. Then \( y_i \)'s are fixed. Then \( y_i \)'s satisfy our independent but not identically distributed. Thus \( y_i \)'s satisfy our independent but non-homogenous set-up and hence the ML\( \varphi \)DE of the parameter \( \theta = (\beta, \sigma^2)^T \) can be obtained by minimizing the expression in Equation 21 with \( f_i = N(x_i^T \beta, \sigma^2) \).

**Real Data Examples in Regression**

We now consider some real data examples to illustrate the above technique in linear regression.

**Hertzsprung-Russel Data**

This example involves a robust regression on the Hertzsprung-Russel data. These data, associated with the Hertzsprung-Russel diagram of the star cluster CYG OB1 containing 47 stars in the direction of Cygnus has been analyzed previously by several authors including Roussev and Leroy (1987).
We fit the simple linear regression model $y = \eta_0 + \eta_1 x + \varepsilon$ under homoscedastic normal errors. Here the independent variable ($x$) is the logarithm of the temperature of the stars and the dependent variable ($y$) is the logarithm of the light intensity of the stars. The initial regression parameter values are the Least Median of Squares (LMS) estimates. The initial scale estimate is the scaled Median Absolute Deviation (MAD) of the LMS residuals. We perform the previously mentioned steps of optimal tuning parameter selection and obtain the estimates for the regression coefficients, which are given in Table 6. The regression lines for LS regression, LMS regression and minimum $\text{L}_\phi \text{DPD}$ regression are given in the Fig. 5. The robust performance of the $\text{ML}_\phi \text{DE}$ is self evident.

Salinity Data

This example involves the Salinity data (Table 5, Chapter 3, Rousseeuw and Leroy, 1987). These data were originally presented by Ruppert and Carroll (1980). The measurements of the salt concentration of the water and the river discharge taken in North Carolina’s Pamlico Sound were recorded as the data. These data represent a multiple linear model with salinity as the dependent variable ($y$) and salinity lagged by two weeks ($x_1$), the number of biweekly periods elapsed since the beginning of the spring season ($x_2$) and the volume of river discharge into the sound ($x_3$) as the dependent variable.

We fit the multiple linear regression model $y = \eta_0 + \eta_1 x_1 + \eta_2 x_2 + \eta_3 x_3 + \varepsilon$ under homoscedastic normal errors. The initial regression parameter values are the Least Median of Squares (LMS) estimates. The initial scale estimate is the scaled Median Absolute Deviation (MAD) of the LMS residuals.

The optimal parameters obtained through our algorithm for optimal parameter selection are presented in Table 7. The residual plots for LS regression, LMS regression and minimum $\text{L}_\phi \text{DPD}$ regression are given in the Fig. 6. Like the LMS method (and unlike the LS method) the $\text{ML}_\phi \text{DE}$ gives a nice outlier resistant fit.

Table 6: Regression estimates for Hertzsprung-Russel data

| Category | $\text{ML}_\phi \text{DE}$ |
|----------|-----------------------------|
| Tuning Parameter | $(\beta, \gamma) = (1,0.9)$ |
| Estimate of $\eta_0$ | $-8.5557324$ |
| Estimate of $\eta_1$ | $3.0590795$ |
| Estimate of $\eta_2$ | $0.4266284$ |

Table 7: Regression estimates for Salinity data

| Category | $\text{ML}_\phi \text{DE}$ |
|----------|-----------------------------|
| Tuning Parameter | $(\beta, \gamma) = (1.0.9)$ |
| Estimate of $\eta_0$ | $57.16780461$ |
| Estimate of $\eta_1$ | $0.06010002$ |
| Estimate of $\eta_2$ | $-0.01301208$ |
| Estimate of $\eta_3$ | $-2.08372562$ |
| Estimate of $\sigma$ | $0.56157558$ |

Fig. 5: Regression fits for the Hertzsprung-Russel data
Hypothesis Testing using LϕDPD

Now we develop the tests of parametric hypothesis based on LϕDPD divergence. The most common problem is that of testing a simple null hypothesis for a parametric family of densities \( \{ f_\theta : \Theta \subseteq \mathbb{R}^p \} \) under the one sample case. Here we test:

\[
H_0 : \theta = \theta_0 \quad \text{Versus} \quad H_1 : \theta \neq \theta_0
\]  

(35)

when a random sample \( X_1, X_2, \ldots, X_n \) is available from the population of interest. We propose our test statistic as:

\[
T = d_{\hat{\theta}, \theta_0}(\hat{\theta}, \theta_0) = 2nd_{\hat{\theta}, \theta_0}(f_\theta, f_{\theta_0})
\]

where:

\[
d_{\hat{\theta}, \theta_0}(f_\theta, f_{\theta_0}) = \int \left[ B(f_\theta(x)) - B(f_{\theta_0}(x)) ight] dx
\]

\[
- \left( f_\theta(x) - f_{\theta_0}(x) \right) B'(f_{\theta_0}(x)) dx.
\]

(36)

with \( \hat{\theta} = \hat{\theta}_{\hat{\theta}} \) being the MLϕDE estimate of \( \theta \) and \( B(\cdot) \) is as defined in Equation 16. We shall find the asymptotic distribution of \( T \) under \( H_0 \) and reject the null hypothesis for large values of \( T \).

We assume the following regularity conditions of the parametric family of distributions:

- (B1) The support of the distribution function \( F_\theta \), i.e., the set \( \mathcal{X} = \{ x : f_\theta(x) > 0 \} \) is independent of \( \theta \)
- (B2) There is an open subset \( \omega \) of the parameter space \( \Theta \), containing the true parameter value \( \theta_0 \) such that for almost all \( x \in \mathcal{X} \) and all \( \theta \in \omega \) the density \( f_\theta(x) \) is three times differentiable with respect to \( \theta \) and the third partial derivatives are continuous with respect to \( \theta \)
- (B3) The integrals \( \int B'(f_\theta(x)) f_{\theta_0}(x) dx \) can be differentiated with respect to \( \theta \) and the derivatives can be taken under the integral sign
- (B4) The \( p \times p \) matrix \( J(\theta) \) defined by:

\[
J_{jkl}(\theta) = E_\theta \left[ \nabla_x \left( \int B'(f_\theta(x)) f_{\theta_0}(x) - B(f_\theta(x)) \right) dx - B'(f_\theta(x)) \right]
\]

is positive definite where \( E_\theta \) represents the expectation under the density \( f_\theta \)
- (B5) There exists functions \( M_{jkl}(x) \) with finite expectation, \( j,k,l = 1, \ldots, p \), such that:

\[
\left| \nabla_x \left( \int B'(f_\theta(x)) f_{\theta_0}(x) - B(f_\theta(x)) \right) dx - B'(f_\theta(x)) \right| \leq M_{jkl}(X) ; \forall \theta \in \omega.
\]
Then we have the following theorem.

**Theorem 5**

Under the assumptions (B1)-(B5) and under the null hypothesis \( H_0: \theta = \theta_0 \), the asymptotic distribution of \( T_{\beta,\gamma} (\hat{\theta}_1, \hat{\theta}_2) \) coincides with the distribution of:

\[
\sum_{i=1}^{r} \lambda_i Z_i^2,
\]

where, \( Z_i \) s are independent standard normals and \( \lambda_i \)'s are non-zero eigenvalues of \( A(\theta_0)\Sigma(\theta_0) \) and:

\[
r = \text{rank}(\Sigma(\theta_0)) = \text{rank}(\Sigma(\theta_0)A(\theta_0)\Sigma(\theta_0)) = \text{rank}(\Sigma(\theta_0))
\]

where, \( \Sigma(\theta_0) \) is the asymptotic covariance matrix of \( \sqrt{n} \hat{\theta}_{\beta,\gamma} \) under the null hypothesis and \( V_{\theta}^2 \) represents second derivative with respect to \( \theta \).

We can extend this theorem and hence the testing result to the general two sample problem of testing \( H_0: \theta_1 = \theta_2 \) against \( H_1: \theta_1 \neq \theta_2 \) where there is a random sample of size \( n \) from population 1 with parameter \( \theta_1 \) and that of size \( m \) from population 2 with parameter \( \theta_2 \). Let \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \) be ML\theta\DEs of the parameter in populations 1 and 2, respectively. Then under the (B1)-(B5) regularity conditions on the model, we have the following results.

**Theorem 6**

Under the null \( H_0: \theta_1 = \theta_2 \), the asymptotic distribution of:

\[
S = S_{\beta,\gamma} (\hat{\theta}_1, \hat{\theta}_2) = \frac{2mn}{m+n} d_{\beta,\gamma} (f_{\theta_1}, f_{\theta_2})
\]

coincides with that of:

\[
\sum_{i=1}^{r} \lambda_i Z_i^2
\]

where, \( Z_i \) s are independent standard normals and \( \lambda_i \)'s are non-zero eigenvalues of \( A(\theta_0)\Sigma(\theta_0) \) and \( r = \text{rank}(\Sigma(\theta_0)A(\theta_0)\Sigma(\theta_0)) \) where \( A(\theta) \) and \( \Sigma(\theta) \) are defined in the statement of Theorem 5.

**Equivalence with the Score Test**

A score test, developed in the same spirit under the same set up as in Theorem 5, also has the same asymptotic null distribution.

**Theorem 7**

The score test statistic using the LB\DPD for testing the simple null in Equation 35 can be given by:

\[
nU^T (\hat{\theta}_1) J^T (\hat{\theta}_1) A(\theta) J^T (\theta) \overline{U} (\theta_1)
\]

where:

\[
U_n (x) = u_n (x) B^* (f_{\theta_1} (x)) f_{\theta_1} (x) - \int u_n (x) B^* (f_{\theta_1} (x)) f_{\theta_1}^2 (x) dx
\]

and:

\[
\overline{U} (\theta) = \frac{1}{n} \sum_{i=1}^{n} U_n (X_i)
\]

with

\[
B^* (x) = \frac{x^2}{2 \gamma} \log \left( 1 + \frac{\gamma}{x} \right)
\]

\[
J_n (\theta_1) = -E_{\theta_1} \frac{\partial}{\partial \theta} U_n (X_i)|_{\theta=\theta_1}
\]

and \( A(\theta_0) \) is as described in Theorem 5. Under the null hypothesis, the asymptotic distribution of this statistic is same as that of \( T_{\beta,\gamma} (\hat{\theta}_1, \hat{\theta}_2) \).

**Divergence Difference Test Statistic**

We assume that we have a parametric model \( \mathcal{F} \) of densities and \( X_1, \ldots, X_n \) be i.i.d. from the true distribution \( G \) with the same support as the distributions in \( \mathcal{F} \). Consider the null hypothesis:

\[
H_0: \theta \in \Theta_0 \text{ versus } H_1: \theta \in \Theta \setminus \Theta_0,
\]

where, \( \Theta_0 \) is a proper subset of \( \Theta \). The Likelihood Ratio Test (LRT) is one of the most common tests that may be employed in this situation. Define:

\[
\lambda = \sup_{\theta \in \Theta_0} \mathcal{D} (\theta | X_1, \ldots, X_n)
\]

\[
\bar{\lambda} = \sup_{\theta \in \Theta} \mathcal{D} (\theta | X_1, \ldots, X_n)
\]

where, \( \mathcal{D} (\theta | X_1, \ldots, X_n) \) is the likelihood of \( q \) given the data. The test statistic in this case is \(-2 \log \lambda\). Assume that the distribution function \( G \) is discrete. In particular let its support be \( \mathcal{X} = \{0, 1, 2, \ldots\} \), which is also the common support of the family \( \mathcal{F} \). Then the test statistic can be expressed in terms of observed relative frequencies \( v_n \):
where, LD (·, ·) stands for the likelihood disparity. Here \( \hat{\theta}_0 \) and \( \hat{\theta} \) stands for unrestricted maximum likelihood estimator and maximum likelihood estimator under null hypothesis respectively. Equation 38 gives a motivation to construct a new test statistic based on \( L(\Phi) \).

As an analog of the likelihood ratio test, we consider the Divergence Difference Test (DDT) based on \( L(\Phi) \) to test the hypothesis given in Equation 37. Note that the test statistic in Equation 38 can be viewed as a difference of the minimized value of likelihood disparity under null and unrestricted minimum of likelihood disparity. In the same spirit one may define the following test statistic:

\[
DDT_{\beta, \gamma}(v_n) = 2n \left[ d_{\beta, \gamma}(v_n, f_{\hat{\theta}}) - d_{\beta, \gamma}(v_n, f_\theta) \right].
\]

(39)

\( \hat{\theta}_0 \) and \( \hat{\theta} \) are ML\( \Phi \)DE under null hypothesis and unrestricted minimum ML\( \Phi \)DE respectively. Also note that:

\[
d_{\beta, \gamma}(v_n, f_\theta) = \sum_{i=1}^{n} [B(v_n(x_i)) - B(f_\theta(x_i)) - (v_n(x_i) - f_\theta(x_i))B'(f_\theta(x_i))].
\]

where, \( B(\cdot) \) is defined as Equation 16. We will show that under certain regularity conditions the asymptotic distribution of the test statistic \( DDT_{\beta, \gamma}(v_n) \) coincides with the distribution of linear combination of independent chi-squared random variables. Suppose that \( \Theta_0 \) is defined by a set of \( r \leq p \) restrictions on \( \Theta \) defined by \( R(\theta) = 0 \). We assume that the parameter space under \( H_0 \) can be described by a parameter \( \xi = (\xi_1, \ldots, \xi_m) \), with \( p-r \) independent components, i.e., \( H_0 \) specifies that there exists a function \( b: \mathbb{R}^{r' \prime} \rightarrow \mathbb{R}^r \) where \( \theta = b(\gamma), \gamma \in \Gamma \subseteq \mathbb{R}^{r' \prime} \). The function \( b(\cdot) \) is assumed to have continuous derivative \( b'(\xi) \) of order \( p \times (p-r) \) with rank \( p-r \). Then the constrained estimator is \( \hat{\theta}_0 = b(\hat{\xi}) \), where \( \hat{\xi} \) is the ML\( \Phi \)DE under the \( \xi \) formulation of the model. Let \( G = F_{\theta_0} \) be the true distribution which belongs to the family \( \mathcal{F} \) with parameter \( \theta \). Under \( H_0 \), let \( \xi \) be the true value of the reduced parameter. So we have \( \theta = b(\xi) \). When the null hypothesis is true under standard regularity conditions it can be easily shown that \( \hat{\xi} \) and \( \hat{\theta}_0 \) are consistent for \( \xi \) and \( \theta \) respectively in the sense that:

\[
\hat{\xi} = \xi + n^{-1/2} \left[ b'(\hat{\xi}) \hat{J}_b(\hat{\xi}) b(\hat{\xi}) \right]^{-1} b'(\hat{\xi}) Z_b(\hat{\xi}) + o_p(n^{-1/2}).
\]

(40)

where, \( Z_b(\hat{\xi}) \) is AN(0, \( K_b(b(\xi)) \)). Here \( J(\cdot) \) and \( K(\cdot) \) is defined as in Theorem 1. Now we will lay out some appropriate regularity conditions under which we will derive the asymptotic distribution of \( DDT_{\beta, \gamma}(v_n) \) under the null hypothesis:

- (C1) The assumptions (A1)-(A5) hold under the model conditions
- (C2) The unconstrained minimum L\( \Phi \)DE estimator \( \hat{\theta} \) satisfies:

\[
\hat{\theta} = \theta + n^{-1/2} J^{-1}_b(\theta) Z_b(\theta) + o_p(n^{-1/2}).
\]

(41)

where, \( Z_b(\theta) \) is AN(0, \( K_b(b(\xi)) \)).
- (C3) The null hypothesis \( H_0 \) is either simple and \( \Theta_0 = \{\theta_0\} \), where \( \theta_0 \) is in the interior of \( \Theta \), or \( H_0 \) is composite and \( \Theta_0 = \{b(\xi): \xi \in \Gamma \subseteq \mathbb{R}^{r' \prime} \} \)
- (C4) If \( H_0 \) is composite then the constrained estimator \( \hat{\theta}_0 = b(\hat{\xi}) \) and \( \hat{\xi} \) satisfies Equation 40. Define:

\[
\Sigma_{\beta, \gamma}(\theta, \xi) = J^{\prime\prime}_{\beta, \gamma}(\theta, \xi) K_b(\theta) J_{\beta, \gamma}^{-1}(\theta, \xi),
\]

where

\[
J_{\beta, \gamma}(\theta, \xi) = \left[ J_b(\theta)^{-1} - b'(\xi)[b(\xi)^T J_b(\theta) b(\xi)]^{-1} b(\xi)^T \right]^{-1}
\]

Theorem 8

Suppose that assumption (C1)-(C4) hold. Under \( f_{\theta_0} \), \( \theta_0 \in \Theta_0 \), the limiting distribution of the distance difference test statistic in Equation 39 coincides with the distribution of:

\[
\sum_{j=1}^{m} \lambda_j Z_{j},
\]

where, \( \lambda_j \)'s are non-zero eigenvalues of \( A(q_0) \Sigma_{\beta, \gamma}(\theta_0, \xi) \) and \( m = \text{rank}(A(q_0) \Sigma_{\beta, \gamma}(\theta_0, \xi)) \). Moreover if \( \Theta_0 = \{\theta_0\} \) then under the null hypothesis the asymptotic distribution of distance difference test statistic in Equation 39 is same as that of \( T_{\beta, \gamma}(\hat{\theta}_0, \hat{\theta}_0) \) in Theorem 5.

Remark

In the above theorems the null distribution of the test statistic turns out to be same as that of a linear
combination of independent chi squared random variables. In general it is hard to get hold of critical values under this distribution for actually performing the test. Also calculations regarding this distribution become numerically hard. This gives the motivation to explore another test statistic which will lead to a simpler null distribution.

#### Wald Type Test

Assume a similar setup of hypothesis testing as in Equation 37. Suppose that the null space $\Theta_0 \subseteq \Theta \subseteq \mathbb{R}^r$ is defined by a set of $r \leq p$ restrictions on $\Theta$ defined by $R_d(\theta) = 0$, $1 \leq i \leq r$. Let $G = F_0$ be the true distribution which belongs to the family $\mathcal{F}$ with parameter $\theta$. Assume $\hat{\theta}$ to be the ML estimate of the true parameter $\theta$. Define $R(\theta) = (R_1(\theta), \ldots, R_r(\hat{\theta}))^T$ and $D(\theta) = \left[ \frac{\partial R(\theta)}{\partial \theta} \right]_{r \times p}$. Under the spirit of the original Wald test statistic, we can construct the following test statistic:

$$W(\hat{\theta}) = R(\hat{\theta})^T D(\hat{\theta}) \Sigma(\hat{\theta})^{-1} D(\hat{\theta})^T R(\hat{\theta}),$$

where, $\Sigma(\theta) = J_\theta(\theta)^{-1} K_{\theta}(\theta) J_\theta(\theta)^{-1}$ under the $B(\cdot)$ function described in Equation 16. Under standard regularity conditions it is easy to prove that the asymptotic distribution of $W(\hat{\theta})$ is $\chi^2$ under the null hypothesis. The proof follows from simple application of delta method theorem on the quantity $R(\hat{\theta})$ and the fact that under the null hypothesis $\sqrt{n} (\hat{\theta} - \theta)$ is $\mathcal{N}(0, \Sigma(\theta))$. The main benefit of this test statistic is that its asymptotic null distribution is simpler. Hence it is easy to perform numerical computations based on these statistics. For example, the critical values of the test statistic can be computed with ease in this case.

#### Real Data Example

Researchers needed to evaluate the effectiveness of an insecticide (dieldrin) in killing Anopheles farauti mosquitoes. The theory was that resistance to dieldrin was due to a single dominant gene and that in an appropriately selected sample of the mosquitoes, there should be $50\%$ susceptibility to insecticide. The hypothesis is:

$$H_0 : p = \frac{1}{2} \text{ versus } H_1 : p \neq \frac{1}{2},$$

where, $p$ is the probability of susceptibility. The results of such experiment is given in Osborn (1979). The sample contains 465 mosquitoes where 264 of them died on being exposed to the insecticide. We can perform this test with test statistic $\DDT_{\beta}(v_a)$ in Equation 39. Here $\beta$ and $\gamma$ are chosen to be 0.3 and 0.05 respectively. The support of the distribution is $x = (0, 1)$, where the digit 1 stands for the death of a mosquito. From here it is evident that $v_a(1) = 264/465$. The null hypothesis is rejected if the value of the test statistic is large. In this case the asymptotic null distribution of the test statistic turns out to be $0.774 \chi^2_{(1)}$. Under the observed data the value of the test statistic turns out to be approximately 6.62. The $95\%$ quantile of the aforementioned scaled chi-squared distribution is 2.97. So, under $5\%$ level of significance the null hypothesis is rejected.

#### Summary

In this paper, we have developed a large class of density based divergences which includes the density power divergence family as a special limiting case. The key philosophy of stronger downweighting effect to construct the new family has been discussed. For application purposes, the family gives the data analyst a larger number of choices of possible divergences for inference purposes. We have shown several asymptotic and distributional properties of the proposed estimator. We have also shown that judicial choice of the tuning parameters leads to highly robust and efficient estimators which can often dominate the MDPDE. Though one of the parameters has a smaller effect on the robustness we have shown that both of them play an important role in the context of finite sample efficiency. We have also presented a possible data driven algorithm to obtain the “optimal” estimator in a given data set. We have also considered several hypothesis testing strategies for parametric models which may serve as robust alternatives to the classical likelihood ratio and other likelihood based tests.

#### Remark

Like the MDPDE, the procedures described in this paper avoid the nonparametric density estimation and associated complications specific to classical minimum distance estimation. Another approach of this type can be found in Toma and Broniatowski (2011).

#### Remark

In creating the test statistics for parametric hypothesis testing using the $L_\theta$DPD, we have restricted ourselves to the case where the same set of tuning parameters have been used for estimation as well as the construction of the subsequent divergences. In practice, one could allow them to vary; see, for example, Basu et al. (2013). In the present context, while this is possible, we do not explore this issue as we feel that there are enough tuning parameters involved already and there are no demonstrated results indicating that such differential choices will necessarily produce improved tests.
Remark

In this paper, most of our illustrations have been with respect to the continuous model. Theoretically, however, there is nothing preventing its successful use in discrete models. All the necessary theories work out satisfactorily in this case.

Proof of Theorems

Proofs of Theorems 2, 5, 6 and 7 are skipped as they can be reproduced along the existing proofs in Basu et al. (2011), Ghosh and Basu (2013) and Ghosh et al. (2015).

Proof of Theorem 2:

Proof

(a) From (P4) we know that \( f_0(x) \phi(f_0(x), \gamma) \) is continuous for \( \gamma \in (0, 1) \) and \( \lim_{\gamma \to 0} f_0(x) \phi(f_0(x), \gamma) = 1 \). By applying dominated convergence theorem (DCT) on Equation 11 at \( \gamma \to 0^+ \) we get:

\[
\frac{1}{n} \sum_{i=1}^{n} u_\beta(X_i) \phi^{1+\beta}(X_i) - \frac{1}{n} \int u_\beta(x) \phi^{1+\beta}(x) \, dx = 0
\]

which is the unbiased estimating equation for DPD with tuning parameter \( \beta \). Hence the result follows.

(b) As \( u_\alpha(x) f_0(x)^{1+\beta} \), \( u_\alpha(x) u_\beta(x) f_0^{1+\beta} \), \( u_\alpha(x) u_\beta(x) f_0^{1+\beta} \) are integrable and \( f_0(x) \phi(f_0(x), \gamma) \) is bounded, by DCT on Equation 12 at \( \gamma \to 0^+ \) we get:

\[
J_\beta = \int u_\alpha u_\beta f_0^{1+\beta}, \quad K_\beta = \int u_\alpha u_\beta f_0^{1+\beta} - \frac{\gamma}{\sqrt{\beta}} \psi_\beta,
\]

\[
\zeta_\beta = \int u_\beta f_0^{1+\beta},
\]

i.e., \( \lim_{\gamma \to -1} J_\beta K_\beta^{-1} = J_\beta^* K_\beta J_\beta^* \). We already know from the assumptions that \( J_\beta^* K_\beta J_\beta^* \preceq J_\alpha^* K_\alpha J_\alpha^* \), i.e., \( J_\beta^* K_\beta J_\beta^* - J_\alpha^* K_\alpha J_\alpha^* \) is positive definite, where \( J_\alpha \) and \( K_\alpha \) are defined in the same fashion as \( J_\beta \) and \( K_\beta \) respectively. The inequality of the asymptotic variances is used here in the sense that AE of DPD with parameter \( \beta \) is greater than that of AE of DPD with parameter \( \alpha \). So there exists a \( \gamma = \gamma(\alpha, \beta) \) such that \( J_\alpha^* K_\alpha J_\alpha^* - J_\beta^* K_\beta J_\beta^* \). Hence the result follows.

Proof of Theorem 3

Proof

First let us assume that breakdown occurs at the model so that there exists sequence \( K_n \) of model densities such that \( \{0_n\} \) as \( n \to \infty \). Now, consider:

\[
D(h_{n.a}, f_{n.a}) = \int_{K_n} d(h_{n.a}, f_{n.a}) + \int_{K_n} d(h_{n.a}, f_{n.a})_{*},
\]

where, \( A_n = \{ x: g(x) > \max\{a(x), f_{n.a}(x)\} \} \). Now since \( g \) belongs to the model family \( \mathcal{M} \), from (BP1) it follows that \( \int_{K_n} k_n(x) \to 0 \) and from (BP2) we get \( \int_{K_n} f_{n.a} \to 0 \), thus under \( k_n \) and \( f_{n.a} \), the set \( A_n \) converges to a set of zero probability as \( n \to \infty \). Thus, on \( A_n \), \( d(h_{n.a}, f_{n.a}) \to d(1-\epsilon)g, 0) \) as \( n \to \infty \) and so by DCT

\[
\int_{K_n} d(h_{n.a}, f_{n.a}) \to \int_{K_n} d(1-\epsilon)g, 0) \to 0. \quad \text{Using (BP1), (BP2)}
\]

and the above result, we have \( \int_{K_n} d(h_{n.a}, f_{n.a}) \to M^{(1)}_{f(1-\epsilon)} \).

Next, by (BP1) and (BP2), \( \int_{K_n} g \to 0 \) as \( n \to \infty \), so under \( g \), the set \( A_n \) converges to a set of zero probability. Hence, similarly, we get

\[
\int_{K_n} d(h_{n.a}, f_{n.a}) \to \int_{K_n} d(\epsilon k_n, f_{n.a}) \to 0. \quad \text{Now by (BP3), we have}
\]

\[
\int d(\epsilon k_n, f_{n.a}) \geq \int d(\epsilon f_{n.a}, f_{n.a}) = M^{(2)}_{f(1-\epsilon)} \cdot \quad \text{Thus combining the equations we get}
\]

\[
\liminf_{n \to \infty} D(h_{n.a}, f_{n.a}) \geq M^{(1)}_{f(1-\epsilon)} - M^{(2)}_{f(1-\epsilon)} + M^{(1)}_{f(1-\epsilon)} = a_1(\epsilon), \quad \text{say.}
\]

We will have a contradiction to our breakdown assumption if we can show that there exists a constant value \( \theta \) in the parameter space such that for the same sequence \( k_n \):

\[
\limsup_{n \to \infty} D(h_{n.a}, f_{n.a}) < a_1(\epsilon)
\]

as then the sequence \( \{0_n\} \) above could not minimize \( D(h_{n.a}, f_{n.a}) \) for every \( n \). We will now show that above equation is true for all \( \epsilon < 1/2 \) under the model when we choose \( \theta = \theta_0 \). For any fixed \( \theta \), let \( B_n = \{ x: k_n(x) > \max\{g(x), f_0(x)\} \} \). Since \( g \) belongs to the model \( \mathcal{M} \), from (BP1) we get \( \int_{K_n} g \to 0 \), \( \int_{K_n} f_0 \to 0 \) and \( \int_{K_n} k_n \to 0 \) as \( n \to \infty \). Thus, under \( k_n \), the set \( B_n \) converges to a set of zero probability, while under \( g \) and \( f_0 \), the set \( B_n \) converges to a set of zero probability. Thus, on \( B_n \), \( d(h_{0n}, f_0) \to d(\epsilon k_n, 0) = B(\epsilon k_n) \) as \( n \to \infty \). So by DCT

\[
\int_{K_n} d(h_{n.a}, f_{n.a}) \to \int B(\epsilon k_n) \to 0. \quad \text{Similarly we have}
\]

\[
\int_{K_n} d(h_{n.a}, f_{n.a}) \to \int d(1-\epsilon)g, f_0 \to 0. \quad \text{Therefore, we have:}
\]

\[
\limsup_{n \to \infty} D(h_{n.a}, f_{n.a}) = \int D(1-\epsilon)g, f_0 \quad \text{limsup}_{n \to \infty} \int B(\epsilon k_n).
\]

Taking \( \theta = \theta_0 \) in Equation 43 and then using (BP3) we get:

\[
\limsup_{n \to \infty} D(h_{n.a}, f_{n.a}) \leq M^{(1)}_{f(1-\epsilon)} - M^{(2)}_{f(1-\epsilon)} + M^{(1)}_{f(1-\epsilon)} = a_1(\epsilon),
\]
say. Consequently, asymptotically there is no breakdown if for ε level contamination when \(a_1(\varepsilon) < a_1(\delta)\). But, notice \(a_1(\varepsilon)\) and \(a_1(\delta)\) are strictly decreasing and increasing functions respectively. To see this for \(a_1(\varepsilon)\), notice \(M_{\varepsilon \rightarrow}^{(1)} - M_{\delta \rightarrow}^{(2)}\). As \(\varepsilon \uparrow 1\) the above expression decreases. \(M_{\varepsilon \rightarrow}^{(1)} = \int B((1-\varepsilon)f)\). From Equation 16 we see that \(B(\cdot)\) is an increasing function on positive half line. Using this it is evident that \(M_{\varepsilon \rightarrow}^{(1)}\) decreases as \(\varepsilon \uparrow 1\).

So, \(a_1(\varepsilon)\) decreases as \(\varepsilon \uparrow 1\). Similarly, it can be shown \(a_\delta(\varepsilon)\) is an increasing function of \(\varepsilon\).

But \(a_\varepsilon(1/2) = a_\delta(1/2)\); thus asymptotically there is no breakdown and \(\limsup_{n \to \infty} |T_{\beta,\delta}H_{\varepsilon, \delta}| < \infty\) for \(\varepsilon < 1/2\). Hence the theorem follows.

**Proof of Theorem 7**

**Proof**

We know the estimating equation for general M-estimators as:

\[
\frac{1}{n} \sum_{i=1}^{n} u_\theta(X_i)B^*(f_\theta(X_i))f_\theta(X_i)
- \int u_\theta(x)B^*(f_\theta(x))f_\theta(x)dx = 0
\]

or equivalently:

\[
\frac{1}{n} \sum_{i=1}^{n} (u_\theta(X_i)B^*(f_\theta(X_i))f_\theta(X_i)
- \int u_\theta(x)B^*(f_\theta(x))f_\theta(x))dx = 0
\]

Viewing this as usual score equation, we take \(U_\theta(X_i)\):

\[
U_\theta(X_i) = u_\theta(X_i)B^*(f_\theta(X_i))f_\theta(X_i)
- \int u_\theta(x)B^*(f_\theta(x))f_\theta(x))dx.
\]

We have already seen that the statistic \(T_{\beta,\theta} = \hat{\beta}, \hat{\theta}\) satisfies:

\[
T_{\beta,\theta} = n(\hat{\beta}, \hat{\theta})^T A(\theta_0) (\hat{\beta} - \theta_0) + o_p(1).
\]

Note that \(\frac{1}{n} \sum_{i=1}^{n} U_\theta(X_i) = 0\) is solved for \(\theta = \hat{\theta}\). By Taylor series expansion:

\[
\sqrt{n} \sum_{i=1}^{n} U_\theta(X_i)
= \sqrt{n} \sum_{i=1}^{n} U_\theta(X_i) + \sqrt{n} (\hat{\beta} - \theta_0) \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} U_\theta(X_i) |_{\theta=\theta_0}
+ \frac{1}{2} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta \partial \theta} U_\theta(X_i) |_{\theta=\theta_0}
\]

for some \(\theta^*\) in between \(\theta_0\) and \(\hat{\theta}\). So we have:

\[
\sqrt{n} \sum_{i=1}^{n} U_\theta(X_i) = \sqrt{n} \tilde{U}(\theta_0)
+ \sqrt{n} (\hat{\beta} - \theta_0) \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} U_\theta(X_i) |_{\theta=\theta_0} + o_p(1).
\]

And hence:

\[
\sqrt{n} \tilde{U}(\theta_0) = \sqrt{n} (\hat{\beta} - \theta_0) \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} U_\theta(X_i) |_{\theta=\theta_0} + o_p(1).
\]

Note \(\frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} U_\theta(X_i) |_{\theta=\theta_0} \rightarrow E_{\theta_0} \frac{\partial}{\partial \theta} U_\theta(X_i) |_{\theta=\theta_0} = -J_{\theta}(\theta_0)\) as \(n \to \infty\). Hence:

\[
\sqrt{n} \tilde{U}(\theta_0) = \sqrt{n} (\hat{\beta} - \theta_0) J_{\theta}(\theta_0) + o_p(1)
\]

So:

\[
n(\hat{\beta} - \theta_0)^T A(\theta_0) (\hat{\beta} - \theta_0)
= n \tilde{U}(\theta_0) J_{\theta}(\theta_0) + o_p(1).
\]

This completes the proof.

**Proof of Theorem 8**

**Proof**

A Taylor expansion of Equation 39 around \(\hat{\theta}\) gives:

\[
DDT_{\beta,\theta}(v_\varepsilon)
= 2n \left[ d_{\beta,\theta}(v_\varepsilon, f_\theta) - d_{\beta,\theta}(v_\varepsilon, f_\delta) \right]
= 2n \left[ \sum_j (\hat{\theta}_0 - \hat{\theta}) \nabla_{\beta} \hat{d}_{\beta,\theta}(v_\varepsilon, f_\theta) |_{v_\varepsilon, f_\delta} \right] + o_p(1)
\]

where the subscripts indicate the denoted components of the vector. Also \(\theta^*\) lies in the line segment joining \(\hat{\theta}_0\) and \(\hat{\theta}\). By definition, \(\nabla_{\beta} \hat{d}_{\beta,\theta}(v_\varepsilon, f_\theta) |_{v_\varepsilon, f_\delta} = 0\). Hence, the Equation 44 reduces to:

\[
DDT_{\beta,\theta}(v_\varepsilon)
= n(\hat{\theta}_0 - \hat{\theta})^T \nabla_{\beta} \hat{d}_{\beta,\theta}(v_\varepsilon, f_\theta)
+ 2n \left[ \sum_j (\hat{\theta}_0 - \hat{\theta}) \nabla_{\beta} \hat{d}_{\beta,\theta}(v_\varepsilon, f_\theta) - \nabla_{\beta} \hat{d}_{\beta,\theta}(v_\varepsilon, f_\theta) |_{v_\varepsilon, f_\delta} \right] (\hat{\theta}_0 - \hat{\theta})
\]

We will show that under the null \(\nabla_{\beta} \hat{d}_{\beta,\theta}(v_\varepsilon, f_\theta) - \nabla_{\beta} \hat{d}_{\beta,\theta}(v_\varepsilon, f_\theta) |_{v_\varepsilon, f_\delta} = 0\) as \(n \to \infty\). By another Taylor
expansion around the true value $\theta_0$, we get for some $\theta^*$ between $\theta_0$ and $\theta^*$:

$$\nabla d_{p, \gamma}(v_*, f_*)$$

$$\nabla d_{p, \gamma}(v_*, f_*) + \sum (\theta^* - \theta_0) \nabla \mu_{d_{p, \gamma}}(v_*, f_*) \quad (46)$$

Under the assumptions (C1)-(C4) it can be easily shown that $\nabla d_{p, \gamma}(v_*, f_*) \rightarrow A(\theta_0)$ as $n \rightarrow \infty$ and $\nabla d_{p, \gamma}(v_*, f_*) = O_p(1)$. By a simple application of delta theorem on Equation 40 it can be shown $\sqrt{n}(\hat{\theta} - \theta_0) = O_p(1)$ under the null hypothesis. Equation 41 yields that $\sqrt{n}(\hat{\theta} - \theta_0) = O_p(1)$. Hence we have $\theta^* = \theta_0 + o_p(1)$. As a result the Equation 46 reduces to $\nabla d_{p, \gamma}(v_*, f_*) = A(\theta_0) + o_p(1)$. So, the Equation 45 becomes:

$$DDT_{p, \gamma}(v_*) = n(\hat{\theta} - \theta_0)^T A(\theta_0)(\hat{\theta} - \theta_0) + o_p(1) \quad (47)$$

To obtain the asymptotic null distribution of $DDT_{p, \gamma}(v_*)$ it is enough to obtain the asymptotic null distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$. Again from Equation 41 and by simple application delta theorem on Equation 40 it is easy to show that:

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow N(0, \Sigma_{p, \gamma}(\theta_0, \xi)),$$

where, $\xi$ is the true value of the parameter under $\xi$ formulation of the model. Hence the result follows. If $\Theta_0 = \{\theta_0\}$, then Equation 47 reduces to:

$$DDT_{p, \gamma}(v_*) = n(\hat{\theta} - \theta_0)^T A(\theta_0)(\hat{\theta} - \theta_0) + o_p(1).$$

We also know:

$$T_{p, \gamma}(\hat{\theta}, \theta_0) = n(\hat{\theta} - \theta_0)^T A(\theta_0)(\hat{\theta} - \theta_0) + o_p(1).$$

under the null hypothesis. Hence the asymptotic null distribution of both the statistics are same. This completes the proof.

Acknowledgement

The authors gratefully acknowledge the comments of three anonymous referees which led to an improved version of the manuscript.

Author’s Contributions

Saptarshi Roy: Developed the Wald type and score type tests for composite hypothesis. Provided real data examples in support of the findings.

Kaustav Chakraborty: Performed the simulation studies and the analysis of real data in different setups in the study. Contributed to the writing of the manuscript.

Somnath Bhadra: Contributed to develop robust regression. Contributed to the writing of the manuscript.

Ayanendranath Basu: Formulated the entire problem, designed the research plan, coordinated the data analysis and contributed to the writing of the manuscript.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Ali, S.M. and S.D. Silvey, 1966. A general class of coefficients of divergence of one distribution from another. J. Royal Stat. Society, 28: 131-142.

Banerjee, A., S. Merugu, I.S. Dhillon and J. Ghosh, 2005. Clustering with Bregman divergences. J. Machine Learn. Res., 6: 1705-1749.

Basu, A., I.R. Harris, N.L. Hjort and M.C. Jones, 1998. Robust and efficient estimation by minimising a density power divergence. Biometrika, 85: 549-559.

Basu, A., A. Mandal, N. Martin and L. Pardo, 2013. Testing statistical hypotheses based on the density power divergence. Annals Inst. Stat. Math., 65: 319-348.

Basu, A., H. Shioya and C. Park, 2011. Statistical Inference: The Minimum Distance Approach. 1st Edn., Chapman and Hall/CRC.

Bregman, L.M., 1967. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Comput. Math. Mathematical Phys., 7: 200-217.

Csiszar, I., 1963. Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten. Magyer Tud. Akad. Mat. Kutato Int. Koezl., 8: 85-108.

Csiszar, I., 1991. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. Annals Stat., 19: 2032-2066.

Ghosh, A. and A. Basu, 2013. Robust estimation for independent non-homogeneous observations using density power divergence with applications to linear regression. Electronic J. Stat., 7: 2420-2456.
Ghosh, A., A. Basu and L. Pardo, 2015. On the robustness of a divergence based test of simple statistical hypotheses. J. Stat. Plann. Inference, 161: 91-108.

Hampel, F.R., E.M. Ronchetti, P.J. Rousseeuw and W.A. Stahel, 1986. Robust Statistics: The Approach Based on Influence Functions. 1st Edn., John Wiley and Sons, New York.

Huber, P.J. and E.M. Ronchetti, 2009. Robust Statistics. 2nd Edn., John Wiley and Sons, New York.

Jones, L.K. and C.L. Byrne, 1990. General entropy criteria for inverse problems, with applications to data compression, pattern classification and cluster analysis. IEEE Trans. Inform. Theory, 36: 23-30.

Lindsay, B.G., 1994. Efficiency versus robustness: The case for minimum Hellinger distance and related methods. Annals Stat., 22: 1081-1114.

Maronna, R.A., R.D. Martin, V.J. Yohai and M. Salibian-Barrera, 2019. Robust Statistics. Theory and Methods (with R). 1st Edn., John Wiley and Sons, New York.

Osborn, J.F., 1979. Statistical Exercises in Medical Research. 1st Edn., John Wiley and Sons, New York.

Pardo, L., 2005. Statistical Inference Based on Divergence Measures. Statistics: A Series of Textbooks and Monographs. 1st Edn., CRC Press.

Rousseeuw, P. and A. Leroy, 1987. Robust Regression and Outlier Detection. 1st Edn., John Wiley and Sons, New York.

Ruppert, D. and R.J. Carroll, 1980. Trimmed least squares estimation in the linear model. J. Am. Stat. Assoc., 75: 828-838.

Stigler, S.M., 1977. Do robust estimators work with real data? Annals Stat., 5: 1055-1098.

Stummer, W. and I. Vajda, 2012. On bregman distances and divergences of probability measures. IEEE Trans. Inform. Theory, 58: 1277-1288.

Toma, A. and M. Broniatowski, 2011. Dual divergence estimators and tests: Robustness results. J. Multivariate Anal., 102: 20-36.

Warwick, J., 2002. Selecting tuning parameters in minimum distance estimators. PhD Thesis, The Open University.

Warwick, J. and M.C. Jones, 2005. Choosing a robustness tuning parameter. J. Stat. Comput. Simulat., 75: 581-588.