Journal of Mathematics and Statistics

Original Research Paper

Density Power Downweighting and Robust Inference: Some New Strategies

[1]Saptarshi Roy, [2]Kaustav Chakraborty, [3]Somnath Bhadra and [4]Ayanendranath Basu

[1]*Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA*
[2]*Department of Statistics, University of Illinois, Urbana-Champaign, IL 61820, USA*
[3]*Department of Statistics, University of Florida, Gainesville, FL 32611, USA*
[4]*Interdisciplinary Statistical Research Unit, Indian Statistical Institute, 203 B. T. Road, Kolkata 700108, India*

Article history
Received: 10-08-2019
Revised: 28-09-2019
Accepted: 14-11-2019

Corresponding Author:
Ayanendranath Basu
Interdisciplinary Statistical Research Unit, Indian Statistical Institute, 203 B. T. Road, Kolkata 700108, India
Email: ayanbasu@isical.ac.in

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 ϕ-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_θ be $\mathcal{X} \subseteq \mathbb{R}$. The Bregman divergence between the density g and model density f_θ is given by:

$$D_B(g, f_\theta) = \int_{\mathcal{X}} \begin{pmatrix} B(g(x)) - B(f_\theta(x)) \\ -(g(x) - f_\theta(x))B'(f_\theta(x)) \end{pmatrix} dx, \quad (1)$$

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_θ is the pdf of the parametric family, g is the true density, the minimization of the above divergence over the parameter space Θ 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 $B(y) = \frac{y^{1+\alpha}}{\alpha}$, $\alpha \geq 0$.

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

$$\int u_\theta(x) B''(f_\theta(x)) f_\theta^2(x) dx - \int u_\theta(x) B''(f_\theta(x)) f_\theta(x) g(x) dx = 0, \quad (2)$$

or equivalently:

$$\int u_\theta(x) B''(f_\theta(x)) f_\theta^2(x) dx - \int u_\theta(x) B''(f_\theta(x)) f_\theta(x) dG(x) = 0, \quad (3)$$

where, $u_\theta(x) = \nabla \log f_\theta(x)$ is the score function of the model $f_\theta(x)$, ∇ represents derivative with respect to θ 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:

$$\int u_\theta(x) B''(f_\theta(x)) f_\theta^2(x) dx - \frac{1}{n}\sum_{i=1}^{n} u_\theta(X_i) B''(f_\theta(X_i)) f_\theta(X_i) = 0. \quad (4)$$

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_θ. For the DPD, this weight becomes $(\alpha+1) f_\theta^\alpha(x)$.

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 $B(y) = \frac{y^{1+\alpha}}{\alpha}$ in

Equation 1. The general form of this divergence, as a function of a nonnegative tuning parameter α, is:

$$DPD_\alpha\left(g, f_\theta\right)=\int\left\{f_\theta^{1+\alpha}-\left(1+\frac{1}{\alpha}\right)gf_\theta^{\alpha}+\frac{1}{\alpha}g^{1+\alpha}\right\}. \quad (5)$$

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

$$DPD_\alpha\left(g, f_{T_\alpha(G)}\right)=\inf_{\theta\in\Theta} DPD_\alpha\left(g, f_\theta\right). \quad (6)$$

Under the estimation set up of this paper, the empirical objective function, ignoring the terms independent of θ, becomes:

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

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

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

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

$$\frac{1}{n}\sum_{i=1}^{n} u_\theta\left(X_i\right)=0. \quad (8)$$

Note that in the part involving real data in Equation 7, a downweighting effect is exerted on the score function $u_\theta(x)$ by the factor $f_\theta^{\alpha}(x)$. 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 $\sum_{i=1}^{n}\psi\left(X_i,\theta\right)=0$, where:

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

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 $u_\theta(x)$ applied by the MDPD estimating equation is $f_\theta^{\alpha}(x)$. 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^α for $\alpha > 0$ fixed. (Note, from Equation 4, the downweighting term for $u_\theta(x)$ in the general Bregman divergence is $f_\theta B''(f_\theta)$). 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 $\beta\in(0,1)$, $\lim_{x\to 0^+}\frac{xB''(x)}{x^\beta} = 0$, 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) $B''(x) = x^\beta\phi(x,\gamma)$ $(0 < \beta \le 1,\ 0 < \gamma \le 1)$. Where $\phi(x,\gamma)$ is a continuous and positive function over $\gamma\in(0,1]$ and $x > 0$. Furthermore, we demand $\lim_{\gamma\to 0^+}\phi(x,\gamma)=\frac{1}{x}$

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 ϕ-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:

- (A1) The pdfs f_θ of X have common support, so that the set $\mathcal{X}=\{x \mid f_\theta(x)>0\}$ is independent of θ. The distribution G is also supported on $\mathcal{X}$, on which the corresponding density g is greater than zero
- (A2) There is an open subset ω of the parameter space Θ, containing the best fitting parameter θ^g $\left(D_B\left(g, f_{\theta^g}\right)=\inf_{\theta\in\Theta} D_B\left(g, f_\theta\right)\right)$ 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 θ and the third partial derivatives are continuous with respect to θ. (The best fitting parameter θ^g depends on the index function $B(\cdot)$ also, but we suppress that notation for brevity)
- (A3) The integrals $\int B''\left(f_\theta(x)\right)f_\theta^2(x)dx$ and $\int B''\left(f_\theta(x)\right)f_\theta(x)g(x)dx$ can be differentiated with respect to θ and the derivatives can be taken under the integral sign
- (A4) The $p\times p$ matrix $J_B(\theta)$ defined by:

$$J_{B,kl}(\theta) = E_g\left\{\nabla_{kl}\begin{pmatrix}\int\left[B'\left(f_\theta(x)\right)f_\theta(x)-B\left(f_\theta(x)\right)\right]dx\\ -B'\left(f_\theta(X)\right)\end{pmatrix}\right\}$$

  is positive definite where E_g represents the expectation under the density g. Where ∇_{kl} represents the partial derivative with respect to the indicated components of θ.
- (A5) There exists functions $M_{jkl}(x)$, $j,k,\ l = 1,\ldots,p$, such that:

$$\left|\nabla_{jkl}\left(\int\left[B'\left(f_\theta(x)\right)f_\theta(x)-B\left(f_\theta(x)\right)\right]dx - B'\left(f_\theta(X)\right)\right)\right| \le M_{jkl}(X);\forall\theta\in\omega$$

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

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) $\sqrt{n}\left(\hat{\theta}_n-\theta^g\right)$ has an asymptotic multivariate normal distribution with mean vector zero and covariance matrix $J_B^{-1}K_BJ_B^{-1}$, where $J_B = J_B(\theta^g)$, $K_B = K_B(\theta^g)$, $K_B(\theta) = Var_g(u_\theta(X)\, f_\theta(X)B''(f_\theta(X)))$.

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

$$J_B=\int u_\theta u_\theta^T f_\theta^2 B''\left(f_\theta\right), K_B=\int u_\theta u_\theta^T f_\theta^3\left(B''\left(f_\theta\right)\right)^2-\zeta_B\zeta_B^T,\quad \zeta_B=\int u_\theta f_\theta^2 B''\left(f_\theta\right). \tag{10}$$

We are now going to establish that the DPD belongs to the class of ϕ-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 ϕ-DPD is:

$$\frac{1}{n}\sum_{i=1}^{n}u_\theta\left(X_i\right)f_\theta^{1+\beta}\left(X_i\right)\phi\left(f_\theta\left(X_i\right),\gamma\right) - \int u_\theta(x)\, f_\theta^{2+\beta}(x)\phi\left(f_\theta(x),\gamma\right)dx=0, \tag{11}$$

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

$$J_\phi=\int u_\theta u_\theta^T f_\theta^{2+\beta}\phi\left(f_\theta,\gamma\right),\quad K_\phi=\int u_\theta u_\theta^T f_\theta^{3+2\beta}\phi^2\left(f_\theta,\gamma\right)-\zeta_\phi\zeta_\phi^T,\quad \zeta_\phi=\int u_\theta f_\theta^{2+\beta}\phi\left(f_\theta,\gamma\right). \tag{12}$$

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

$$K_\phi=\int u_\theta u_\theta^T f_\theta^{2+2\beta}\phi^2\left(f_\theta,\gamma\right)g-\zeta_\phi\zeta_\phi^T,\quad \zeta_\phi=\int u_\theta f_\theta^{1+\beta}\phi\left(f_\theta,\gamma\right)g, \tag{13}$$

and:

$$J_\phi=\int u_\theta u_\theta^T f_\theta^{2+\beta}\phi\left(f_\theta,\gamma\right), +\int\left(\kappa_\theta-\beta u_\theta u_\theta^T\right)\left(g-f_\theta\right)f_\theta^{1+\beta}\phi\left(f_\theta,\gamma\right), \tag{14}$$

$$\kappa_\theta=-\left[\nabla_\theta u_\theta+\left(1+\frac{f_\theta\phi'\left(f_\theta,\gamma\right)}{\phi\left(f_\theta,\gamma\right)}\right)u_\theta u_\theta^T\right],\quad \phi'(x,\gamma)=\frac{\partial\phi(x,\gamma)}{\partial x}. \tag{15}$$

Remark

Notice that $1+\frac{x\phi'(x,\gamma)}{\phi(x,\gamma)}=\frac{1}{\phi(x,\gamma)}\frac{\partial}{\partial x}\left[x\phi(x,\gamma)\right]$. If $\lim_{\gamma\to0^+}$ and $\frac{\partial}{\partial x}$ are interchangeable for $\phi(\cdot)$ then by (P4) it can be concluded that κ_θ converges to i_θ as $\gamma\to0^+$.

Theorem 2

If $u_\theta(x)\ f_\theta(x)^{1+\beta}$, $u_\theta(x)\ u_\theta(x)^T\ f_\theta(x)^{1+2\beta}$, $u_\theta(x)u_\theta(x)^T f_\theta(x)^{1+\beta}$ are integrable and $f_\theta(x)\phi(f_\theta(x), \gamma)$ is bounded by some universal constant then the following hold:

(a) The usual DPD defined in Equation 5 is a special limiting case of ϕ-DPD
(b) If $g = f_\theta$ for some $\theta\in\Theta$ and if for the DPD there exists α, β such that the Asymptotic Relative Efficiency (ARE) of the estimator under tuning parameter β is greater than that of the estimator under tuning parameter α, then there exists γ such that ϕ-DPD with tuning parameter (β, γ) generates an estimator with higher ARE than the MDPDE with tuning parameter α

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

Our choice for $B''(\cdot)$ in the LϕDPD case is $B''(x) = \frac{1}{\gamma}x^\beta\log\left(1+\frac{\gamma}{x}\right) 0 < \beta\le 1,\ 0 < \gamma\le 1,\ x > 0$. The corresponding B function may be expressed in the integral form as:

$$B(x) = \frac{1}{\gamma}\int_0^x\int_0^t s^\beta\log\left(1+\frac{\gamma}{s}\right)ds\,dt. \tag{16}$$

Obviously other choices are possible, but we have found the LϕDPD 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 LϕDPD. We will refer to the corresponding minimum distance estimator as MLϕDE.

The Influence Function of MLϕDE

It is easy to see that the MLϕDE is also an M-estimator. Let the minimum LϕDPD functional $T_{\beta,\gamma}(G)$ be defined as:

$$L\phi DPD_{\beta,\gamma}\left(g, f_{T_{\beta,\gamma}(G)}\right) = \inf_{\theta\in\Theta} L\phi DPD_{\beta,\gamma}\left(g, f_\theta\right).$$

Under $g = f_\theta$ the influence function of this minimum distance estimator simplifies to:

$$IF\left(y, T_{\beta,\gamma}, G\right) = \left[\int u_\theta u_\theta^T f_\theta^{2+\beta}\log\left(1+\frac{\gamma}{f_\theta}\right)\right]^{-1} \left[u_\theta(y)\, f_\theta^{1+\beta}(y)\log\left(1+\frac{\gamma}{f_\theta(y)}\right) - \int u_\theta f_\theta^{2+\beta}\log\left(1+\frac{\gamma}{f_\theta}\right)\right]$$

for $0 < \beta\le 1, 0 < \gamma\le 1$. If $\int u_\theta u_\theta^T f_\theta^{2+\beta}\log\left(1+\frac{\gamma}{f_\theta}\right)$ and $\int u_\theta f_\theta^{2+\beta}\log\left(1+\frac{\gamma}{f_\theta}\right)$ are finite then the expressions in Equation 13 are finite if $u_\theta(y)\, f_\theta^{1+\beta}(y)\log\left(1+\frac{\gamma}{f_\theta(y)}\right)$ is finite which is indeed the case for most parametric models suggesting the observed robustness of the MLϕDE under those parametric models.

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

The Breakdown Point Under the Location Model

Now we will establish the breakdown point of the minimum LϕDPD functional under the location family of densities $\mathcal{F} = \{f_\theta(x) = f(x-\theta): \theta\in\Theta\}$. Let $B(\cdot)$ be the function defined in Equation 16. Define the quantities:

$$\int B\left(\varepsilon f\left(x-\theta\right)\right)dx = \int B\left(\varepsilon f\left(x\right)\right)dx := M_{f,\varepsilon}^{(1)},$$
$$\int\left[B\left(f\left(x-\theta\right)\right) + \left(\varepsilon-1\right)f\left(x-\theta\right)B'\left(f\left(x-\theta\right)\right)\right]dx = \int\left[B\left(f\left(x\right)\right) + \left(\varepsilon-1\right)f\left(x\right)B'\left(f\left(x\right)\right)\right]dx := M_{f,(\varepsilon-1)}^{(2)}.$$

Define $d(g, f) = B(g)-B(f)-(g-f)B'(f)$ and let $D(g, f) = \int d(g, f)$. From Equation 16 we have $d(g, 0) = \lim_{f\to 0^+} d(g, f) = B(g)$.

Consider the contamination model $H_{\varepsilon,n} = (1-e)G + \varepsilon K_n$, where $\{K_n\}$ is a sequence of contaminating distributions. Let $h_{\varepsilon,n}$, g and k_n be the corresponding densities. We say that there is breakdown in the minimum LϕDPD functional for ε level contamination if there exists a sequence K_n such that $|T_{\beta,\gamma}(H_{\varepsilon,n})-T_{\beta,\gamma}(G)| \to \infty$ as $n\to\infty$. We write below $\theta_n = T_{\beta,\gamma}(H_{\varepsilon,n})$ and assume that the true distribution belongs to the model family, i.e., $g = f_{\theta^0}$. We make the following assumptions:

- (BP1) $\int \min\{f_\theta(x), k_n(x)\}dx \to 0$ as $n\to\infty$ uniformly for $|\theta|\le c$ 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) $\int \min\{f_{\theta^g}(x), f_{\theta_n}(x)\}dx \to 0$ as $n\to\infty$ if $|\theta|\to\infty$ as $n\to\infty$, i.e., large values of θ give distributions which become asymptotically singular to the true distribution
- (BP3) The contaminating sequence $[k_n\}$ is such that:

$$D(\varepsilon k_n, f_\theta) \ge D(\varepsilon f_\theta, f_\theta) = M^{(1)}_{f,\varepsilon} - M^{(2)}_{f,(\varepsilon-1)}$$

for any $\theta\in\Theta$ and $0 < \varepsilon < 1$ and $\limsup_{n\to\infty} \int B(\varepsilon k_n) \le M^{(1)}_{f,\varepsilon}$.

Theorem 3

Under the assumptions (BP1)-(BP3) above, the asymptotic breakdown point ε^* of the LϕDPD functional is at least 0.5 at the location model.

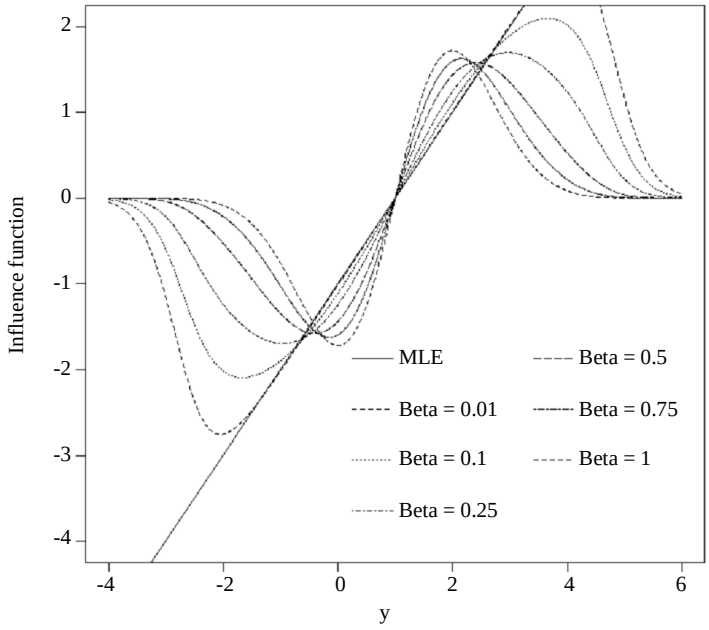

Fig. 1: Influence function of the MLϕDE for various values of b with fixed $\gamma = 0.001$ under $N(\mu,1)$ model at $N(1,1)$

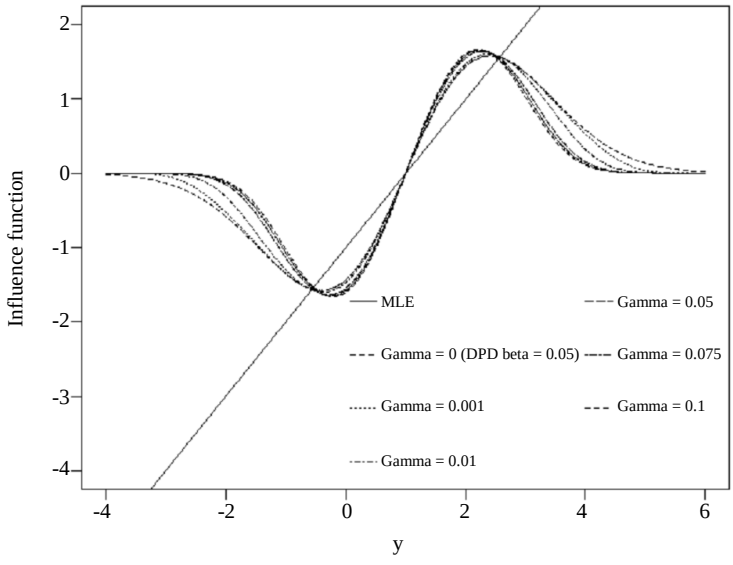

Fig. 2: Influence function of the MLϕDEs for various values of γ with fixed $\beta = 0.5$ under the $N(\mu,1)$ model at $N(1,1)$

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 $N(\mu,1)$ model at a given level of contamination. In the following study data are generated from two normal mixtures, $0.9N(0,1) + 0.1N(5,1)$ and $0.8N(0,1) + 0.2N(5,1)$, 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., $\mu = 0$. 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 $\alpha\in(0, 1)$, it may be possible to choose $\beta\in(0, \alpha)$ and $\gamma\in(0, 1)$ so that, in relation to our numerical study, MLϕDE$_{\beta, \gamma}$ beats MDPDE$_\alpha$ both in terms of asymptotic model efficiency and the empirical mean square error under contamination. As an illustration, consider MDPDE$_{0.5}$ in the first contaminated model. The corresponding MSE and asymptotic relative efficiency are 0.0294 and 83.8% respectively. Now choose the LϕDPD parameter $(\beta, \gamma) = (0.3, 0.01)$. In this case, the corresponding MSE and efficiency of the MLϕDE are 0.0281 and 89% respectively. Similarly MLϕDE$_{0.2,0.04}$ appears to dominate MDPDE$_{0.4}$ 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 theγ = 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	0.02900	0.0259	0.0271	0.02780	0.0282	0.0259
0.2	0.0560	0.0273	0.0277	0.02540	0.0252	0.0248	0.02680	0.0260	0.0266
0.3	0.0360	0.0281	0.0257	0.02670	0.0277	0.0264	0.02790	0.0268	0.0266
0.4	0.0268	0.0267	0.0261	0.02650	0.0270	0.0261	0.02890	0.0273	0.0265
0.5	0.0294	0.0277	0.0276	0.02760	0.0276	0.0307	0.02840	0.0296	0.0291
0.6	0.0275	0.0272	0.0298	0.03070	0.0293	0.0305	0.02950	0.0293	0.0296
0.7	0.0277	0.0276	0.0296	0.02940	0.0301	0.0320	0.03110	0.0311	0.0300
0.8	0.0292	0.0305	0.0327	0.03080	0.0342	0.0300	0.02880	0.0336	0.0315
0.9	0.0309	0.0299	0.0313	0.03445	0.0309	0.0295	0.03580	0.0326	0.0320
1	0.0313	0.0320	0.0350	0.03620	0.0369	0.0361	0.03368	0.0340	0.0335

Table 3: Empirical MSE of the MLϕDE and DPDE for different values of (β, γ) under 20% contaminated data for location model. Here the $\gamma = 0$ column represents the MDPDE

β	$\gamma = 0$	$\gamma = 0.01$	$\gamma = 0.02$	$\gamma = 0.03$	$\gamma = 0.04$	$\gamma = 0.05$	$\gamma = 0.06$	$\gamma = 0.07$	$\gamma = 0.08$
0.1	0.3214	0.0329	0.0314	0.0308	0.0306	0.0305	0.0304	0.0304	0.0305
0.2	0.0786	0.0312	0.0306	0.0305	0.0305	0.0305	0.0306	0.0307	0.0306
0.3	0.0414	0.0308	0.0306	0.0307	0.0309	0.0310	0.0311	0.0309	0.0313
0.4	0.0342	0.0311	0.0312	0.0314	0.0316	0.0318	0.0320	0.0322	0.0322
0.5	0.0327	0.0317	0.0320	0.0323	0.0326	0.0328	0.0330	0.0323	0.0324
0.6	0.0329	0.0327	0.0331	0.0334	0.0351	0.0372	0.0226	0.0303	0.0310
0.7	0.0366	0.0413	0.0409	0.0346	0.0366	0.0364	0.0418	0.0421	0.0423
0.8	0.0382	0.0388	0.0394	0.0399	0.0403	0.0407	0.0410	0.0408	0.0412
0.9	0.0424	0.0428	0.0432	0.0436	0.0438	0.0442	0.0445	0.0447	0.0309
1	0.0437	0.0442	0.0446	0.0390	0.0293	0.0445	0.0267	0.0452	0.0467

Algorithm for Finding the Optimal (β, γ)

The LϕDPD can generate many different kinds of estimators, starting from the most efficient estimator to highly robust estimators. For example, in the limit $\gamma \to 0$ and $\beta \to 0$, one gets the likelihood disparity which is minimized by the classical maximum likelihood estimator. On the other hand, relatively larger values of β and γ 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 (β,γ) 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 θ_0 be the parameter for the model component. Although the data are generated by a contaminated version, the parameter θ_0 of the model component is our target parameter. The spirit of such a set up is described in Warwick and Jones (2005). Let $\theta_{\beta,\gamma} = T_{\beta,\gamma}(G)$ be the corresponding minimum distance functional and $\hat{\theta}_{\beta,\gamma}$ is the solution of the unbiased equation of LϕDPD with tuning parameter (β, γ) based on the data. The summed mean square error of the minimum LϕDPD estimator has the asymptotic formula:

$$E\left[\left(\hat{\theta}_{\beta,\gamma}-\theta^*\right)^T\left(\hat{\theta}_{\beta,\gamma}-\theta^*\right)\right] = \left(\theta_{\beta,\gamma}-\theta^*\right)^T\left(\theta_{\beta,\gamma}-\theta^*\right)+n^{-1}tr\left\{\mathrm{var}\left(\hat{\theta}_{\beta,\gamma}\right)\right\}. \quad (17)$$

Here θ^* is the pilot estimator playing the role of θ_0 and tr{·} represents the trace of matrix. The asymptotic covariance matrix of $\sqrt{n}\left(\hat{\theta}_{\beta,\gamma}-\theta_{\beta,\gamma}\right)$ is $J^{-1}KJ^{-1}$, where J and K are as in Equation 12 with $\phi(x, g) = \frac{1}{\gamma}\log\left(1+\frac{\gamma}{x}\right)$. So the estimated asymptotic summed mean square of the MLϕDE is:

$$\left(\theta_{\beta,\gamma}-\theta^*\right)^T\left(\theta_{\beta,\gamma}-\theta^*\right)+\frac{1}{n}J^{-1}KJ^{-1}. \quad (18)$$

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 (μ, σ) for $N(\mu, \sigma)$ model) then the expression to be minimized is:

$$n^{-1}tr\left\{J^{-1}\left(\theta_{\beta,\gamma}\right)K\left(\theta_{\beta,\gamma}\right)\right)J^{-1}\left(\theta_{\beta,\gamma}\right)\right\} +\left(\mu-\mu^*\right)^2+\left(\sigma-\sigma^*\right)^2. \quad (19)$$

The optimal value of (β, γ) is the minimizer of Equation 19 under certain conditions. One important note is that in the first stage of minimization our pilot estimate for θ^* 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 θ:

1. Apply the method suggested in Warwick (2002) to get an optimal α for MDPDE. Suppose this value is α_w. This step is the 1st stage of optimization by assuming an initial pilot estimate of θ^*.

2. Consider the interval (0, α_w). Update the pilot estimate for $\theta^* = \hat{\theta}_{\alpha_w}$, which is MDPDE of θ with α_w as the tuning parameter
3. Perform a two dimensional optimization which selects the value of (β, γ) for which the minimum:

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

is attained under the constraint $AMSE\left(\hat{\theta}_{\beta,\gamma}\right) < AMSE\left(\hat{\theta}_{\alpha_w}\right)$.

An alternative to this approach could be to perform an unrestricted minimization of $AMSE\left(\hat{\theta}_{\beta,\gamma}\right)$ with respect to (β, γ) over the set (0,1)×(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 (μ, σ).

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 LϕDPD, as well as the parameter estimates at these optimal values. The estimators are extremely close, but the estimated asympmtotic summed mean square, for whatever it is worth, is lower in case of the MLϕDE.

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 MLϕDE 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 MLϕDE 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 MLϕDE is better than that of the MDPDE, although the order of improvement is small.

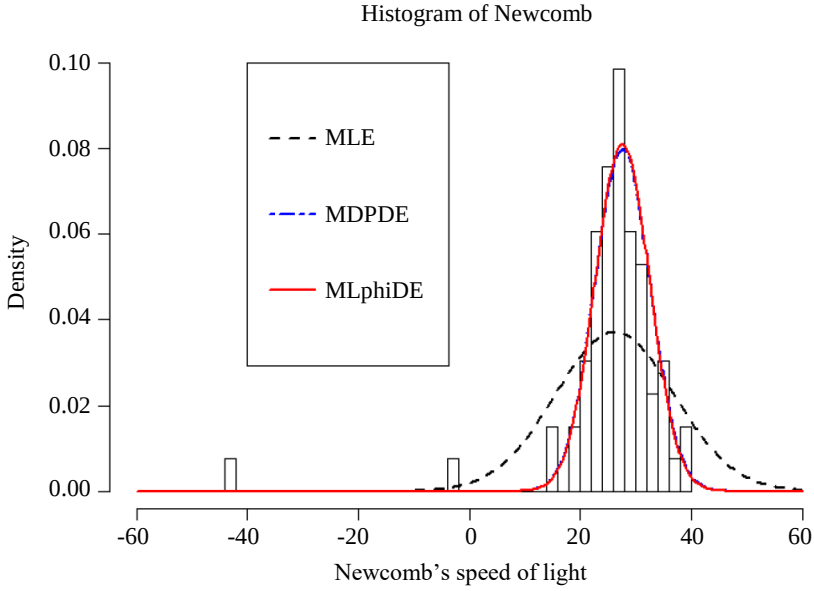

Fig. 3: Normal density fits for Newcomb's data

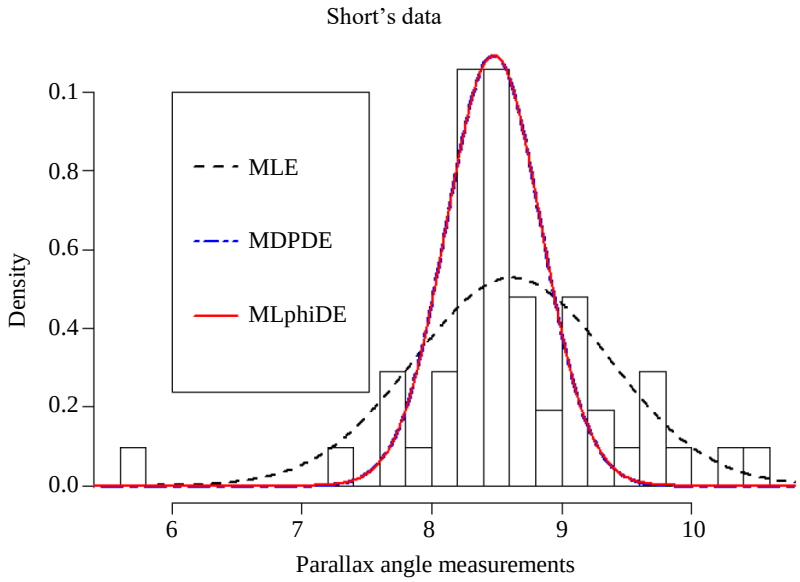

Fig. 4: Normal density fits for Short's data

Table 4: Parameter estimates: Newcomb's light speed data

Category	MDPDE	MLϕDE
Optimal tuning parameter	$\alpha = 0.3$	$(\beta, \gamma) = (0.1, 0.03)$
Estimate of μ	27.62	27.57
Estimate of σ	5.01	4.93
AMSE	0.7	0.64

Table 5: Parameter estimates: Short's data

Category	MDPDE	MLϕDE
Optimal tuning parameter	$\alpha = 0.96$	$(\beta, \gamma) = (0.55, 1)$
Estimate of μ	8.477	8.478
Estimate of σ	0.365	0.365
AMSE	0.0058	0.0057

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 $Y_1,\ldots,Y_n$ are independent but for each i, $Y_i \sim g_i$ where the densities $g_1,\ldots,g_n$ may not be same. We want to model g_i by the family $\mathcal{F}_{i,\theta} = \{f_i(\cdot;\theta) | \theta \in \Theta\}$ for all $i = 1,2,\ldots,n$. We want to estimate θ 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:

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

with respect to $\theta \in \Theta$, where $d\left(\hat{g}_i, f_i\ (\cdot;\theta)\right)$ 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:

$$\begin{aligned} & d\left(\hat{g}_i, f_i\ (\cdot;\theta)\right) \\ &= \frac{1}{\gamma}\int\left[f_i\ (y;\theta)\int_0^{f_i\ (y;\theta)} s^{\beta}\log\left(1+\frac{\gamma}{s}\right)ds \right. \\ &\left. -\int_0^{f_i\ (y;\theta)}\int_0^{t} s^{\beta}\log\left(1+\frac{\gamma}{s}\right)ds\,dt \right]dy \\ &-\frac{1}{\gamma}\int_0^{f_i\ (Y_i;\theta)} s^{\beta}\log\left(1+\frac{\gamma}{s}\right)ds + K. \end{aligned}$$

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

$$H_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} V_i\ (Y_i;\theta), \qquad (21)$$

where:

$$
\begin{aligned}
&V_i\left(Y_i;\theta\right)\\
&=\frac{1}{\gamma}\int\Big[f_i\left(y;\theta\right)\int_0^{f_i\left(y;\theta\right)}s^{\beta}\log\left(1+\frac{\gamma}{s}\right)ds\\
&-\int_0^{f_i\left(y;\theta\right)}\int_0^{t}s^{\beta}\log\left(1+\frac{\gamma}{s}\right)ds\,dt\Big]dy.\\
&-\frac{1}{\gamma}\int_0^{f_i\left(Y_i;\theta\right)}s^{\beta}\log\left(1+\frac{\gamma}{s}\right)ds.
\end{aligned}
\tag{22}
$$

Differentiating the above with respect to θ we get the estimating equation of the minimum LϕDPD estimator for non-homogeneous observations as:

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\Big[f_i\left(Y_i;\theta\right)^{\beta+1}\log\left(1+\frac{\gamma}{f_i\left(Y_i;\theta\right)}\right)u_i\left(Y_i;\theta\right)\\
&-\int f_i\left(y;\theta\right)^{\beta+2}\log\left(1+\frac{\gamma}{f_i\left(y;\theta\right)}\right)u_i\left(y;\theta\right)ds\Big]=0,
\end{aligned}
\tag{23}
$$

where, $u_i(\cdot)$ is the score function for $f_i(\cdot)$.

Asymptotic Properties

We will now derive the asymptotic distribution of the minimum LϕDPD estimator $\hat{\theta}_n$ defined by the relation:

$$H_n\left(\hat{\theta}_n\right)=\min_{\theta\in\Theta}H_n\left(\theta\right)$$

provided such a minimum exists. Let us first present the necessary set up and conditions. Let the parametric model $\mathcal{F}_{i,\theta}$ be as defined above. We also assume that there exists a best fitting parameter of θ which is independent of the index i of the different densities. Let us denote it by θ^g. The assumptions hold if all the true densities g_i belong to the model family so that $g_i = f_i(\cdot;\theta)$ for some common θ and in that case the best fitting parameter is nothing but the true parameter θ.

Next, recall that the MLϕDE $\hat{\theta}_n$ is obtained as a solution of the estimating Equation 23. This equation is satisfied by the minimizer of $\underline{H_n}(\theta)$ in Equation 21. Similarly, we also define, for $i = 1, 2,\cdots$:

$$
\begin{aligned}
&H^{(i)}\left(\theta\right)\\
&=\frac{1}{\gamma}\int\Big[f_i\left(y;\theta\right)\int_0^{f_i\left(y;\theta\right)}s^{\beta}\log\left(1+\frac{\gamma}{s}\right)ds\\
&-\int_0^{f_i\left(y;\theta\right)}\int_0^{t}s^{\beta}\log\left(1+\frac{\gamma}{s}\right)ds\,dt\Big]dy.\\
&-\frac{1}{\gamma}\Big[\int_0^{f_i\left(y;\theta\right)}s^{\beta}\log\left(1+\frac{\gamma}{s}\right)ds\Big]g_i\left(y\right)dy.
\end{aligned}
\tag{24}
$$

Note, at the best fitting parameter θ^g, we must have:

$$\nabla H^{(i)}\left(\theta^g\right)=0, i=1,2,\cdots$$

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

$$J_{kl}^{(i)}=E_{g_i}\left[\nabla_{kl}V_i\left(Y_i;\theta\right)\right],\tag{25}$$

where, ∇kl represents the partial derivative with respect to the indicated components of θ. We further define the quantities:

$$\Psi_n=\frac{1}{n}\sum_{i=1}^{n}J^{(i)},\tag{26}$$

$$\Omega_n=\frac{1}{n}\sum_{i=1}^{n}Var_{gi}\left[\nabla V_i\left(Y_i;\theta\right)\right].\tag{27}$$

A simple calculation shows that:

$$
\begin{aligned}
&J^{(i)}\\
&=\frac{1}{\gamma}\int u_i\left(y;\theta^g\right)u_i^T\left(y;\theta^g\right)f_i^{\beta+2}\left(y;\theta^g\right)\log\left(1+\frac{\gamma}{f_i\left(y;\theta^g\right)}\right)dy\\
&-\frac{1}{\gamma}\int\Big[\Big\{\nabla u_i\left(y;\theta^g\right)+\left(\beta+1\right)u_i\left(y;\theta^g\right)u_i^T\left(y;\theta^g\right)\Big\}\\
&\log\left(1+\frac{\gamma}{f_i\left(y;\theta^g\right)}\right)-u_i\left(y;\theta^g\right)u_i^T\left(y;\theta^g\right)\left(\frac{\gamma}{\gamma+f_i\left(y;\theta^g\right)}\right)\Big]\\
&\Big\{g_i\left(y;\theta^g\right)-f_i\left(y;\theta^g\right)\Big\}f_i^{\beta+1}\left(y;\theta^g\right)dy
\end{aligned}
\tag{28}
$$

and:

$$
\begin{aligned}
&\Omega_n\\
&=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\gamma}\int\left\{\int_0^{f_i\left(y;\theta\right)}s^{\beta}\log\left(1+\frac{\gamma}{s}\right)ds\right\}^2g_i\left(y;\theta\right)dy\\
&-\frac{1}{n}\sum_{i=1}^{n}\xi_i\,\xi_i^T,
\end{aligned}
\tag{29}
$$

where:

$$\xi_i=\frac{1}{\gamma}\int\left\{\int_0^{f_i\left(y;\theta\right)}s^{\beta}\log\left(1+\frac{\gamma}{s}\right)ds\right\}g_i\left(y;\theta\right)dy.\tag{30}$$

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

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

- (G2) There is an open subset ω of the parameter space Θ, containing the best fitting parameter θ^g such that for almost all $y \in \mathcal{Y}$ and all $\theta \in \Theta$, all i = 1,2,⋯, the density $f_i(y;\theta)$ is thrice differentiable with respect to θ and the third partial derivatives are continuous with respect to θ
- (G3) For each i = 1,2,⋯, the three integrals $\int f_i(y;\theta)\int_0^{f_i(y;\theta)} s^\beta \log\left(1+\frac{\gamma}{s}\right) ds\, dy$, $\int\int_0^{f_i(y;\theta)}\int_0^t s^\beta \log\left(1+\frac{\gamma}{s}\right) ds\, dt\, dy$ and $\int\left[\int_0^{f_i(y;\theta)} s^\beta \log\left(1+\frac{\gamma}{s}\right) ds\right] g_i(y)\, dy$ can be differentiated thrice with respect to θ and the derivatives can be taken under the integral sign (the first indefinite integral)
- (G4) For each i = 1,2,⋯, the matrices $J^{(i)}$ are positive definite and:

$$\lambda_0 = \inf_n \left[\min eigenvalue of\ \Psi_n\right] > 0.$$

- (G5) There exists functions $M_{jkl}^{(i)}(Y)$ such that:

$$|\nabla_{jkl} V_i(Y;\theta)| \le M_{jkl}^{(i)}(Y) \forall \theta \in \Theta, \forall i$$

with $E_{g_i}|M_{jkl}^{(i)}(Y)| < \infty\ \forall j, k, l$.

- (G6) For all j,k, we have:

$$\limsup_{N\to\infty}{}_{n\ge 1}\left\{\frac{1}{n}\sum_{i=1}^n E_{g_i}\left[|\nabla_j V_i(Y;\theta)| I\left(|\nabla_j V_i(Y;\theta)| > N\right)\right]\right\} = 0, \quad (31)$$

$$\limsup_{N\to\infty}{}_{n\ge 1}\left\{\frac{1}{n}\sum_{i=1}^n E_{g_i}\left[|\nabla_{jk} V_i(Y;\theta) - E_{g_i}\left(\nabla_{jk} V_i(Y;\theta)\right)| \times I\left(|\nabla_{jk} V_i(Y;\theta) - E_{g_i}\left(\nabla_{jk} V_i(Y;\theta)\right)| > N\right)\right]\right\} = 0. \quad (32)$$

Here $I(\cdot)$ stands for indicator function.

- (G7) For all $\varepsilon > 0$, we have:

$$\lim_{N\to\infty}\left\{\frac{1}{n}\sum_{i=1}^n E_{g_i}\left[\begin{array}{l}\|\Omega_n^{-1/2}\nabla V_i(Y;\theta)\|^2 \\ I\left(\|\Omega_n^{-1/2}\nabla V_i(Y;\theta)\| > \varepsilon\sqrt{n}\right)\end{array}\right]\right\} = 0 \quad (33)$$

Theorem 4

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

(i) There exists a consistent sequence θ_n of roots to the minimum LϕDPD estimating Equation 23

(ii) The asymptotic distribution of $\Omega_n^{-\frac{1}{2}}\Psi_n\left[\sqrt{n}\left(\theta_n - \theta^g\right)\right]$ is p-dimensional normal with (vector) mean 0 and covariance matrix I_p, the p-dimensional identity matrix

Note that, putting $f_i = f$ for all i, we get back the corresponding asymptotic properties of the minimum LϕDPD estimator for the i.i.d. case. If $f_i = f$, i = 1,2,⋯, 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 LϕDPD estimator $\hat{\theta}$ with $n^{1/2}\left(\hat{\theta} - \theta^g\right)$ having the asymptotic covariance matrix $\Psi_n^{-1}\Omega_n\Psi_n^{-1} = J^{-1}KJ^{-1}$, 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:

$$y_i = x_i^T\beta + \varepsilon_i,\ i = 1,\cdots,n, \quad (34)$$

where the error ε_i's are i.i.d. normal variables with mean zero and variance σ^2, $x_i^T = \left(x_{ii},\cdots x_{ip}\right)$ is the vector of the independent variables corresponding to the i-th observation and $\beta = (\beta_1,\cdots \beta_p)^T$ represents the regression coefficients. We will assume that x_i's are fixed. Then $y_i \sim N\left(x_i^T\beta, \sigma^2\right)$ and hence the y_i's are independent but not identically distributed. Thus y_i's satisfy our independent but non-homogenous set-up and hence the MLϕDE of the parameter $\theta = (\beta^T, \sigma^2)^T$ can be obtained by minimizing the expression in Equation 21 with $f_i \equiv N\left(x_i^T\beta, \sigma^2\right)$.

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 Rousseeuw 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 LϕDPD regression are given in the Fig. 5. The robust performance of the MLϕDE 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 LϕDPD regression are given in the Fig. 6. Like the LMS method (and unlike the LS method) the MLϕDE gives a nice outlier resistant fit.

Table 6: Regression estimates for Hertzsprung-Russel data

Category	MLϕDE
Tuning Parameter	$(\beta, \gamma) = (1,0.9)$
Estimate of η_0	−8.5557324
Estimate of η_1	3.0590795
Estimate of σ	0.4266284

Table 7: Regression estimates for Salinity data

Category	MLϕDE
Tuning Parameter	$(\beta, \gamma) = (1,0.9)$
Estimate of η_0	57.16780461
Estimate of η_1	0.06010002
Estimate of η_2	−0.01301208
Estimate of η_3	−2.08372562
Estimate of σ	0.56157558

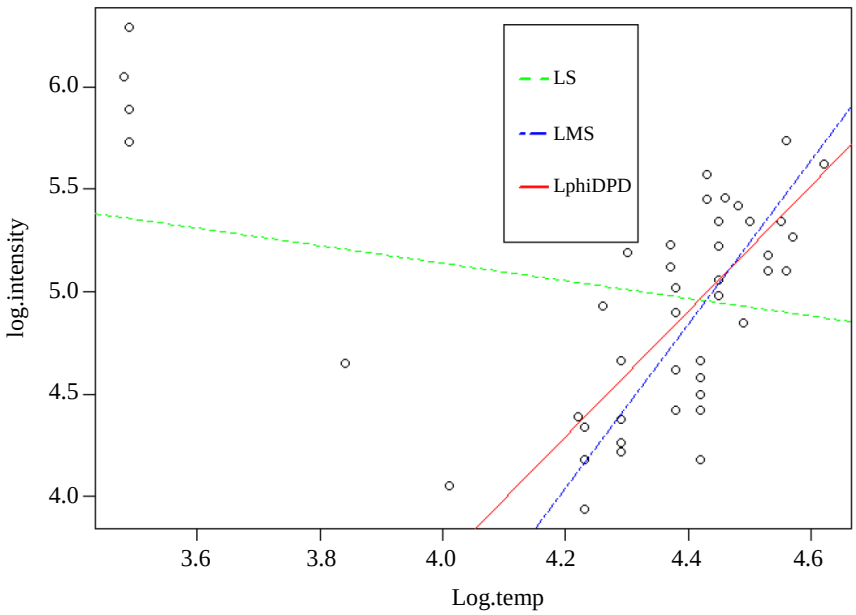

Fig. 5: Regression fits for the Hertzsprung-Russel data

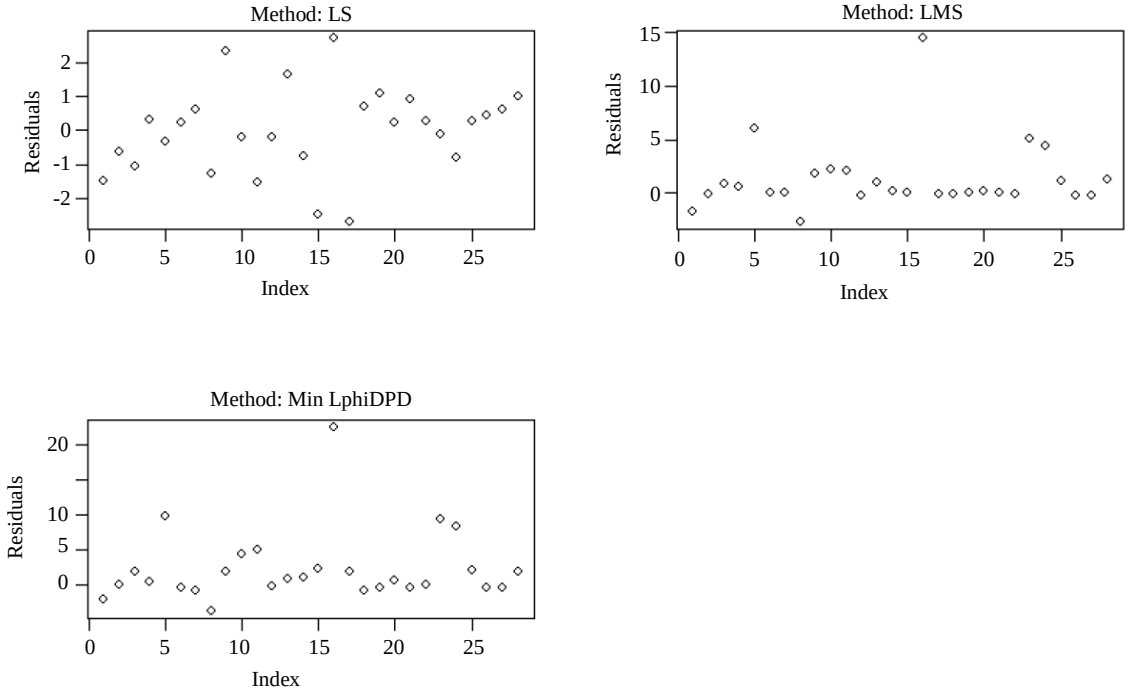

Fig. 6: Residual plots of the fitted regression models for Salinity data using LS, LMS and minimum LϕDPD estimation

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 $\{f_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$ under the one sample case. Here we test:

$$H_0 : \theta = \theta_0 \; Versus \; H_1 : \theta \neq \theta_0 \tag{35}$$

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

$$T = T_{\beta,\gamma}\left(\hat{\theta}, \theta_0\right) = 2n d_{\beta,\gamma}\left(f_{\hat{\theta}}, f_{\theta_0}\right)$$

where:

$$\begin{aligned} & d_{\beta,\gamma}\left(f_{\hat{\theta}}, f_{\theta_0}\right) \\ & = \int \Big[B\left(f_{\hat{\theta}}(x)\right) - B\left(f_{\theta_0}(x)\right) \\ & - \left(f_{\hat{\theta}}(x) - f_{\theta_0}(x)\right) B'\left(f_{\theta_0}(x)\right) \Big] dx, \end{aligned} \tag{36}$$

with $\hat{\theta} = \hat{\theta}_{\beta,\gamma}$ being the MLϕDE estimate of θ 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_θ, i.e., the set $\mathcal{X} = \{x \mid f_\theta(x) > 0\}$ is independent of θ
- (B2) There is an open subset ω of the parameter space Θ, containing the true parameter value θ_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 θ and the third partial derivatives are continuous with respect to θ
- (B3) The integrals $\int B''\left(f_\theta(x)\right) f_\theta^2(x) dx$ can be differentiated with respect to θ and the derivatives can be taken under the integral sign
- (B4) The $p \times p$ matrix $J(\theta)$ defined by:

$$\begin{aligned} & J_{B,kl}(\theta) = \\ & E_\theta \left\{ \nabla_{kl} \left(\int \Big[B'\left(f_\theta(x)\right) f_\theta(x) - B\left(f_\theta(x)\right) \Big] dx - B'\left(f_\theta(X)\right) \right) \right\} \end{aligned}$$

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

$$\begin{aligned} & \left| \nabla_{jkl} \left(\int \Big[B'\left(f_\theta(x)\right) f_\theta(x) - B\left(f_\theta(x)\right) \Big] dx - B'\left(f_\theta(x)\right) \right) \right| \\ & \leq M_{jkl}(X); \forall \theta \in \omega. \end{aligned}$$

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 $T_{\beta,\gamma}\left(\hat{\theta},\theta_0\right)$ coincides with the distribution of:

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

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

$$r = rank\left(\Sigma\left(\theta_0\right)A\left(\theta_0\right)\Sigma\left(\theta_0\right)\right)$$
$$A\left(\theta_0\right)\nabla_\theta^2 d_{\beta,\gamma}\left(f_\theta, f_{\theta_0}\right)|_{\theta=\theta_0}$$

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

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 θ_1 and that of size m from population 2 with parameter θ_2. Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be MLϕDEs 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:

$$S = S_{\beta,\gamma}\left(\hat{\theta}_1,\hat{\theta}_2\right) = \frac{2mn}{m+n} d_{\beta,\gamma}\left(f_{\hat{\theta}_1}, f_{\hat{\theta}_2}\right)$$

coincides with that of:

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

where, Z_is are independent standard normals and λ_i's are non-zero eigenvalues of $A(\theta_0)\Sigma(\theta_0)$ and r = rank$(\Sigma(\theta_1)A(\theta_1)\ \Sigma(\theta_1))$ 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 LϕDPD for testing the simple null in Equation 35 can be given by:

$$n\bar{U}^T\left(\theta_0\right)J_B^{-1}\left(\theta_0\right)A\left(\theta_0\right)J_B^{-1}\left(\theta_0\right)\bar{U}\left(\theta_0\right)$$

where:

$$U_\theta(x) = u_\theta(x)B''\left(f_\theta(x)\right)f_\theta(x) - \int u_\theta(x)B''\left(f_\theta^2(x)\right)f_\theta^2(x)dx$$

and:

$$\bar{U}(\theta) = \frac{1}{n}\sum_{i=1}^{n}U_\theta\left(X_i\right)$$

with

$$B''(x) = \frac{x^\beta}{\gamma}\log\left(1+\frac{\gamma}{x}\right)$$
$$J_B\left(\theta_0\right) = -E_{\theta_0}\frac{\partial}{\partial_\theta}U_\theta\left(X_1\right)|_{\theta=\theta_0}$$

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 $T_{\beta,\gamma}\left(\hat{\theta},\theta_0\right)$.

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:

$$H_0 : \theta \in \Theta_0 \; versus \; H_1 : \theta \in \Theta \setminus \Theta_0, \tag{37}$$

where, Θ_0 is a proper subset of Θ. The Likelihood Ratio Test (LRT) is one of the most common tests that may be employed in this situation. Define:

$$\lambda = \frac{\sup_{\theta\in\Theta_0}\mathcal{L}\left(\theta \mid X_1,\ldots,X_n\right)}{\sup_{\theta\in\Theta}\mathcal{L}\left(\theta \mid X_1,\ldots,X_n\right)},$$

where, $\mathcal{L}\left(\theta \mid X_1,\ldots,X_n\right)$ is the likelihood of q given the data. The test statistic in this case is -2logλ. 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 as:

$$
\begin{aligned}
&-2\log\lambda \\
&= 2\left[\log\left(\prod_{i=1}^{n} f_{\hat{\theta}}\left(X_i\right)\right) - \log\left(\prod_{i=1}^{n} f_{\hat{\theta}_0}\left(X_i\right)\right)\right] \\
&= 2n\left[LD\left(v_n, f_{\hat{\theta}_0}\right) - LD\left(v_n, f_{\hat{\theta}}\right)\right],
\end{aligned} \tag{38}
$$

where, $LD(\cdot,\cdot)$ stands for the likelihood disparity. Here $\hat{\theta}$ and $\hat{\theta}_0$ 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 LϕDPD.

As an analog of the likelihood ratio test, we consider the Divergence Difference Test (DDT) based on LϕDPD 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:

$$
DDT_{\beta,\gamma}\left(v_n\right) = 2n\left[d_{\beta,\gamma}\left(v_n, f_{\hat{\theta}_0}\right) - d_{\beta,\gamma}\left(v_n, f_{\hat{\theta}}\right)\right], \tag{39}
$$

$\hat{\theta}_0$ and $\hat{\theta}$ are MLϕDE under null hypothesis and unrestricted minimum MLϕDE respectively. Also note that:

$$
\begin{aligned}
&d_{\beta,\gamma}\left(v_n, f_\theta\right) \\
&= \sum_{x\in\mathfrak{X}}\left[B\left(v_n(x)\right) - B\left(f_\theta(x)\right) - \left(v_n(x) - f_\theta(x)\right)B'\left(f_\theta(x)\right)\right],
\end{aligned}
$$

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 Θ_0 is defined by a set of $r \le p$ restrictions on Θ defined by $R_i(\theta) = 0$, $1 \le i \le r$. We assume that the parameter space under H_0 can be described through a parameter $\xi = (\xi_1, \ldots, \xi_{p-r})$, with p-r independent components, i.e., H0 specifies that there exists a function b: $\mathbb{R}^{p-r} \to \mathbb{R}^p$ where $\theta = b(\gamma)$, $\gamma \in \Gamma \subseteq \mathbb{R}^{p-r}$. The function b is assumed to have continuous derivative $\dot{b}(\xi)$ of order $p \times (p\text{-}r)$ with rank p-r. Then the constrained estimator is $\hat{\theta}_0 = b\left(\hat{\xi}\right)$, where $\hat{\xi}$ is the MLϕDE under the ξ formulation of the model. Let $G = F_\theta$ be the true distribution which belongs to the family $\mathcal{F}$ with parameter θ. Under H_0, let ξ 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 ξ and θ respectively in the sense that:

$$
\begin{aligned}
\hat{\xi} &= \xi + n^{-1/2}\left[\dot{b}(\xi)^T J_B\left(b(\xi)\right)\dot{b}(\xi)\right]^{-1}\dot{b}(\xi)^T Z_n b(\xi) \\
&+ o_P\left(n^{-1/2}\right),
\end{aligned} \tag{40}
$$

where, $Z_n(b(\xi))$ is $AN(0, K_B(b(\xi)))$. Here $J_B(\cdot)$ and $K_B(\cdot)$ 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 LϕDPD estimator $\hat{\theta}$ satisfies:

$$
\hat{\theta} = \theta + n^{-1/2} J_B^{-1}(\theta) Z_n(\theta) + o_P\left(n^{-1/2}\right), \tag{41}
$$

  where, $Z_n(\theta)$ is $AN(0, K_B(\theta))$.
- (C3) The null hypothesis H_0 is either simple and $\Theta_0 = \{\theta_0\}$, where θ_0 is in the interior of Θ, or H_0 is composite and $\Theta_0 = \{b(\xi): \xi \in \Gamma \subseteq \mathbb{R}^{p-r}\}$
- (C4) If H_0 is composite then the constrained estimator $\hat{\theta}_0 = b\left(\hat{\xi}\right)$ and $\hat{\xi}$ satisfies Equation 40. Define:

$$
\Sigma_{B,b}(\theta,\xi) = \tilde{J}_{B,b}^{-1}(\theta,\xi) K_B(\theta) \tilde{J}_{B,b}^{-1}(\theta,\xi),
$$

  where

$$
\tilde{J}_{B,b}(\theta,\xi), = \left[J_B(\theta)^{-1} - \dot{b}(\xi)\left[\dot{b}(\xi)^T J_B(\theta)\dot{b}(\xi)\right]^{-1}\dot{b}(\xi)^T\right]^{-1}
$$

Theorem 8

Suppose that assumption (C1)-(C4) hold. Under f_{θ_0}, $\theta_0 \in \Theta_0$, the limiting distribution of the distance difference test statistic in Equation 39 coincides with the distribution of:

$$
\sum_{i=1}^{m} \lambda_i Z_i^2,
$$

where, λ_i's are non-zero eigenvalues of $A(q_0)\Sigma_{B,b}(\theta_0, \xi)$ and $m = \text{rank}(A(q_0)\Sigma_{B,b}(\theta_0, \xi))$. 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 $T_{\beta,\gamma}\left(\hat{\theta}, \theta_0\right)$ 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 $\Theta_0 \subseteq \Theta \subseteq \mathbb{R}^p$ is defined by a set of $r \leq p$ restrictions on Θ defined by $R_i(\theta) = 0$, $1 \leq i \leq r$. Let $G = F_\theta$ be the true distribution which belongs to the family $\mathcal{F}$ with parameter θ. Assume $\hat{\theta}$ to be the MLϕDE of the true parameter θ. Define $R(\theta) = (R_1(\theta),\ldots,R_r(\theta))^T$ and $D(\theta) = \left[\frac{\partial R_i(\theta)}{\partial \theta_j}\right]_{r\times p}$. Under the spirit of the original Wald test statistic, we can construct the following test statistic:

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

where, $\Sigma(\theta) = J_B(\theta)^{-1}K_B(\theta)J_B(\theta)^{-1}$ under the $B(\cdot)$ function described in Equation 16. Under standard regularity conditions it is easy to prove that the asymptotic distribution of $W\left(\hat{\theta}\right)$ is χ^2_r under the null hypothesis. The proof follows from simple application of delta method theorem on the quantity $R\left(\hat{\theta}\right)$ and the fact that under the null hypothesis $\sqrt{n}\left(\hat{\theta}-\theta\right)$ is $AN(0, \Sigma(\theta))$. 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:

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

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 $DDT_{\beta,\gamma}(v_n)$ in Equation 39. Here β and γ are chosen to be 0.3 and 0.05 respectively. The support of the distribution is $\mathcal{X} = (0,1)$, where the digit 1 stands for the death of a mosquito. From here it is evident that $v_n(1) = 264/465$. 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 parameteric 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 tesing using the LϕDPD, 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 Theorem 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 $f_\theta(x)\phi(f_\theta(x), \gamma)$ is continuous for $\gamma\in(0,1]$ and $\lim_{\gamma\to\infty} f_\theta(x)\phi(f_\theta(x), \gamma) = 1$. By applying dominated convergence theorem (DCT) on Equation 11 at $\gamma\to 0^+$ we get:

$$\frac{1}{n}\sum_{i=1}^{n} u_\theta\left(X_i\right) f_\theta^{\beta}\left(X_i\right) - \int u_\theta(x) f_\theta^{1+\beta}(x)dx = 0$$

which is the unbiased estimating equation for DPD with tuning parameter β. Hence the result follows.

(b) As $u_\theta(x)\ f_\theta(x)^{1+\beta}$, $u_\theta(x)\ u_\theta(x)^T\ f_\theta(x)^{1+2\beta}$, $u_\theta(x)u_\theta(x)^T\ f_\theta(x)^{1+\beta}$ are integrable and $f_\theta(x)\phi(f_\theta(x),\gamma)$ is bounded, by DCT on Equation 12 at $\gamma\to 0^+$ we get:

$$J_\beta = \int u_\theta u_\theta^T f_\theta^{1+\beta}, K_\beta = \int u_\theta u_\theta^T f_\theta^{1+2\beta} - \zeta_\beta\zeta_\beta^T,$$
$$\zeta_\beta = \int u_\theta f_\theta^{1+\beta},$$

i.e., $\lim_{\gamma\to 0^+} J_\phi^{-1}K_\phi J_\phi^{-1} = J_\beta^{-1}K_\beta J_\beta^{-1}$. We already know from the assumptions that $J_\beta^{-1}K_\beta J_\beta^{-1} \prec J_\alpha^{-1}K_\alpha J_\alpha^{-1}$, i.e., $\left(J_\alpha^{-1}K_\alpha J_\alpha^{-1} - J_\beta^{-1}K_\beta J_\beta^{-1}\right)$ is positive definite, where J_α and K_α are defined in the same fashion as J_β and K_β respectively. The inequality of the asymptotic variances is used here in the sense that AE of DPD with parameter β is greater than that of AE of DPD with parameter a. So there exists a $\gamma = \gamma(\alpha,\beta)$ such that $J_\phi^{-1}K_\phi J_\phi^{-1} \prec J_\alpha^{-1}K_\alpha J_\alpha^{-1}$. 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 $|\theta_n|$ as $n\to\infty$. Now, consider:

$$D\left(h_{\varepsilon,n}, f_{\theta_n}\right) = \int_{A_n} d\left(h_{\varepsilon,n}, f_{\theta_n}\right) + \int_{A_n^c} d\left(h_{\varepsilon,n}, f_{\theta_n}\right), \qquad (42)$$

where, $A_n = \{x: g(x) > \max\{k_n(x),\ f_{\theta_n}(x)\}\}$. Now since g belongs to the model family $\mathcal{F}$, from (BP1) it follows that $\int_{A_n} k_n(x)\to 0$ and from (BP2) we get $\int_{A_n} f_{\theta_n}\to 0$, thus under k_n and f_{θ_n}, the set A_n converges to a set of zero probability as $n\to\infty$. Thus, on A_n, $d(h_{\varepsilon,n})\to d((1-\varepsilon)g,0)$ as $n\to\infty$ and so by DCT $\left|\int_{A_n} d\left(h_{\varepsilon,n}, f_{\theta_n}\right) - \int_{A_n} d\left((1-\varepsilon)g,0\right)\right|\to 0$. Using (BP1), (BP2) and the above result, we have $\int_{A_n} d\left(h_{\varepsilon,n}, f_{\theta_n}\right)\to M^{(1)}_{f,(1-\varepsilon)}$. Next, by (BP1) and (BP2), $\int_{A_n^c} g\to 0$ as $n\to\infty$, so under g, the set A_n^c converges to a set of zero probability. Hence, similarly, we get $\left|\int_{A_n^c} d\left(h_{\varepsilon,n}, f_{\theta_n}\right) - \int_{A_n^c} d\left(\varepsilon k_n, f_{\theta_n}\right)\right|\to 0$. Now by (BP3), we have $\int d\left(\varepsilon k_n, f_{\theta_n}\right) \ge \int d\left(\varepsilon f_{\theta_n}, f_{\theta_n}\right) = M^{(1)}_{f,\varepsilon} - M^{(2)}_{f,(\varepsilon-1)}$. Thus combining the equations we get $\liminf_{n\to\infty} D\left(h_{\varepsilon,n}, f_{\theta_n}\right) \ge M^{(1)}_{f,\varepsilon} - M^{(2)}_{f,(\varepsilon-1)} + M^{(1)}_{f,(1-\varepsilon)} = a_1(\varepsilon)$, say.

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

$$\limsup_{n\to\infty} D\left(h_\varepsilon, f_{\theta_n}, f_{\theta_n}\right) < a_1(\varepsilon)$$

as then the sequence $\{\theta_n\}$ above could not minimize $D\left(h_{\varepsilon,f_{\theta_n}}, f_{\theta_n}\right)$ for every n. We will now show that above equation is true for all $\varepsilon < 1/2$ under the model when we choose $\theta^* = \theta^g$. For any fixed θ, let $B_n = \{x: k_n(x) > \max\{g(x), f_\theta(x)\}\}$. Since g belongs to the model $\mathcal{F}$, from (BP1) we get $\int_{B_n} g\to 0$, $\int_{B_n} f_\theta\to 0$ and $\int_{B_n^c} k_n\to 0$ as $n\to\infty$. Thus, under k_n, the set B_n^c converges to a set of zero probability, while under g and f_θ, the set B_n converges to a set of zero probability. Thus, on B_n, $d(h_{\varepsilon,n}, f_\theta)\to d(\varepsilon k_n,\ 0) = B(\varepsilon k_n)$ as $n\to\infty$. So by DCT $\left|\int_{B_n} d\left(h_{\varepsilon,n}, f_\theta\right) - \int B\left(\varepsilon k_n\right)\right|\to 0$. Similarly we have $\left|\int_{B_n^c} d\left(h_{\varepsilon,n}, f_\theta\right) - \int d\left((1-\varepsilon)g, f_\theta\right)\right|\to 0$. Therefore, we have:

$$\limsup_{n\to\infty} D\left(h_{\varepsilon,n}, f_\theta\right) = \int D\left((1-\varepsilon)g, f_\theta\right) + \limsup_{n\to\infty}\int B\left(\varepsilon k_n\right). \quad (43)$$

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

$$\limsup_{n\to\infty} D\left(h_{\varepsilon,n}, f_{\theta^g}\right) \le M^{(1)}_{f,(1-\varepsilon)} - M^{(2)}_{f,(-\varepsilon)} + M^{(1)}_{f,\varepsilon} = a_3(\varepsilon),$$

say. Consequently, asymptotically there is no breakdown if for ε level contamination when $a_3(\varepsilon) < a_1(\varepsilon)$. But, notice $a_1(\varepsilon)$ and $a_3(\varepsilon)$ are strictly decreasing and increasing functions respectively. To see this for $a_1(\varepsilon)$, notice $M^{(1)}_{f,\varepsilon} - M^{(2)}_{f,(\varepsilon-1)}$. As $\varepsilon\uparrow 1$ the above expression decreases. $M^{(1)}_{f,(1-\varepsilon)} = \int B\left((1-\varepsilon) f\right)$. From Equation 16 we see that $B(\cdot)$ is an increasing function on positive half line. Using this it is evident that $M^{(1)}_{f,(1-\varepsilon)}$ decreases as $\varepsilon\uparrow 1$. So, $a_1(\varepsilon)$ decreases as $\varepsilon\uparrow 1$. Similarly

it can be shown $a_3(\varepsilon)$ is an increasing function of e. But $a_1(1/2) = a_3(1/2)$; thus asymptotically there is no breakdown and $\limsup_{n\to\infty} |T_{\beta,\gamma}(H_{\varepsilon,n})| < \infty$ for $\varepsilon < 1/2$. Hence the theorem follows.

Proof of Theorem 7

Proof

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

$$\frac{1}{n}\sum_{i=1}^{n} u_\theta\left(X_i\right) B''\left(f_\theta\left(X_i\right)\right) f_\theta\left(X_i\right) - \int u_\theta(x) B''\left(f_\theta(x)\right) f_\theta^2(x)\,dx = 0$$

or equivalently:

$$\frac{1}{n}\sum_{i=1}^{n}\left(u_\theta\left(X_i\right) B''\left(f_\theta\left(X_i\right)\right) f_\theta\left(X_i\right) - \int u_\theta(x) B''\left(f_\theta(x)\right) f_\theta^2(x)\,dx\right) = 0$$

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

$$U_\theta\left(X_i\right) = u_\theta\left(X_i\right) B''\left(f_\theta\left(X_i\right)\right) f_\theta\left(X_i\right) - \int u_\theta(x) B''\left(f_\theta(x)\right) f_\theta^2(x)\,dx.$$

We have already seen that the statistic $T_{\beta,\gamma}\left(\hat{\theta},\theta_0\right)$ satisfies:

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

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

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} U_{\hat{\theta}}\left(X_i\right) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} U_{\theta_0}\left(X_i\right) + \sqrt{n}\left(\hat{\theta}-\theta_0\right)\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta} U_\theta\left(X_i\right)|_{\theta=\theta_0} + \sqrt{n}\left(\hat{\theta}-\theta_0\right)^2\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2} U_\theta\left(X_i\right)|_{\theta=\theta'}$$

for some θ' in between θ_0 and $\hat{\theta}$. So we have:

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} U_{\hat{\theta}}\left(X_i\right) = \sqrt{n}\bar{U}\left(\theta_0\right) + \sqrt{n}\left(\hat{\theta}-\theta_0\right)\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta} U_\theta\left(X_i\right)|_{\theta=\theta_0} + o_p(1).$$

And hence:

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

Note $\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta} U_\theta\left(X_i\right)|_{\theta=\theta_0} \to E_{\theta_0}\frac{\partial}{\partial\theta} U_\theta\left(X_1\right)|_{\theta=\theta_0} = -J_B(\theta_0)$ as $n\to\infty$. Hence:

$$\sqrt{n}\bar{U}\left(\theta_0\right) = \sqrt{n}\left(\hat{\theta}-\theta_0\right) J_B\left(\theta_0\right) + o_p(1)$$

So:

$$n\left(\hat{\theta}-\theta_0\right)^T A\left(\theta_0\right)\left(\hat{\theta}-\theta_0\right) = n\bar{U}\left(\theta_0\right) J_B^{-1}\left(\theta_0\right) A\left(\theta_0\right) J_B^{-1}\left(\theta_0\right)\bar{U}\left(\theta_0\right) + o_p(1).$$

This completes the proof.

Proof of Theorem 8

Proof

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

$$\begin{aligned} &DDT_{\beta,\gamma}\left(v_n\right) \\ &= 2n\left[d_{\beta,\gamma}\left(v_n, f_{\hat{\theta}_0}\right) - d_{\beta,\gamma}\left(v_n, f_{\hat{\theta}}\right)\right] \\ &= 2n\left[\sum_j\left(\hat{\theta}_{0j}-\hat{\theta}_j\right)\nabla_j d_{\beta,\gamma}\left(v_n, f_\theta\right)|_{\theta=\hat{\theta}}\right. \\ &\left.+\frac{1}{2}\sum_{j,k}\left(\hat{\theta}_{0j}-\hat{\theta}_j\right)\left(\hat{\theta}_{0k}-\hat{\theta}_k\right)\nabla_{jk} d_{\beta,\gamma}\left(v_n, f_\theta\right)|_{\theta=\theta*}\right] \end{aligned} \tag{44}$$

where the subscripts denote the indicated components of the vector. Also θ^* lies in the line segment joining $\hat{\theta}_0$ and $\hat{\theta}$. By definition, $\nabla_j d_{\beta,\gamma}\left(v_n, f_\theta\right)|_{\theta=\hat{\theta}} = 0$. Hence, the Equation 44 reduces to:

$$\begin{aligned} &DDT_{\beta,\gamma}\left(v_n\right) \\ &= n\left(\hat{\theta}_0-\hat{\theta}\right)^T A\left(\theta_0\right)\left(\hat{\theta}_0-\hat{\theta}\right) \\ &+ n\left(\hat{\theta}_0-\hat{\theta}\right)\left[\nabla_2 d_{\beta,\gamma}\left(v_n, f_{\theta^*}\right) - A\left(\theta_0\right)\right]\left(\hat{\theta}_0-\hat{\theta}\right) \end{aligned} \tag{45}$$

We will show that under the null $\nabla_2 d_{\beta,\gamma}\left(v_n, f_{\theta^*}\right) - A\left(\theta_0\right)$ as $n\to\infty$. By another Taylor

expansion around the true value θ_0, we get for some θ^{**} between θ_0 and θ^*:

$$\begin{aligned}&\nabla_{jk} d_{\beta,\gamma}\left(v_n, f_{\theta^*}\right)\\ &\nabla_{jk} d_{\beta,\gamma}\left(v_n, f_{\theta_0}\right)+\sum_l\left(\theta_l^*-\theta_{0l}\right)\nabla_{jkl} d_{\beta,\gamma}\left(v_n, f_{\theta^{**}}\right).\end{aligned} \tag{46}$$

Under the assumptions (C1)-(C4) it can be easily shown that $\nabla_2 d_{\beta,\gamma}\left(v_n, f_{\theta_0}\right) \to A\left(\theta_0\right)$ as $n \to \infty$ and $\nabla_{jkl} d_{\beta,\gamma}\left(v_n, f_{\theta^{**}}\right)=O_P(1)$. By a simple application of delta theorem on Equation 40 it can be shown $\sqrt{n}\left(\hat{\theta}_0-\theta_0\right)=O_P(1)$ under the null hypothesis. Equation 41 yields that $\sqrt{n}\left(\hat{\theta}_0-\theta_0\right)=O_P(1)$. Hence we have $\theta^* = \theta_0 + o_P(1)$. As a result the Equation 46 reduces to $\nabla_2 d_{\beta,\gamma}\left(v_n, f_{\theta^*}\right) = A(\theta_0) + o_P(1)$. So, the Equation 45 becomes:

$$DDT_{\beta,\gamma}\left(v_n\right)=n\left(\hat{\theta}_0-\hat{\theta}\right)^T A\left(\theta_0\right)\left(\hat{\theta}_0-\hat{\theta}\right)+o_P(1) \tag{47}$$

To obtain the asymptotic null distribution of $DDT_{\beta,\gamma}(v_n)$ it is enough to obtain the asymptotic null distribution of $\sqrt{n}\left(\hat{\theta}_0-\hat{\theta}\right)$. Again from Equation 41 and by simple application delta theorem on Equation 40 it is easy to show that:

$$\sqrt{n}\left(\hat{\theta}_0-\hat{\theta}\right) \xrightarrow{w} N\left(0, \Sigma_{B,b}\left(\theta_0, \xi_0\right)\right),$$

where, ξ_0 is the true value of the parameter under ξ formulation of the model. Hence the result follows. If $\Theta_0 = \{\theta_0\}$, then Equation 47 reduces to:

$$DDT_{\beta,\gamma}\left(v_n\right)=n\left(\theta_0-\hat{\theta}\right)^T A\left(\theta_0\right)\left(\theta_0-\hat{\theta}\right)+o_P(1).$$

We also know:

$$T_{\beta,\gamma}\left(\hat{\theta}, \theta_0\right)=n\left(\theta_0-\hat{\theta}\right)^T A\left(\theta_0\right)\left(\theta_0-\hat{\theta}\right)+o_P(1).$$

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.