The Density Power Divergence: Some recent extensions

Anirban Nath
Indian Statistical Institute, Kolkata

Supervisor: Ayanendranath Basu
Interdisciplinary Statistical Research Unit,
203 Barrackpore Trunk Rd, Kolkata, WB 700108, India.

February 21, 2022

Abstract

In robust estimation, density power divergence (DPD) and logarithmic density power divergence (LDPD) holds a significant position due to their significant robustness properties at the cost of very minimal efficiency. LDPD, being the normalised version of DPD also has a greater control over bias in heavy contamination. But the former sometimes shows a spurious minimum for some particular occurrence of data points that lead to multiple roots of the estimating equation hindering the standard theories of estimation procedures. In this dissertation, we want to propose some new estimators that will be free of this spurious behaviour but retaining those desired properties at the same time.

Keywords: DPD, LDPD, Divergence, Bregman Divergence, Asymptotic distribution, Influence function.

1 Introduction

A statistical inference procedure is based on two things. One is a set of observations representing the data; the other is the set of underlying assumptions – either implicit or explicit – on the data which represent the premises on which the analysis is being done. Classical parametric inference is based on maximum likelihood, which performs best when all these underlying assumptions hold perfectly. In reality, however, these assumptions cannot hold in a perfect sense. Even when they are reasonable, they are only met in an approximate sense and small deviations can never be totally eliminated. The assumption that all the data points follow the model strictly is a very vulnerable assumption. There are often data points that lie outside of the overall pattern of a distribution [1] i.e. it is distant from the majority of the other observations [2]. These points may be called “outliers”, at least in a geometric sense. They lead to inefficient and highly unstable performance of the classical techniques resulting in inaccurate statistical inference. The initial attempts to overcome these inefficient inference procedure consisted of the subjective rejection of these so called “bad” data points. However, with time, researchers have come to realize that these outliers often contain valuable information about the system and merit further scrutiny, rather than being subjectively deleted from the data. Moreover, detecting these outliers in the present era of big and high-dimensional data may be extremely difficult on the basis of traditional methods alone. So the inference procedures must be equipped with suitable robustness properties which will control the influence of these data points on the inference decision.

The word ‘robust’ is loaded with many inconsistent connotations, but we will process with the idea of ’robustness’ given in [3] which is, “insensitivity to small deviations from the assumptions”. The robust approach to statistical modelling and data analysis aims at introducing techniques which produce stable statistics leading to reliable parameter estimates, associated tests and confidence intervals, not only when the data follow a given distribution exactly, but also when there are mild violations to the parametric assumptions or contamination is present in the data. However, the utility of robust methods will be limited if they come at a cost of a high efficiency loss at the model. So the current line of research in robust statistics wants to come up with robust inference procedures that lead to minimal compromise in model efficiency.

2 The Minimum distance approach

In our work, we shall focus on the minimum distance approach to robust inference which was pioneered by [4], who described the desirable properties of this method under suitable conditions. The idea behind this procedure is the quantification of a measure of discrepancy between the data and the model. Donoho and Liu [5] later showed that this approach sometimes has an important role in generating inference procedures with natural robustness properties. This quantification of the amount of discrepancy between data and the model is usually done through an appropriate divergence or statistical distance. The distance can be constructed both between two distribution functions or two densities. It is important to make it clear at this stage that many of the density-based divergences (and some of the other divergences as well) that are utilized for different purposes in the statistical literature are not mathematical distances in the sense of being metrics. Most of them are not symmetric in their arguments. This is not very important for statistical purposes.

In fact, in many cases it is the asymmetry in the structure of these divergences which has a major role in imparting some of the desirable properties to the estimators generated by them.

Definition 2.1. A divergence is a measure of distance that satisfies the following two properties -

- The measure should be nonnegative.
- It equals to zero if and only if the data match the model exactly (i.e., the two arguments of the measure are identically equal).

Any divergence which satisfies the above two properties will be referred to here as a "statistical distance."

For example, we are familiar with some of the popular distances based on the distribution functions of the data and the model. These include

1. **Weighted Kolmogorov-Smirnov distance**: The Weighted K-S distance between the empirical distribution function of the data G_n and the distribution function of the model F_θ is given by -
$$\rho_{KS}(G_n, F_\theta) = \sup_{-\infty<z<\infty} |G_n(z) - F_\theta(z)| \sqrt{\psi(F_\theta(z))} \tag{2.1}$$
where, $\psi(u) = 1$ gives the usual Kolmogorov–Smirnov distance measure. A minimum distance estimator corresponding to the Kolmogorov–Smirnov distance (or the weighted Kolmogorov–Smirnov distance if appropriate) can be obtained by minimizing the above distance over the parameter space Θ.

2. **Weighted Cramér–von Mises distance** The Weighted C-M distance between the empirical distribution function of the data G_n and the distribution function of the model F_θ is given by -
$$\rho_{CM}(G_n, F_\theta) = \int_{-\infty}^{\infty} (G_n(z) - F_\theta(z))^2 \psi(F_\theta(z)) dF_\theta(z) \tag{2.2}$$
where, $\psi(u) = 1$ gives the usual Cramér–von Mises distance measure.

3. **Anderson-Darling measure** If $\psi(u) = [u(1-u)]^{-1}$ in (2.2), the measure generated is known as Anderson-Darling distance.

3 Density based distances

One of the important members of the class of density-based distances is the family of chi-square type distances, generally called ϕ-divergences. All the minimum distance procedures based on these distances generate estimators which are asymptotically fully efficient and many of them have remarkably strong robustness properties. First, we introduce them for discrete models. This class is also known as the class of "disparity" measures, where the term "disparity" is used exchangeably with "distances".

Let $X_1, X_2, \ldots, X_n$ represent a sequence of independent and identically distributed observations from a distribution G, with support $\mathcal{X} = \{0, 1, 2, \ldots\}$, having a probability density function g with respect to the counting measure. Let $d_n(x)$ represent the relative frequency of the value x in the random sample. We will denote by $\mathcal{G}$ the class of all distributions having densities with respect to the counting measure (or the appropriate dominating measure in other cases), and we will assume this class to be convex. We will also assume that both G and F belong to $\mathcal{G}$.

Definition 3.1. The disparity between $\boldsymbol{d}$ and $\boldsymbol{f_\theta}$ generated by C is given by

$$\rho_C(d_n, f_\theta) = \sum_{x=0}^{\infty} C(\delta(x)) f_\theta(x) \tag{3.1}$$

where C is a thrice differentiable strictly convex function satisfying $C(0) = 0$ and $\delta(x)$ is the Pearson residual at value x defined as

$$\delta(x) = \frac{d_n(x)}{f_\theta(x)} - 1.$$

In this setup, the minimum distance estimator $\hat{\theta}$ of θ based on the disparity ρ_C will be defined by

$$\rho_C(d_n, f_{\hat{\theta}}) = \min_{\theta \in \Theta} \rho_C(d_n, f_\theta).$$

Different choices of C generates different disparity measures. Some of the well known disparities are listed below.

1. **Likelihood disparity** (LD): If $C(\delta) = (\delta + 1)\log(\delta + 1) - \delta$, we get the likelihood disparity

   $$\mathrm{LD}(d_n, f_\theta) = \sum d_n \log\left(\frac{d_n}{f_\theta}\right).$$

   It is important to note that, the minimizer of the likelihood disparity is actually the maximum likelihood estimator (MLE). Hence, the MLE is a member of the class of Minimum Disparity Estimators (MDE).

2. **Kullback-Leibler divergence**: If $C(\delta) = \delta - \log(\delta + 1)$, we get the Kullback-Leibler divergence

   $$\mathrm{KLD}(d_n, f_\theta) = \sum f_\theta \log\left(\frac{f_\theta}{d_n}\right).$$

This divergence is the symmetric opposite of LD (which is itself a version of the Kullback-Leibler divergence).

3. **Hellinger distance** (HD): The (twice, squared) Hellinger distance is generated by the function $C(\delta) = 2(\sqrt{\delta+1}-1)^2$ and is defined as

$$\mathrm{HD}(d_n, f_\theta) = 2\sum(\sqrt{d_n}-\sqrt{f_\theta})^2.$$

4. **Power divergence** (PD): A very important subfamily of disparities is the PD family introduced by Cressie and Read (1984) [6] as

$$\mathrm{PD}_\lambda(d_n, f_\theta) = \frac{1}{\lambda(1+\lambda)}\sum d_n \left[\left(\frac{d_n}{f_\theta}^\lambda\right) - 1\right]$$

which has a tuning parameter $\lambda \in \mathbb{R}$. It is important to note that, for $\lambda = 1, -\frac{1}{2}, -2$, the family coincides with Pearson chi-square, the Hellinger distance and Neyman modified chi-square, respectively. Also, it generates LD and KLD for the limiting cases $\lambda \to 0$ and $\lambda \to -1$, respectively.

But things take a turn for the continuous models. Here constructing the disparity (or divergence in general) is difficult, as the data are discrete and the model is continuous, so there is an obvious mismatch of measures. Instead of discretizing the model, Beran (1977) [7] proposed to construct a continuous density estimate using some appropriate nonparametric density estimation method such as kernel density estimation. In this case, let

$$g_n^*(x) = \frac{1}{n}\sum_{i=1}^{n} K(x, X_i, h_n) = \int K(x, y, h_n)dG_n(y)$$

denote a nonparametric kernel density estimator where $K(x, y, h_n)$ is a smooth kernel function with bandwidth h_n and G_n is the empirical distribution function obtained from the data. We can then estimate θ by minimizing the disparity

$$\rho_C(g_n^*, f_\theta) = \int C(\delta(x))f_\theta(x) \tag{3.2}$$

where the Person residual now equals

$$\delta(x) = \frac{g_n^*(x)}{f_\theta(x)} - 1.$$

In practice, however, the addition of the kernel density estimation process leads to substantial difficulties. The theoretical derivation of the asymptotic normality of the minimum distance estimators based on disparities and the description of their other asymptotic properties are far more complex in this case. The kernel smoothing introduces a bias in the density estimate, which has to be asymptotically corrected by choosing the bandwidth $h = h_n$ to be a function of the sample size and letting it slide to zero at the appropriate rate with increasing n. This smoothing component, which is an intermediate step in our estimation scheme, adds an additional layer of theoretical complexity to the procedure, as the choice of the bandwidth now becomes crucial.

Although under suitable conditions, the minimum distance estimators based on disparities continue to be first order efficient at the model, it seems to be difficult to achieve a completely general structure where the theory flows freely for the whole class of minimum distance estimators for the continuous case for all disparities under some general conditions. Park and Basu (2004) [8, 9] has proved the asymptotic normality for the continuous models under a set of strong conditions. However the conditions are not satisfied by several common disparities. Hence, we follow another approach proposed by Basu and Lindsay(1994) [10]. Instead of smoothing the data alone, this approach suggests the convolution of the model density with the same kernel as well. The rationale behind this proposal is that it introduces the same bias in th e data and the model, which can then compensate each other. Hence, the importance of the kernel is diminished in this estimation procedure as compared to Beran's approach.

For a suitable kernel function $K(x, y, h)$, let $g_n^*(x)$ be the kernel density estimate obtained from the data, and f_θ^* be the kernel smoothed model density of the model defined as

$$g_n^*(x) = \int K(x, y, h) dG_n(y)$$

and

$$f_\theta^*(x) = \int K(x, y, h) dF_\theta(y).$$

A typical disparity based on a disparity generating function C can now be constructed as

$$\rho_C(g_n^*, f_\theta^*) = \int C(\delta(x)) f_\theta^*(x) \tag{3.3}$$

where the Person residual now equals

$$\delta(x) = \frac{g_n^*(x)}{f_\theta^*(x)} - 1.$$

It is obvious that we have fixed h consistency for the minimum disparity estimator under this approach.

4 Escaping the curse of kernel smoothing

Although there are several benefits of minimum distance estimation based on disparities, at the same time, we have pointed out that in case of continuous models one is forced to use some form of nonparametric smoothing such as kernel density estimation to produce a continuous estimate of the true density. As a result, the minimum distance estimation method based on disparities inherits all the associated complications such as those related to bandwidth selection in continuous models. Although the Basu and Lindsay procedure largely reduces the effect of the bandwidth, the process still involves a nonparametric smoothing component.

To get past this, we need to develop some inference procedure where we can approximate the density of the data without using the kernel density estimator but still possesses high asymptotic efficiency. Basu et al. (1998) [11] developed a class of divergences called Density Power Divergence (DPD) by emphasizing on the pattern of density power downweighting of the score function. Before going into the details, we provide the motivation behind this inference procedure.

4.1 Minimum $\mathcal{L}_2$ distance estimator

Consider a parametric family of distributions $\mathcal{F}$ as $\mathcal{F} := \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$. Let $\{f_\theta\}$ be the associated class of model densities with respect to an appropriate dominating measure, and let G be the class of all distributions having densities with respect to the given measure. The $\mathcal{L}_2$ distance between the densities g and f_θ is given by $[\int\{g(x) - f_\theta(x)\}^2 dx]^{\frac{1}{2}}$. The minimum $\mathcal{L}_2$ distance functional $T(G)$ is defined, for every G in $\mathcal{G}$, by the relation,

$$\int \{g(x) - f_{T(G)}(x)\}^2 dx = \min_{\theta\in\Theta} \int \{g(x) - f_\theta(x)\}^2 dx \tag{4.1}$$

where g is the density of G. Note that given a distribution $G \in \mathcal{G}$, the squared $\mathcal{L}_2$ distance between the corresponding density g and f_θ can be represented as

$$\int f_\theta^2(x)dx - 2 \int f(x)dG(x) + K$$

where K is a constant independent of θ and can be removed from the objective function as it does not contribute to the minimization procedure. Notice that once the constant term is removed, the true density g appears linearly in this objective function (only through the $dG(x)$ term), unlike the disparities discussed earlier. Thus, given a random sample $X_1, X_2, \ldots, X_n$ from the true density g, one can actually minimize

$$\int f_\theta^2(x)dx - 2 \int f(x)dG_n(x) = \int f_\theta^2(x)dx - 2\, n^{-1} \sum_{i=1}^{n} f_\theta(X_i) \tag{4.2}$$

with respect to θ, where G_n is the empirical distribution function. (This may be viewed as replacing the population mean with the sample mean to get an empirical estimate of the unknown objective function). Remarkably, one does not need a smooth nonparametric estimate of g for this inference process, in contrast to the disparity based methods. The estimating equation for this divergence will be

$$n^{-1} \sum_i u_\theta(X_i)\ f_\theta(X_i) - \int\ u_\theta(x) f_\theta^2(x) dx = 0. \tag{4.3}$$

Now, consider a location model $\{\mathcal{F}_\theta\}$, and let θ be the location parameter. Notice that for this model, $\int f_\theta^2(x)dx$ the first term in (4.2), is independent of θ and hence does not affect the minimization process. The minimum $\mathcal{L}_2$ distance estimator is now simply the maximizer of $\sum_{i=1}^n f_\theta(Xi)$; this provides an interesting contrast with the maximum likelihood estimator, which maximizes the product of the model densities $\prod_{i=1}^n f_\theta(X_i)$ rather than their sum.

For comparison we present below the estimating equations of the minimum $\mathcal{L}_2$ distance estimator and maximum likelihood estimator in this case, which are

$$\sum_i u_\theta(X_i)\ f_\theta(X_i) = 0 \quad \text{and} \quad \sum_i u_\theta(X_i) = 0, \tag{4.4}$$

respectively. Notice that the score function is weighted by the model density in the estimating equation of the minimum $\mathcal{L}_2$ distance estimator, which may be viewed as a form of weighted likelihood equation. This provides an automatic downweighting for the scores $u_\theta(\cdot)$ of observations that are unlikely under the model. Such downweighting is not available under the ordinary likelihood approach,

4.2 Density Power Divergence

The idea of the density power divergence (DPD) comes from the link between the estimating equations presented in (4.4). Both of them belong to a class of generalised estimating equations

$$\sum_i u_\theta(X_i)\ f_\theta^\alpha(X_i) = 0 \quad \alpha \in [0, 1]. \tag{4.5}$$

The aforementioned estimating equations can be recovered for $\alpha = 0, 1$ and intermediate values of α provide a smooth bridge between these two estimating equations, and the degree of downweighting increases with increasing α and hence the degree of robustness and outlier stability increases. However, larger values of α are also associated with more inefficient estimators.

Using the same generalization for models beyond location models, the estimating equation in (4.3) can be generalized to the following form

$$\frac{1}{n}\sum_i u_\theta(X_i)\ f_\theta^\alpha(X_i) - \int u_\theta(x) f_\theta^{1+\alpha}(x)dx = 0, \quad \alpha \geq 0 \tag{4.6}$$

which is an unbiased estimating equation when the true distribution G belongs to the model $\{F_\theta\}$. From efficiency considerations one would rarely use a distance with $\alpha > 1$ for estimating the unknown model parameter, even though such estimators would have solid robustness properties. From this estimating equation, we can retrieve the underlying divergence by integrating the estimating equation which will result to the DPD class with tuning parameter α.

Definition 4.1. Given two densities g and f, the density power divergence $d_\alpha(g,f)$ corresponding to index $\alpha \in [0,1]$ between g and f is defined to be

$$d_\alpha(g,f) = \int \left\{ f^{1+\alpha}(x) - \left(1 + \frac{1}{\alpha}\right) g(x) f^\alpha(x) + \frac{1}{\alpha} g^{1+\alpha}(x) \right\} dx. \tag{4.7}$$

When $\alpha \to 0$, we recover the LD from the above form. We can define it as

$$d_0(g,f) = \lim_{\alpha \to 0} d_\alpha(g,f).$$

It is remarkable that $d_0(g,f)$ is the only common member between the class of disparities and the DPD family.

4.3 Bregman Divergence and related inference

Let $B : \mathbb{R} \to \mathbb{R}$ be a twice continuously differentiable, strictly convex function defined on a closed convex set in $\mathbb{R}$. For two given density functions g and f, the Bregman divergence between them (generated by the function B) is defined as

$$D_B(g,f) = \int \left\{ B(g(x)) - B(f(x)) - (g(x) - f(x))B'(f(x)) \right\} dx. \tag{4.8}$$

The function B, in the above case, is called the Bregman function. This function is clearly not uniquely defined due to the linearity property of the integral, as both $B(y)$ and $B(y) + ay + b$ give rise to the exact same divergence for any real constants a and b.

To go through the estimation procedure, the most important thing about this divergence is to note that as in the case of the DPD, here also the true data density $g(x)$ occurs linearly in the divergence. (The DPD family is, in fact, a subclass of Bregman divergences). Hence, we can empirically estimate that term from the data through empirical distribution.

Suppose that we have an independent and identically distributed sample $X_1, X_2, \ldots, X_n$ from the true distribution G, and we try to model this distribution by a parametric family $\mathcal{F} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ where θ is unknown but the functional form of F_θ is known to us. If we wish to use the minimum Brègman divergence approach for the estimation of the unknown parameter θ, the functional $T(G)$ will be defined as

$$D_B(g, f_{T(G)}) = \arg\min_{\theta\in\Theta} D_B(g, f_\theta).$$

After discarding the terms independent of θ, we observe that the only term involving true data density is $\int B'(f_\theta(x))g(x)dx$, which can be estimated by the corresponding sample mean $\frac{1}{n}\sum B'(f_\theta(X_i))$. Hence, the empirical objective function for the minimization of $D_B(g, f_\theta)$ is now given by

$$\int_x \left\{B'\left(f_\theta(x)\right) f_\theta(x) - B\left(f_\theta(x)\right)\right\} dx - \frac{1}{n}\sum_{i=1}^{n} B'\left(f_\theta\left(X_i\right)\right) \tag{4.9}$$

Under appropriate differentiability conditions, the minimizer of this empirical divergence over $\theta \in \Theta$ is obtained as a solution to the estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} u_\theta\left(X_i\right) B''\left(f_\theta\left(X_i\right)\right) f_\theta\left(X_i\right) - \int_x u_\theta(x) B''\left(f_\theta(x)\right) f_\theta^2(x) dx = 0. \tag{4.10}$$

This may be viewed as being in the general weighted likelihood equation form given by

$$\frac{1}{n}\sum_{i=1}^{n} u_\theta\left(X_i\right) w_\theta\left(X_i\right) - \int_x u_\theta(x) w_\theta(x) f_\theta(x) dx = 0 \tag{4.11}$$

where the relation between the Bregman function B and the weight function w_θ is given as

$$w_\theta(x) = w\left(f_\theta(x)\right) = B''\left(f_\theta(x)\right) f_\theta(x)$$

The non-negativity of the above weight function is secured by the convexity of the B function with the non-negativity of the density function. Also, it is easy to check that the estimating equation in (4.10) is unbiased and hence the asymptotic properties can be obtained through the theory for M-estimators.

The Bregman family some of the well known divergences for some particular choices of the Brègman function $B(\cdot)$. For example,

1. $B(y) = y\log(y) - y$: This generates the Kullback-Leibler divergence given by

$$D_{KL}\left(g, f_\theta\right) = \int_x g(x) \log\left(\frac{g(x)}{f_\theta(x)}\right) dx. \tag{4.12}$$

   Under our parametric setup, its estimating equation and weight function are, respectively,

$$\frac{1}{n}\sum_{i=1}^{n} u_\theta\left(X_i\right) - \int_x u_\theta(x) f_\theta(x) dx = 0, \quad w_\theta(x) = w\left(f_\theta(x)\right) = 1.$$

2. $B(y) = y^2$: This leads to the squared L_2 distance

$$L_2(g, f_\theta) = \int_x [g(x) - f_\theta(x)]^2 \, dx$$

generating, respectively, estimating equation and weight function as

$$\frac{1}{n}\sum_{i=1}^{n} u_\theta(X_i) f_\theta(X_i) = \int_x u_\theta(x) f_\theta^2(x) dx, \quad w_\theta(x) = w(f_\theta(x)) = f_\theta(x)$$

3. $B(y) = \left(y^{1+\alpha} - 1\right)/\alpha$: This generates the DPD (α) family given by

$$d_\alpha(g, f_\theta) = \int_x \left\{ f_\theta^{1+\alpha}(x) - \left(1 + \frac{1}{\alpha}\right) g(x) f_\theta^\alpha(x) + \frac{1}{\alpha} g^{1+\alpha}(x) \right\} dx. \tag{4.13}$$

In this case its estimating equation and weight function are given by

$$\frac{1}{n}\sum_{i=1}^{n} u_\theta(X_i) f_\theta^\alpha(X_i) - \int_x u_\theta(x) f_\theta^{1+\alpha}(x) dx = 0, \quad w_\theta(x) = w(f_\theta(x)) = f_\theta^\alpha(x).$$

It may be noted that the estimating equations are all unbiased under the model and have the same general structure as given in Equation (4.11). The equations differ only in the form of the weight function $w_\theta(x)$. And it is this weight function which determines to what extent the estimating equation is able to control the contribution of the score to the equation.

4.4 LDPD and its advantages over DPD

Windham (1995) [12] considered weighting the data using weights "proportional to a power of the density," and constructed weighted moment equations. This is similar to the philosophy of the DPD, although Windham (1995) [12] did not define a divergence or view this as a minimum distance problem. It turns out, however, if we utilize the likelihood score function as the choice of the moment functional, this approach reduces to choosing the estimator $\hat{\theta}$ as the solution of the equation

$$\frac{\sum_{i=1}^{n} f_\theta^\beta(X_i) u_\theta(X_i)}{\sum_{i=1}^{n} f_\theta^\beta(X_i)} = \frac{\int f_\theta^{1+\beta} u_\theta dx}{\int f_\theta^{1+\beta} dx}. \tag{4.14}$$

in θ. However, the above estimating equation is nothing but a normalized version of the estimating equation (4.6) of the minimum DPD estimator. By viewing equation (4.14) as a differential equation and integrating we recover another popular family of divergences – the logarithmic density power divergence (LDPD) – as the function of a nonnegative tuning parameter β.

Definition 4.2. Given two densities g and f, the logarithmic density power divergence $d_\beta(g,f)$ corresponding to index β between the densities g and f is defined to be

$$d_\beta(g,f) = \log\left(\int f^{1+\alpha}dx\right) - \left(1+\frac{1}{\alpha}\right)\log\left(\int gf^{\alpha}dx\right) + \frac{1}{\alpha}\log\left(\int g^{1+\alpha}dx\right). \tag{4.15}$$

As in the case of the DPD, for the minimum LDPD estimator also, robustness and outlier stability increases with β, while model efficiency drops with increasing β. Also note that both the DPD and the LDPD belong to the generalized class of divergences having the form

$$\psi\left(\int f^{1+\alpha}dx\right) - \left(1+\frac{1}{\alpha}\right)\psi\left(\int gf^{\alpha}dx\right) + \frac{1}{\alpha}\psi\left(\int g^{1+\alpha}dx\right).$$

We recover the DPD from the above when ψ is the identity function and LDPD when ψ is the log function.

There is some debate in the literature as to which one (among the minimum DPD estimator and the minimum LDPD estimator) should be the generally preferred estimator. [13] and [14] argue that the LDPD based estimators fare much better in keeping the latent bias of the estimator within control – in fact very close to zero – under heavy contamination. In particular, Fujisawa argue that this is a general property of normalized estimating equations. (Recall that the estimating equation of the LDPD is a normalized version of the same for the DPD; the latter does not have normalized estimating equations).

On the other hand, some counter-views are provided by [15]. These authors demonstrate that while there is some truth in the assertions of [13] and [14], the minimum LDPD estimator is vulnerable to a *spurious behaviour of the root* under *inner contamination.* In some situations the minimum LDPD objective function demonstrate a strange behaviour where there is a reasonable local minimum, but the global minimum represents a spurious root (nonsensical value) of the estimating equation. There is reason to suspect that the algorithms for computing the minimum LDPD estimator often leads to the local minimum. While this is a reasonable solution, there is no theory to suggest that this should be selected over the global minimum (although the latter returns a meaningless value). We discuss this phenomenon in the next section.

4.5 Spurious behaviour under inner contamination

Definition 4.3. We generally consider two kinds of contamination for the parametric model. The first case will correspond to the approximate singularity idea of the gross error model where the contaminating (minor) component is well-separated from the target (major) component. We will refer to this as *outer contamination*; in this case the contaminating values will be *surprising observations* in the sense of Lindsay (1994). In the second case of contamination, we will choose the contaminating component near the mode of the major component so that these observations are no longer surprising observations but will nevertheless distort the shape of the distribution relative to the parametric model. We will refer to this case as *inner contamination*.

[15] has shown that such spurious behaviour can also be observed under pure data although it is relatively rare in scale models. An actual random sample of size 20 from $\mathcal{N}(0,1)$ with seed 129 was obtained in R as

$$\begin{array}{rrrrrr}
-1.120900, & -0.989724, & -1.374697, & -1.355645, & 1.996755, & 0.695870 \\
0.771968, & -0.002847, & 1.008549, & -0.990280, & 1.131772, & -0.244929, \\
1.186625, & -1.671537, & -0.081999, & -1.831365, & 0.358867, & 0.891639, \\
0.489801, & 0.000010. & & & &
\end{array}$$

The exact value of the last observation is 1.048653×10^{-5}, which forces a spurious behaviour in the LDPD objective function plotted in figure 1 for $\alpha = 0.8$. where X_{20} is the last observation which is close to zero and $n = 20$. Figure 1 shows the spurious (global) minimum near zero (at $\sigma = 0.00001309906$ to be exact), although there is a reasonable local minimum at $\sigma = 1.265882$. In contrast, the global minimum of the DPD objective function (also plotted in Figure 1) for $\alpha = 0.8$ is obtained at $\sigma = 1.20833$. If the last observation 1.048653×10^{-5} were removed from this data set, this spurious minimum behaviour of the LDPD objective function disappears. The minimum LDPD estimator now equals $\sigma = 1.321807$ at $\alpha = 0.8$, an entirely reasonable value. (The corresponding minimum DPD estimator is 1.298228). Thus one single observation can bring about an absolutely drastic change in the minimum LDPD estimator which is against the spirit of stability that robust estimators should have; so far we (or indeed anybody else), have not detected such spurious behaviour in any scenario involving the DPD.

The main reason for this phenomenon in the LDPD and the bridge divergences is the behaviour of the log function at 0, where it diverges to $-\infty$. The validity of this reasoning is confirmed by examining the behaviour of the LDPD and the BDPD close to it in case of the $\mathcal{N}(0,\sigma^2)$ model, where it can be expected that the estimate of σ will be driven to zero if there are some inner contamination observations near 0 in the sample.

Thus while we recognize the merit in the findings of Fujisawa and Eguchi (2008) and Fujisawa (2013), we also note that it would be prudent to be aware of the pitfalls of this method of inference. Clearly these procedures can do with further refinement. Here we strive to improve the procedures

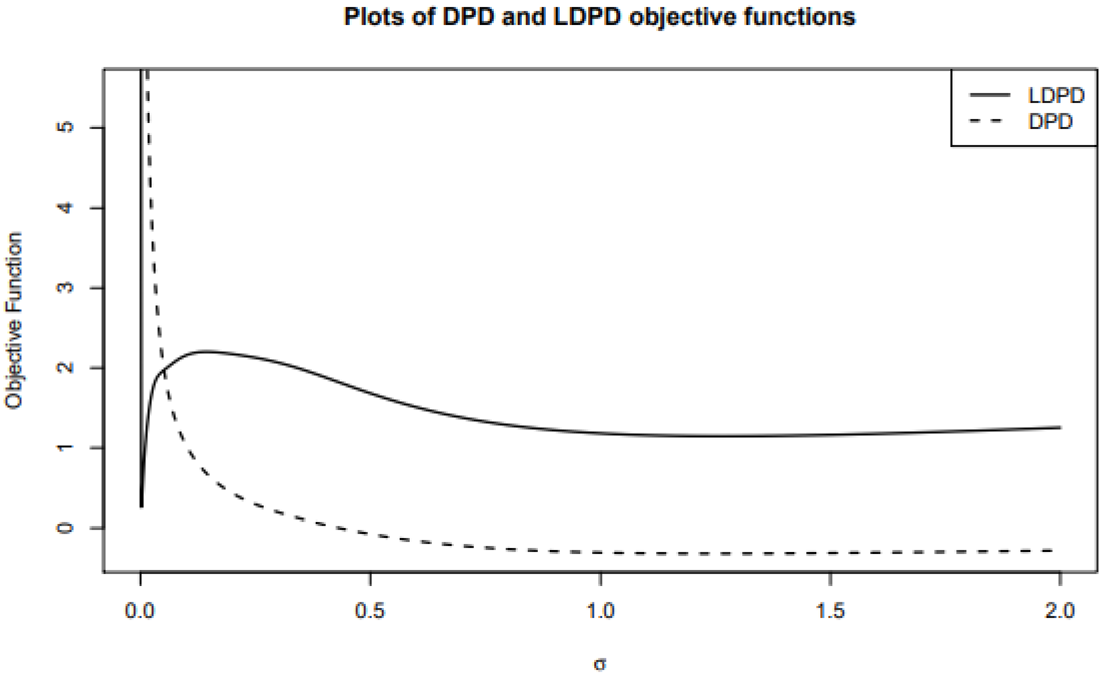

Figure 1: Plots of the sample based DPD and LDPD objective functions under the $\mathcal{N}(0, \sigma^2)$ model based on a sample of size 20 from $\mathcal{N}(0, 1)$.

based on the DPD and the LDPD in a way such that,

- We get good robustness properties like LDPD.
- No spurious minima occur.

If successful, this would combine the positive properties of these two classes.

5 Methodology

We have taken two approaches to deal with the above problem. Before we describe them, it is necessary to briefly introduce another divergence. Mukherjee et al. (2018) [16] have proposed a new divergence called the B-exponential divergence which is essentially a Bregman divergence for the Bregman function

$$B(y) = \frac{2(e^{\alpha y} - \alpha y - 1)}{\alpha^2},$$

This divergence has the form

$$d_\alpha(g, f) = \frac{2}{\alpha} \int \left[e^{\alpha f(x)} \left(f(x) - \frac{1}{\alpha} \right) - e^{\alpha f(x)} g(x) + \frac{1}{\alpha} e^{\alpha g(x)} \right], \quad \alpha \in \mathbb{R}, \tag{5.1}$$

where α is the tuning parameter of the divergence. For a random sample $X_1, X_2, \ldots, X_n$ from G, the optimization of the objective function in (5.1) leads to an estimating equation of the form

$$\frac{1}{n}\sum_{i=1}^{n} u_\theta(X_i) f_\theta(X_i) e^{\alpha f_\theta(X_i)} - \int u_\theta(x) f_\theta^2(x) e^{\alpha f_\theta(x)} dx = 0 \tag{5.2}$$

which is an unbiased estimating equation under the model. Refining this, Mukherjee et al. (2018) [16] have proposed a generalised estimator arising from (5.2) through the extended estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} u_\theta(X_i) f_\theta^\beta(X_i) e^{\alpha f_\theta(X_i)} - \int u_\theta(x) f_\theta^{1+\beta}(x) e^{\alpha f_\theta(x)} dx = 0, \quad \alpha \in \mathbb{R},\ \beta \geq 0. \tag{5.3}$$

We refer to the associated estimator as generalized B-exponential divergence estimator with tuning parameter α and β and refer to it as GBEDE(α, β). This two parameter family represented by (5.3) gives us more flexibility to come up with an estimator having better balance between robustness and efficiency. Mukherjee et al. (2018) present some simulation results demonstraing improvements (over the inference based on DPD alone) by using this generalized form in equation (5.3).

This estimating equation does not correspond to a explicit divergence. However one may find an approximate measure based on numerical integration. If we take a closer look at this estimating equation, we can see a close resemblance with the estimating equation of DPD. The weights on the score function is given by the term

$$w_\theta(x) = w(f_\theta(x)) = f_\theta^\beta(x)\ e^{\alpha f_\theta(x)}.$$

So the weight function is the product of the weight function of the DPD and an exponential weight. Hence, when $\alpha = 0$, the estimating equation corresponds to the DPD estimating equation.

5.1 Modifying the LDPD estimating equation

We want to implement this idea of the modified weight function to the LDPD estimator. The LDPD estimator comes through normalizing the estimating function of the DPD estimator. Then, we propose an new generalised estimating equation by normalizing the extended estimating equation in (5.3) as

$$\frac{\sum_{i=1}^{n} u_\theta(X_i) f_\theta^\beta(X_i) e^{\alpha f_\theta(X_i)}}{\sum_{i=1}^{n} f_\theta^\beta(X_i) e^{\alpha f_\theta(X_i)}} = \frac{\int u_\theta(x) f_\theta^{1+\beta}(x) e^{\alpha f_\theta(x)} dx}{\int f_\theta^{1+\beta}(x) e^{\alpha f_\theta(x)} dx} \tag{5.4}$$

Observe that, here also if we put $\alpha = 0$ to this estimating equation, the equation matches with the estimating equation of LDPD estimator.

We want to explore its robustness properties and compare the bias to the non-normalised version proposed in the paper. We would investigate the proposition by Fujisawa (2013) [14] about the increased robustness properties of the normalised estimating equation.

On this occasion note that, when the tuning parameter α is set to 0, the equation (5.4) matches the estimating equation of LDPD estimator (4.14). But that can not escape the spurious minimum phenomenon pointed out by Jones et al (2001) [17] and Kuchibhotla et al (2019) [15]. Hence, we want to investigate the behaviour of this estimating equation for positive tuning parameter β. Basically, we want to add something to the weight function which will prevent the logarithm from going to $-\infty$.

5.2 A new Bregman divergence

Here we take another approach. Here our aim is to find a suitable convex function so that we can propose a generalized class of Brègman divergences that generates the DPD class as a special case. Define the Bregman function as

$$B_{\alpha,\beta}(x) = \frac{x^\alpha - 1}{\alpha} \, \frac{e^{\beta x} - 1}{\beta} \tag{5.5}$$

where α and β are the tuning parameters of the system. It is easy to check that the function is strictly convex and twice differentiable. Hence, this function will generate a Bregman divergence. Note that, this family of divergences contain DPD(α) as a subfamily when $\beta = 0$. It is important to note that, at $\beta = 0$ the function is not defined. Hence we take it as a limiting case of $\beta \to 0$.

$$B_{\alpha,0}(x) = \lim_{\beta \to 0} B_{\alpha,\beta}(x) = \frac{x^\alpha - 1}{\alpha} \lim_{\beta \to 0} \frac{e^{\beta x} - 1}{\beta} = \frac{x^{1+\alpha} - x}{\alpha}. \tag{5.6}$$

Check that $B_{\alpha,0}(\cdot)$ matches with the generting Bregman function for DPD as in (4.13). Also, we can obtain the LD when $\alpha, \beta \to 0$. We can write this as

$$B_{0,0}(x) = \lim_{\alpha \to 0, \beta \to 0} B_{\alpha,\beta}(x) = \lim_{\alpha \to 0} \frac{x^\alpha - 1}{\alpha} \lim_{\beta \to 0} \frac{e^{\beta x} - 1}{\beta} = x \ \log x. \tag{5.7}$$

Suppose we have an independent and identically distributed sample $X_1, X_2, \dots, X_n$ from the true distribution G, and we try to model this distribution by a parametric family $F = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ where θ is unknown but the functional form of F_θ is known to us. If we wish to use the minimum Brègman divergence approach for the estimation of the unknown parameter θ, the functional $T(G)$ will be defined as

$$D_B(g, f_{T(G)}) = \arg\min_{\theta \in \Theta} D_B(g, f_\theta).$$

Then, using the general Bregman Divergence objective function (4.9) for the Bregman function in (5.5), for the estimation procedure we need to minimize

$$\int \left[f_\theta^\alpha \frac{e^{\beta f_\theta} - 1}{\beta} - \frac{f_\theta^\alpha - 1}{\alpha} \frac{e^{\beta f_\theta} - 1}{\beta} + \frac{f_\theta^\alpha - 1}{\alpha} e^{\beta f_\theta} \ f_\theta \right] dx \\ - n^{-1} \sum_{i=1}^{n} \left[\frac{f_\theta^\alpha(X_i) - 1}{\alpha} e^{\beta f_\theta(X_i)} + f_\theta^{\alpha-1}(X_i) \ \frac{e^{\beta f_\theta(X_i)} - 1}{\beta} \right] \tag{5.8}$$

with respect to θ to get the desired estimate. Now, consider the function

$$\phi_{\alpha,\beta}(x) = (\alpha - 1)x^{\alpha-1}\frac{e^{\beta x} - 1}{\beta} + 2 \ x^\alpha e^{\beta x} + \frac{x^\alpha - 1}{\alpha} \ e^{\beta x} \beta x \tag{5.9}$$

Then we get the estimating equation for this divergence as

$$\int u_\theta(x) f_\theta(x) \phi_{\alpha,\beta} \left(f_\theta(x)\right) dx - n^{-1} \sum_{i=1}^{n} u_\theta(X_i) \phi_{\alpha,\beta} \left(f_\theta(X_i)\right) = 0. \tag{5.10}$$

We can easily see that this is an unbiased estimating equation and matches the generalised weighted estimating equation (4.11) with weights $w_\theta(x) = w(f_\theta(x)) = \phi_{\alpha,\beta}(f_\theta(x))$. When $\beta \to 0$, the estimating equation weights the scores with a constant multiple of f^α which matches the estimating equation of DPD.

5.2.1 Link with M-estimation

All the estimators described in the previous two subsections are M-estimators. **For an M-estimator, the estimating equation can be written as $\sum_i \psi(X_i, \theta) = 0$, for a suitable function $\psi(\cdot, \cdot) : \mathbb{R}^2 \to \mathbb{R}$. From (5.10) we can obtain the ψ function as**

$$\psi(x, \theta) = u_\theta(x) \phi_{\alpha,\beta} \left(f_\theta(x)\right) - \int u_\theta(y) f_\theta(y) \phi_{\alpha,\beta} \left(f_\theta(y)\right) dy \tag{5.11}$$

Notice that this is an unbiased estimating equation under the model. Therefore, the theoretical properties of the GBEDE can be derived directly from the properties of M-estimators.

5.2.2 The Asymptotic Distribution

Let g be the true data generating density. By exploring the theoretical estimating equation $\int \psi(x, \theta) dG(x) = 0$ where $\psi(\cdot, \cdot)$ is as in (5.11), it is immediately seen that the functional $T_{\alpha,\beta}(G)$ associated with the minimum distance estimator is Fisher consistent; it recovers the true value of θ when the data generating density is in the model, i.e.

$g = f_\theta$ for some $\theta \in \Theta$. When g is not in the model, $\theta_{\alpha,\beta} = T_{\alpha,\beta}(G)$ will be the root of the theoretical version of the generalized estimating equation (5.10) which may be expressed as

$$\int u_\theta(x)\phi_{\alpha,\beta}\left(f_\theta(x)\right) g(x)dx - \int u_\theta(x) f_\theta(x)\phi_{\alpha,\beta}\left(f_\theta(x)\right) dx = 0. \tag{5.12}$$

Now, let us define,

$$\begin{aligned} J_{\alpha,\beta}(\theta) = & \int_x u_\theta(x)u_\theta^\top(x)f_\theta(x)\phi_{\alpha,\beta}(f_\theta(x))dx \\ & + \int_x \Big\{ I_\theta(x)\phi_{\alpha,\beta}(x)(f_\theta(x)) - u_\theta(x)u_\theta^\top(x)f_\theta(x)\phi'_{\alpha,\beta}(f_\theta(x))\Big\}\Big\{g(x) - f_\theta(x)\Big\} = 0 \end{aligned} \tag{5.13}$$

and

$$K_{\alpha,\beta}(\theta) = \int_x u_\theta(x)u_\theta^\top(x)\phi^2_{\alpha,\beta}(f_\theta(x))g(x)dx - \xi_{\alpha,\beta}(\theta)\xi^\top_{\alpha,\beta}(\theta) \tag{5.14}$$

where

$$\xi_{\alpha,\beta}(\theta) = \int_x u_\theta(x)\phi_{\alpha,\beta}(f_\theta(x))g(x)dx \tag{5.15}$$

and $I_\theta(x) = -\frac{\partial}{\partial\theta^\top}u_\theta(x)$ is the information function of the model. The following theorem provides the asymptotic distribution of the minimum distance estimator.

The following theorem is provided under the set of assumptions given below. These may be viewed as generalizations of the conditions presented in Basu et al. (2011) [18] (which were designed specifically for the DPD class). The details of the proof are not presented here, as it mimics the approach of Theorem 9.2 of Basu et al. (2011) [18] exactly.

- **(A1)** The distributions F_θ of X have common support, so that the set $\chi = \{x : f_\theta(x) > 0\}$ is independent of $\boldsymbol{\theta}$. The distribution of G is also supported on χ, on which the corresponding density g is greater than zero.

- **(A2)** There is an open subset ω of the s-dimensional parameter space Ω containing the best fitting parameter $\boldsymbol{\theta}_g$ such that for almost all $x \in \chi$ and all $\boldsymbol{\theta} \in \omega$, the density $f_{\boldsymbol{\theta}}(x)$ is three times differentiable with respect to $\boldsymbol{\theta}$ and the third partial derivatives are continuous with respect to θ.

- The integrals $\int \left[f_\theta(x)B'\left(f_\theta(x)\right) - B\left(f_\theta(x)\right)\right]dx$ and $\int B'\left(f_\theta(x)\right)g(x)dx$ can be differentiated three times with respect to θ and the derivatives can be taken under the integral sign.

- **The $s \times s$ matrix $J(\boldsymbol{\theta})$, with its (k,l) entry defined as**

$$J_{kl}(\boldsymbol{\theta}) = \mathrm{E}_g \left[\nabla_{kl} \left\{ \int \left[f_{\boldsymbol{\theta}}(x) B'\left(f_{\boldsymbol{\theta}}(x)\right) - B\left(f_{\boldsymbol{\theta}}(x)\right)\right] dx - B'\left(f_{\boldsymbol{\theta}}(x)\right) \right\} \right],$$

  is positive definite. Here ∇_{kl} denotes the partial derivative of a function with respect to the k th and l th components of its argument and E_g represents the expectation under the density g.

- **There exists a function $M_{jkl}(x)$ such that**

$$|\nabla_{jkl} V_{\boldsymbol{\theta}}(x)| \leq M_{jkl}(x) \quad \forall \boldsymbol{\theta} \in \omega,$$

  where $V_{\boldsymbol{\theta}}(x) = \int \left[f_{\boldsymbol{\theta}}(x) B'\left(f_{\boldsymbol{\theta}}(x)\right) - B\left(f_{\boldsymbol{\theta}}(x)\right)\right] dx - B'\left(f_{\boldsymbol{\theta}}(x)\right)$ **and**

$$\mathrm{E}_g\left[M_{jkl}(X)\right] = m_{jkl} < \infty \qquad \textbf{for all } j, k \textbf{ and } l.$$

Theorem 5.1. Assuming that conditions $A1 - A5$ hold,

1. The estimating equation given by (5.10) has a consistent sequence of roots $\hat{\theta}_{\alpha,\beta} = T_{\alpha,\beta}(G_n)$
2. $n^{1/2}\left(\widehat{\theta}_{\alpha,\beta} - \theta_{\alpha,\beta}\right)$ has an asymptotic multivariate normal distribution with mean zero (vector) and covariance matrix $J^{-1}_{\alpha,\beta} K_{\alpha,\beta} J^{-1}_{\alpha,\beta}$, where $J_{\alpha,\beta}$ and $K_{\alpha,\beta}$ are as in equations (5.13) and (5.14) respectively.

When the true distribution G belongs to the model so that $G = F_\theta$ for some $\theta \in \Theta$, the formula for $J = J_{\alpha,\beta}(\theta), K = K_{\alpha,\beta}(\theta)$ and $\xi = \xi_{\alpha,\beta}(\theta)$ simplify to

$$\begin{aligned} &J = \int_x u_\theta(x) u_\theta^T(x) f_\theta(x) \phi_{\alpha,\beta}(f_\theta(x)) dx, \quad K = \int_x u_\theta(x) u_\theta^\top(x) \phi^2_{\alpha,\beta}(f_\theta(x)) f_\theta dx - \xi_{\alpha,\beta}(\theta) \xi^\top_{\alpha,\beta}(\theta) \\ &\xi = \int_x u_\theta(x) \phi_{\alpha,\beta}(f_\theta(x)) f_\theta(x) dx \end{aligned} \tag{5.16}$$

Note that the matrix J is the expectation of the partial derivative matrix, and K is the covariance matrix, of the ψ function given in Equation (5.15). We use the above formulas to evaluate the theoretical asymptotic efficiencies for the normal location parameter. The results are presented in Table 1 . At $\alpha = 0$, the estimators are simply the minimum DPD estimators, and hence the asymptotic efficiencies are decreasing with β.

β/α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	100.00	98.76	95.86	92.12	88.01	83.81	79.67	75.68	71.89	68.31	64.95
0.2	99.95	98.60	95.55	91.64	87.37	83.00	78.69	74.53	70.56	66.81	63.26
0.4	99.70	98.42	95.22	91.16	86.72	82.17	77.69	73.34	69.18	65.22	61.46
0.6	98.90	98.23	94.89	90.68	86.07	81.34	76.65	72.10	67.71	63.52	59.51
0.8	95.64	98.03	94.57	90.20	85.42	80.50	75.60	70.81	66.17	61.70	57.39
1	61.23	97.83	94.26	89.75	84.79	79.66	74.52	69.46	64.52	59.71	55.03

Table 1: Asymptotic Relative Efficiency under normal location model

5.2.3 Influence function

We are concerned about M-estimators. The influence function of these estimators is readily available. In the general setup this is given by

$$IF(y, T_{\alpha,\beta}, G) = J^{-1}_{\alpha,\beta}(\theta)\left\{u_\theta(y) f^{\beta}_{\theta}(y) e^{\alpha f_\theta(y)} - \xi_{\alpha,\beta}(\theta)\right\}$$

where $J_{\alpha,\beta}$ and $\xi_{\alpha,\beta}$ are given in Equations (5.13) and (5.15), respectively. When the true distribution belongs to the model, a simplified form for the influence function is obtained by replacing $J_{\alpha,\beta}$ with J and ξ as given in Equation (5.16). If we assume that J and ξ are finite, then the influence function is bounded whenever $u_\theta(y) f^{\beta}_{\theta}(y) e^{\alpha f_\theta(y)}$ is bounded in y. This is true for the most standard parametric models, including the normal location scale model, for all finite values of $\alpha, \beta \geq 0$. On the other hand, the influence function is not bounded in case of the maximum likelihood estimator (corresponding to $\alpha = \beta = 0$).

6 Simulation

6.1 Pure model

For simulation studies, first we have calculated the efficiency of the Bregman estimator under the true model. We have considered two scenarios, one for the estimation of location parameter and the other one for the estimation of scale parameter. For both the cases, the true model is considered as $\mathcal{N}(\prime, \infty$ and a moderately large sample size $n = 75$ has been used. We have simulated the data with seed 2017 for 500 replications and took the ratio of the maximum likelihood estimator to that of the minimum distance estimator and the results are listed in table 2 and table 3. For all the simulation studies, we have considered eleven values of the tuning parameter α, namely from 0 to 1 with interval of 0.1 and six values of the tuning parameter β, namely from 0 to 1 with interval of 0.2. For the estimation procedure, we have considered the estimating equation and to find the roots, we have used a bisection algorithm.

β/α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	100.00	97.25	95.31	94.46	88.62	89.32	76.34	76.73	69.51	70.04	64.19
0.2	99.98	99.27	93.4	90.49	89.66	84.55	84.76	70.95	71.57	64.69	67.53
0.4	99.29	99.27	98.12	94.6	93.08	74.68	84.56	75.57	73.85	69.55	57.47
0.6	98.68	96.62	91.93	89.94	87.35	86.38	77.12	80.21	69.83	61.93	63.7
0.8	96.31	98.24	93.64	86.14	85.45	80.46	79.79	64.93	66.95	59.91	58.96
1	65.81	96.64	94.24	89.36	83.32	80.03	70.47	64.64	67.28	59.53	57.55

Table 2: Empirical Efficiency of the estimator under the normal location model

β/α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	100.00	96.29	89.30	86.82	79.26	75.84	65.75	59.95	56.32	49.87	55.36
0.2	89.83	97.84	89.53	81.02	80.48	72.91	69.46	60.57	55.31	52.18	46.41
0.4	74.25	97.24	91.08	85.36	77.42	69.97	67.40	70.60	60.28	56.77	49.50
0.6	70.76	96.30	91.15	86.40	71.36	69.50	60.18	61.63	59.29	50.54	54.27
0.8	60.11	99.09	88.19	89.27	70.82	69.10	61.75	50.48	58.00	49.69	50.53
1	55.16	96.26	89.32	83.90	75.25	68.95	61.56	55.08	54.60	49.69	47.76

Table 3: Efficiency of estimator under normal scale model

6.2 Contaminated model

For the contaminated model, we have considered the true distribution as $\mathcal{N}(\prime, \infty)$ and the contaminating distribution as $\mathcal{N}(5, 1)$. We have taken the contamination proportion as $\varepsilon = 0.05$. For this contaminated data, we have simulated from the mixture model $0.95\mathcal{N}(\prime, \infty) + 0.05\mathcal{N}(\bigtriangledown, \infty)$ and a moderately small size $n = 40$ has been used as was used in [14]. We have simulated the data with seed 2017 for 500 replications and took the ratio of the maximum likelihood estimator to that of the minimum distance estimator . We have considered the location model and the scale model. For the estimation of the scale parameter, we have used both the normalised and non-normalised estimating equations. The results are listed in table 4 and table 5 and 6. For all the simulation studies, we have considered eleven values of the tuning parameter α, namely from 0 to 1 with interval of 0.1 and six values of the tuning parameter β, namely from 0 to 1 with interval of 0.2. For the estimation procedure, we have considered the estimating equation and to find the roots, we have used a bisection algorithm.

β/α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	1.00	2.95	3.54	3.53	4.13	3.57	3.44	3.51	2.87	3.11	2.71
0.2	1.05	2.96	3.86	3.63	3.44	3.92	3.35	3.18	3.31	2.67	2.92
0.4	1.11	2.84	3.83	3.68	3.76	3.33	3.21	3.11	2.77	3.04	2.91
0.6	1.20	2.88	3.75	4.11	3.45	3.62	3.53	3.26	2.80	3.00	2.18
0.8	1.46	2.95	3.52	3.64	3.97	3.90	3.47	2.68	2.83	2.74	2.52
1	0.48	2.94	4.16	3.67	4.33	3.48	3.33	3.32	2.59	2.49	2.16

Table 4: Efficiency under Normal location model with contaminated data

β/α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	1.00	2.23	6.12	15.44	21.51	21.83	21.45	19.78	19.47	20.90	21.82
0.2	0.81	1.32	2.32	3.52	4.57	4.88	5.02	4.64	4.63	4.35	4.21
0.4	0.83	1.37	2.19	3.62	4.74	4.88	5.20	4.82	4.25	4.35	4.15
0.6	0.94	1.28	2.40	3.27	4.57	5.18	4.82	4.86	4.39	4.40	3.56
0.8	1.21	1.26	2.32	3.59	4.15	4.74	4.94	4.37	4.36	4.13	3.83
1	15.89	1.17	2.06	3.55	4.83	4.64	4.44	4.67	4.19	3.93	3.63

Table 5: Efficiency under normal scale model with contaminated data for normalised estimating equation

β/α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	1.00	2.23	6.12	15.44	21.51	21.83	21.45	19.78	19.47	20.90	21.82
0.2	0.81	1.34	2.29	3.41	4.41	4.65	4.84	4.40	4.59	4.28	4.36
0.4	0.83	1.37	2.18	3.49	4.56	4.60	4.92	4.78	4.21	4.43	4.41
0.6	0.94	1.30	2.38	3.17	4.35	4.96	4.69	4.71	4.56	4.67	3.92
0.8	1.21	1.28	2.30	3.47	3.94	4.52	4.90	4.25	4.44	4.34	4.35
1	15.89	1.20	2.07	3.44	4.58	4.42	4.33	4.69	4.21	4.46	4.08

Table 6: Efficiency under normal scale model with contaminated data for normalised estimating equation

7 Future plan of study

In this report we have described the problem of combining the nice properties of the DPD and LDPD and produce a general estimation procedure. We have made some proposals but have not yet gone that far as to claim success. In the second half of this dissertation we hope to deal with actual implementation of our proposals and the related issues.

References

[1] D. V. Lindley. “Introduction to the practice of statistics, (3rd edition), by David S. Moore and George P. McCabe. Pp. 825 (with appendices and CD-ROM). £27.95. 1999. ISBN 0 7167 3502 4 (W. H. Freeman).” In: *The Mathematical Gazette* 83.497 (1999), pp. 374–375. DOI: 10.2307/3619120.

[2] Frank E. Grubbs. “Procedures for Detecting Outlying Observations in Samples”. In: *Technometrics* 11.1 (1969), pp. 1–21. DOI: 10.1080/00401706.1969.10490657. eprint: https://www.tandfonline.com/doi/pdf/10.1080/00401706.1969.10490657. URL: https://www.tandfonline.com/doi/abs/10.1080/00401706.1969.10490657.

[3] Peter J. Huber. *Robust statistics.* English. John Wiley & Sons, Hoboken, NJ, 1981. DOI: 10.1002/0471725250.

[4] J. Kiefer and J. Wolfowitz. “Stochastic Estimation of the Maximum of a Regression Function”. In: *The Annals of Mathematical Statistics* 23.3 (1952), pp. 462–466. DOI: 10.1214/aoms/1177729392. URL: https://doi.org/10.1214/aoms/1177729392.

[5] David L. Donoho and Richard C. Liu. “The ”Automatic” Robustness of Minimum Distance Functionals”. In: *The Annals of Statistics* 16.2 (1988), pp. 552–586. DOI: 10.1214/aos/1176350820. URL: https://doi.org/10.1214/aos/1176350820.

[6] Noel Cressie and Timothy R. C. Read. “Multinomial Goodness-of-Fit Tests”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 46.3 (1984), pp. 440–464. ISSN: 00359246. URL: http://www.jstor.org/stable/2345686.

[7] Rudolf Beran. “Minimum Hellinger Distance Estimates for Parametric Models”. In: *The Annals of Statistics* 5.3 (1977), pp. 445–463. DOI: 10.1214/aos/1176343842. URL: https://doi.org/10.1214/aos/1176343842.

[8] Chanseok Park and Ayanendranath Basu. “Minimum disparity estimation: asymptotic normality and breakdown point results”. In: *Bulletin of Informatics and Cybernetics* 36 (Dec. 2004), pp. 19–33. DOI: 10.5109/12576.

[9] Ayanendranath Basu, Chanseok Park, and Bruce Lindsay. “Some variants of minimum disparity estimation”. In: *Computational Statistics & Data Analysis* 45 (May 2004), pp. 741–763. DOI: 10.1016/S0167-9473(03)00098-7.

[10] Ayanendranath Basu and Bruce Lindsay. “Minimum disparity estimation for continuous models: Efficiency, distributions and robustness”. In: *Annals of the Institute of Statistical Mathematics* 46 (Feb. 1994), pp. 683–705. DOI: 10.1007/BF00773476.

[11] Ayanendranath Basu et al. “Robust and efficient estimation by minimising a density power divergence”. In: *Biometrika* 85 (Sept. 1998). DOI: 10.1093/biomet/85.3.549.

[12] Michael P. Windham. “Robustifying Model Fitting”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.3 (1995), pp. 599–609. DOI: https://doi.org/10.1111/j.2517-6161.1995.tb02050.x. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1995.tb02050.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1995.tb02050.x.

[13] Hironori Fujisawa and Shinto Eguchi. “Robust parameter estimation with a small bias against heavy contamination”. In: *Journal of Multivariate Analysis* 99.9 (2008), pp. 2053–2081.

[14] Hironori Fujisawa. “Normalized estimating equation for robust parameter estimation”. In: *Electronic Journal of Statistics* 7 (Jan. 2013). DOI: 10.1214/13-EJS817.

[15] Arun Kuchibhotla, Somabha Mukherjee, and Ayanendranath Basu. “Statistical Inference based on Bridge Divergences”. In: *Annals of the Institute of Statistical Mathematics* 71 (June 2017). DOI: 10.1007/s10463-018-0665-x.

[16] Taranga Mukherjee, Abhijit Mandal, and Ayanendranath Basu. “The B-exponential divergence and its generalizations with applications to parametric estimation”. In: *Statistical Methods & Applications* 28 (Nov. 2018). DOI: 10.1007/s10260-018-00444-8.

[17] M. Jones et al. “A Comparison of related density-based minimum divergence estimators”. In: *Biometrika* 88 (Oct. 2001). DOI: 10.1093/biomet/88.3.865.

[18] Ayanendranath Basu, Hiroyuki Shioya, and Chanseok Park. *Statistical Inference: The Minimum Distance Approach.* June 2011. ISBN: 9781420099652. DOI: 10.1201/b10956.