On Minimum Bregman Divergence Inference

Soumik Purkayastha[1*] and Ayanendranath Basu[2]

[1*]Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, 48109-2029, MI, USA https://orcid.org/0000-0002-3619-2804.

[2]Interdisciplinary Statistical Research Unit, Indian Statistical Institute, 205 B.T. Road, Kolkata, 700108, WB, INDIA https://orcid.org/0000-0003-1416-9109.

*Corresponding author(s). E-mail(s): soumikp@umich.edu; Contributing authors: ayanendranath.basu@gmail.com;

Abstract

The popular density power divergence (DPD) is a sub-class of Bregman divergences, and it generates an useful class of minimum divergence estimators. In this paper, we propose and study a new sub-class of Bregman divergences called the exponentially weighted divergence (EWD) with the aim of generating better or competitive inference procedures. The EWD is specifically constructed to constrain the corresponding weight function within the $[\mathbf{0}, \mathbf{1}]$ interval, which helps to intuitively understand and perceive the nature and degree of robustness of the resulting inference procedures; such interpretations are less obvious for the DPD. Performances of the two classes in the area of parametric estimation are compared – both through extensive simulations as well as through real life examples. Our parametric estimation procedure, while motivated by i.i.d. data applications, are extended to handle independent but non-homogeneous data structures. General tests of parametric hypotheses based on the Bregman divergences are also considered. We establish the asymptotic null distribution of our proposed test statistic and explore its behaviour when applied to real data. The inference procedures generated by the new EWD divergence appear to provide better alternatives to parametric statistical inference based on DPD-based procedures.

The authors declare no competing interests.

Keywords: Minimum distance inference, density power divergence, robustness, influence function, M-estimation, robust regression, robust testing

1 Introduction

Density based minimum divergence methods are popular tools in statistical inference. In parametric estimation, this amounts to choosing the model density closest (in terms of the selected divergence) to the empirical data density. This approach often combines strong robustness properties with high asymptotic efficiency. An important class of density based divergences is the class of ϕ divergences (see Csiszár (1963)). Under standard regularity conditions, all minimum ϕ divergence estimators have full asymptotic efficiency at the model (Lindsay (1994)); many also have attractive robustness properties. The seminal Hellinger distance study of Beran (1977) appears to be the first which demonstrated that strong robustness properties may be achieved simultaneously with full asymptotic efficiency. Later, the same has been demonstrated with respect to much of the ϕ divergence class (see, eg., Basu, Shioya, and Park (2011)). The usefulness of the corresponding procedures in providing robust alternatives to the likelihood ratio test has also been explored in the literature (Basu et al., 2011; Lindsay, 1994; Simpson, 1989). The approach has been further refined and extended in many directions by later authors. On the whole, the utility of the minimum divergence procedures based on ϕ divergences is well established in the literature.

One major criticism of this inference procedure is that it inevitably involves the use of some form of non-parametric smoothing (such as kernel density estimation) to produce a continuous estimate of the true density. This can throw up several potential difficulties including the problematic bandwidth selection issue and the slow convergence of the of kernel density estimator to the 'truth' (particularly for high dimensional data). The theoretical derivations are also harder. Development of methods which eliminate these difficulties may be worthwhile even if that involves a marginal loss in asymptotic efficiency.

An alternative class of minimum divergence estimators which avoids nonparametric smoothing in the construction of the empirical divergence is the class of minimum Bregman divergence estimators. An important example is the family of density power divergences (henceforth DPD(α), where α is the tuning parameter); the corresponding minimum density power divergence estimators (henceforth MDPDE(α)) have been shown to combine strong robustness properties with high asymptotic efficiency (see Basu, Harris, Hjort, and Jones (1998)). Divergences within the Bregman class have been called decomposable divergences by Broniatowski, Toma, and Vajda (2012) and non-kernel divergences by Jana and Basu (2019). These divergences have simple estimating equations and much of their asymptotic properties can be obtained from the M-estimation theory. The Kullback-Leibler divergence, which is a decomposable divergence, is the only common member between the ϕ divergence and the Bregman divergence classes.

In the context of density-based minimum divergence estimation, we have several 'good' choices available. To justify the development of another family of estimators, one must demonstrate that the new estimators are competitive, if not better than the existing standard. Within the class of minimum divergence estimators which do not require any nonparametric smoothing, the MDPDE(α) is the current standard. In this paper, we will develop a family of divergences yielding minimum divergence estimators which satisfy this requirement; at the least, this family provides a highly competitive standard. Our proposed class of divergences will be called the *exponentially weighted divergence* family, indexed by a tuning parameter β (henceforth referred to as EWD(β)). The corresponding minimum divergence estimator will be denoted by MEWDE(β).

This divergence will also be useful in the field of hypothesis testing. Although the likelihood ratio test has several asymptotic optimality properties at the model, it is also known to have very poor robustness properties. Many density based minimum distance procedures yield robust tests of hypothesis with high efficiency, eg., Basu, Mandal, Martin, and Pardo (2013, 2018); Pardo (2006). Some of these papers address the general problem of parametric hypothesis testing based on the density power divergence. The present work considers this problem in the context of a general Bregman divergence, with special emphasis on our proposed EWD(β) class.

2 The Bregman Divergence

2.1 Introduction to minimum Bregman divergence estimation

The Bregman divergence between two densities g and f is defined as

$$D_B(g,f) = \int_x [B(g(x)) - B(f(x)) - (g(x) - f(x))B'(f(x))]dx, \qquad (1)$$

where $B : \mathbb{R} \to \mathbb{R}$ is a strictly convex function and $B'(\cdot)$ is the derivative of B with respect to its argument. The strict convexity of B assures that the measure $D_B(g,f)$ is non-negative, and equals zero if and only if the arguments are identically equal. Csiszár et al. (1991) discuss this and other measures in greater detail. We note that the convex functions $B(y)$ and $B^*(y) = B(y) + ay + c$ generate identical divergences in Equation (1) for $a, c \in \mathbb{R}$. In this section we discuss minimum Bregman divergence inference in case of data that are independent, but not necessarily identically distributed. The i.i.d. data case (see Section 4 of the supplement) emerges as a special case of this setup.

We assume that the data $X_1, \ldots, X_n$ are independent. For $i = 1, 2, \ldots, n$ we have $X_i \sim g_i$, where g_i are possibly different densities with respect to some common dominating measure. We model g_i by the family $\mathscr{F}_{i,\boldsymbol{\theta}} = \{f_{i,\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \Omega \subseteq \mathbb{R}^s\}$ for each $i = 1, 2, \ldots, n$. An estimate of the Bregman divergence between the density g_i and the associated model density $f_{i,\boldsymbol{\theta}}$ ($\in \mathscr{F}_{i,\boldsymbol{\theta}}$) is given by $D_B(\hat{g}_i, f_{i,\boldsymbol{\theta}})$, where $\hat{g}_i$ is a non-parametric density estimate of g_i.

Since our aim is to reach some 'common' value of $\boldsymbol{\theta}$ (if it exists) which can be used to model each g_i individually, it is intuitive to minimize the average divergence between the data points and the models, given by $H_n(\boldsymbol{\theta}) =$

$n^{-1}\sum_{i=1}^{n} D_B(\hat{g}_i, f_{i,\boldsymbol{\theta}})$ with respect to $\boldsymbol{\theta}$. In presence of only one data point X_i from density g_i, the best density estimate of g_i is the (degenerate) density which puts the entire mass on the observed value of X_i and this yields the objective function (ignoring the terms independent of $\boldsymbol{\theta}$) given by

$$\begin{aligned} H_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n}\left[\int \Big(f_{i,\boldsymbol{\theta}}(x)B'(f_{i,\boldsymbol{\theta}}(x)) - B(f_{i,\boldsymbol{\theta}}(x))\Big)dx - B'(f_{i,\boldsymbol{\theta}}(X_i))\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} V_{i,\boldsymbol{\theta}}(X_i). \end{aligned} \tag{2}$$

Let $\nabla_{\boldsymbol{\theta}}$ represent the gradient with respect to $\boldsymbol{\theta}$. Considering partial derivatives of the objective function in Equation (2), we arrive at the estimating equation

$$\sum_{i=1}^{n}\left[u_{i,\boldsymbol{\theta}}(X_i)w(f_{i,\boldsymbol{\theta}}(X_i)) - \int u_{i,\boldsymbol{\theta}}(x)w(f_{i,\boldsymbol{\theta}}(x))f_{i,\boldsymbol{\theta}}(x)dx\right] = \mathbf{0}, \tag{3}$$

where $u_i(x) = \nabla_{\boldsymbol{\theta}} \log(f_{i,\boldsymbol{\theta}}(x))$ is the likelihood score function of the density $f_{i,\boldsymbol{\theta}}(x)$ used to model the i-th data point, and $w(t) = B''(t) \times t$. If the data were i.i.d. with common density g, we would choose a common model density $f_{\boldsymbol{\theta}}$ for modeling and inference. Consequently, Equation (3) would yield

$$\sum_{i=1}^{n}\left[u_{\boldsymbol{\theta}}(X_i)w(f_{\boldsymbol{\theta}}(X_i)) - \int u_{\boldsymbol{\theta}}(x)w(f_{\boldsymbol{\theta}}(x))f_{\boldsymbol{\theta}}(x)dx\right] = \mathbf{0},$$

where $u_{\boldsymbol{\theta}}(x) = \nabla_{\boldsymbol{\theta}} \log(f_{\boldsymbol{\theta}}(x))$ is the likelihood score function of the model density $f_{\boldsymbol{\theta}}(x)$. Equation (3) is a generalisation of weighted likelihood estimating equation for the case of independent and non-homogeneous data.

2.2 The exponentially weighted divergence

To develop new estimation procedures based on Bregman divergences, one can either (a) start with a specific convex function B as given in Equation (2) and construct a weighted likelihood equation, or (b) begin with a suitable robust weight representation as in Equation (3) and backtrack to recover the corresponding convex function B. We take the latter approach.

Philosophically, our treatment of outliers is probabilistic; an outlying point is one which has a small probability of occurrence under the model $f_{\boldsymbol{\theta}} \in \mathscr{F}_{\boldsymbol{\theta}}$. We downweight those observations in the estimating equation for which the value of $f_{\boldsymbol{\theta}}(x)$ is small. We plot the weight functions for some members of the DPD(α) family at different values of α in Figure 1. We note that the strength of downweighting increases with increasing α. For $f_{\boldsymbol{\theta}}(x) > 1$, these weights grow unboundedly for all $\alpha > 0$ as the argument increases. As an alternative, we propose a new class of divergences based on a different choice of the weight

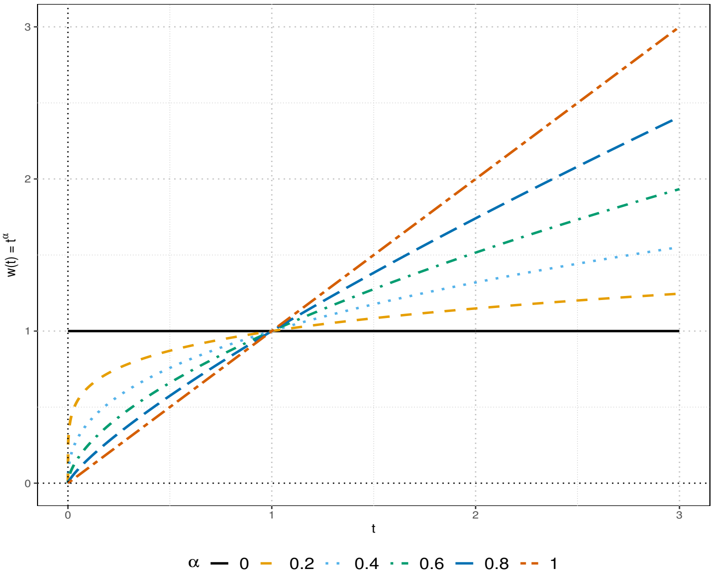

Fig. 1 Weight functions of some DPD(α) members.

function given by (depending on a non-negative index β)

$$w_\beta(t) = \begin{cases} 1 - \exp(-t/\beta) & \text{if } \beta > 0, \\ 1 & \text{if } \beta = 0. \end{cases} \tag{4}$$

These weights smoothly drop to zero for decreasing values of the pdf $f_{\boldsymbol{\theta}}(x)$ for $\beta > 0$. Unlike the DPD(α) weights, they are bounded above by 1. We plot the weight functions given by Equation (4) for specific values of β in Figure 2. The likelihood equation may be recovered at $\beta = 0$, where, to avoid the complications of division by zero, the weights have been defined by the corresponding limiting case as $\beta \to 0$. Using the relation $w(t) = B''(t) \times t$, we recover the associated B function

$$B_\beta(x) = \frac{x^2}{\beta}\Big[\sum_{n=0}^{\infty} \frac{(-x/\beta)^n}{(n+2)!(n+1)}\Big].$$

In Section 1 of the supplement, we show that this can be further simplified to

$$B_\beta(x) = -x + \gamma x + \beta - \beta\exp(-x/\beta) + x\Gamma(0, x/\beta) + x\log(x/\beta),$$

where γ is the Euler-Mascheroni constant and $\Gamma(\alpha, \beta)$ is the incomplete Gamma integral (see Section 3 of the supplement for more details). We can equivalently consider the following simplified form of the defining function

$$B^*_\beta(x) = -\beta\exp(-x/\beta) + x\Gamma(0, x/\beta) + x\log(x/\beta). \tag{5}$$

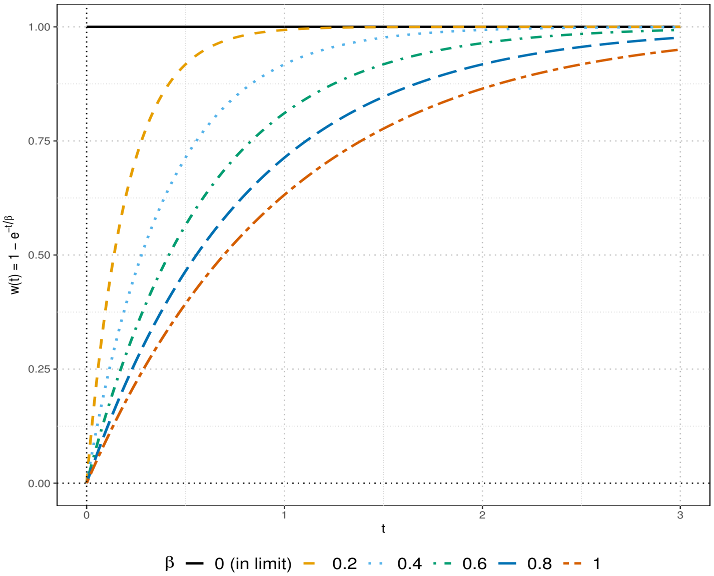

Fig. 2 Weight functions of some EWD(β) members.

The associated Bregman divergence (which we will refer to as the exponentially weighted divergence EWD(β)) has the form

$$\frac{1}{n}\sum_{i=1}^{n}\Big[\int\Big(f_{i,\boldsymbol{\theta}}(x)B'(f_{i,\boldsymbol{\theta}}(x)) - B(f_{i,\boldsymbol{\theta}}(x))\Big)dx - B'(f_{i,\boldsymbol{\theta}}(X_i))\Big], \tag{6}$$

where $B = B^*_\beta$ is as in Equation (5). The MEWDE(β) solves the estimating equation

$$\begin{aligned}\sum_{i=1}^{n}\Bigg[&u_{i,\boldsymbol{\theta}}(X_i)\Big[1-\exp\Big(-f_{i,\boldsymbol{\theta}}(X_i)/\beta\Big)\Big] - \\ &\int u_{i,\boldsymbol{\theta}}(x)\Big[1-\exp\Big(-f_{i,\boldsymbol{\theta}}(x)/\beta\Big)\Big]f_{i,\boldsymbol{\theta}}(x)dx\Bigg] = \mathbf{0}.\end{aligned} \tag{7}$$

3 Properties

3.1 Fisher consistency of minimum Bregman divergence estimators

We consider independent data $X_1, \ldots, X_n$ with $X_i \sim g_i$, where g_i are (possibly) different densities with respect to some common dominating measure. Let G_i be the distribution function associated with g_i for $i = 1, 2, \ldots, n$. The minimum Bregman divergence functional $T_B(G_1, G_2, \ldots, G_n)$ for non-homogeneous observations is given by the relation $T_B(G_1, G_2, \ldots, G_n) = \underset{\boldsymbol{\theta}\in\Omega}{\operatorname{argmin}}\ H_n^*(\boldsymbol{\theta})$, where $H_n^*(\theta) = \frac{1}{n}\sum_{i=1}^n D_B(g_i, f_{i,\theta})$ is the theoretical version of the expression is defined by Equation (2). As the Bregman divergence so defined is a genuine divergence in the sense that it is non-negative and attains its minimum if and only if each data generating distribution G_i equals the model counterpart $F_{i,\theta}$, the functional T is Fisher consistent in the sense $T(F_{1,\theta}, F_{2,\theta}, \ldots, F_{n,\theta}) = \boldsymbol{\theta}$.

3.2 Asymptotic properties

We derive the asymptotic distribution of the minimum Bregman divergence estimator $\hat{\boldsymbol{\theta}}_n$ defined by the relation $\hat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta}\in\Omega}{\operatorname{argmin}}\ H_n(\boldsymbol{\theta})$, provided such a minimum exists, where $H_n(\theta)$ is as defined in Equation (2). In particular the presented results hold for MEWDE(β).

We assume that the data are generated from the setup described in Section 3.1. We model g_i by the parametric family $\mathscr{F}_{i,\boldsymbol{\theta}} = \{f_{i,\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Omega, \Omega \subseteq \mathbb{R}^s\}$ for each $i = 1, 2, \ldots, n$. We also assume that there exists a best fitting value of $\boldsymbol{\theta}$ which is independent of the index i of the different densities and let us denote it by $\boldsymbol{\theta}_g$. It is important to note that this assumption is satisfied if all the true densities g_i belong to the model family so that $g_i = f_{i,\boldsymbol{\theta}}$ for some common $\boldsymbol{\theta} = \boldsymbol{\theta}_g$. The minimum Bregman divergence estimating equation i given by Equation (3), which is solved by the minimizer of $H_n(\boldsymbol{\theta})$ as defined in Equation (2). We now define, for $i = 1, 2, \ldots$

$$H^{(i)}(\boldsymbol{\theta}) = \int \Big(f_{i,\boldsymbol{\theta}}(x)B'(f_{i,\boldsymbol{\theta}}(x)) - B(f_{i,\boldsymbol{\theta}}(x)) - B'(f_{i,\boldsymbol{\theta}}(x))g_i(x)\Big)dx, \tag{8}$$

so that for the best fitting parameter $\boldsymbol{\theta}_g$, we have $\nabla_{\boldsymbol{\theta}} H^{(i)}(\boldsymbol{\theta}_g) = \mathbf{0}$, for $i = 1, 2, \ldots$ We also define, for each $i = 1, 2, \ldots$, the $s \times s$ matrix $J^{(i)}$ with (k, l)-th entry given by

$$J_{kl}^{(i)} = E_{g_i}[\nabla_{kl} V_{i,\boldsymbol{\theta}}(X_i)], \tag{9}$$

where $V_{i,\boldsymbol{\theta}}(X_i)$ is as defined in Equation (2), $\nabla_{kl} V_{i,\boldsymbol{\theta}}(X_i) = \frac{\partial^2 V_{i,\boldsymbol{\theta}}(X_i)}{\partial\theta_k \partial\theta_l}$ is the (k, l)-th component of $\nabla_\theta V_{i,\theta}(X_i)$, $E_{g_i}[\cdot]$ denotes expectation under the distribution specified by g_i and θ_i denotes the i-th component of $\boldsymbol{\theta}$. We also define

$$\Psi_n = \frac{1}{n}\sum_{i=1}^n J^{(i)}, \tag{10}$$

and the matrix

$$\Omega_n = \frac{1}{n}\sum_{i=1}^n V_{g_i}[\nabla_{\boldsymbol{\theta}} V_{i,\boldsymbol{\theta}}(X_i)], \tag{11}$$

where V_{g_i} denotes taking variance under the distribution specified by g_i. The matrix defined in Equation (9) has the following expression

$$J^{(i)} = \int u_{i,\boldsymbol{\theta}_g}(x)u^T_{i,\boldsymbol{\theta}_g}(x)w(f_{i,\boldsymbol{\theta}_g}(x))f_{i,\boldsymbol{\theta}_g}(x)dx \quad + \\ \int \Big(-\nabla_{\boldsymbol{\theta}} u_{i,\boldsymbol{\theta}_g}(x) - u_{i,\boldsymbol{\theta}_g}(x)u^T_{i,\boldsymbol{\theta}_g}(x)h_i(x)\Big)\Big(g_i(x) - f_{i,\boldsymbol{\theta}_g}(x)\Big)w(f_{i,\boldsymbol{\theta}_g}(x))dx, \tag{12}$$

where $w(t) = B''(t)\times t$, w' is the derivative of w w.r.t. its argument and $h_i(t) = (w'(f_{i,\boldsymbol{\theta}}(t))f_{i,\boldsymbol{\theta}}(t))\,/w(f_{i,\boldsymbol{\theta}}(t))$. Similarly, the matrix defined in Equation (11), has the expression

$$\Omega_n = \frac{1}{n}\sum_{i=1}^{n} \int \Big(u_{i,\boldsymbol{\theta}_g}(x)u^T_{i,\boldsymbol{\theta}_g}(x)w^2(f_{i,\boldsymbol{\theta}_g}(x))g_i(x)dx - \xi_i\xi_i^T\Big), \tag{13}$$

where $\xi_i = \int u_{i,\boldsymbol{\theta}_g}(x)w(f_{i,\boldsymbol{\theta}_g}(x))g_i(x)dx$. We will make the following assumptions to establish the asymptotic properties of the minimum Bregman divergence estimators. These are in the spirit of Basu et al. (1998), with appropriate modifications to cover the general independent but non-homogeneous data case.

Assumption (A1): The support $\chi = \{x : f_{i,\boldsymbol{\theta}}(x) > 0\}$ is independent of i and $\boldsymbol{\theta}$ for all i; the true densities g_i are also supported on χ for all i.

Assumption (A2): There is an open subset ω of the parameter space Ω, containing the best fitting parameter $\boldsymbol{\theta}_g$ such that for almost all $x \in \chi$, and all $\boldsymbol{\theta} \in \Omega$, all $i = 1, 2, \ldots$, the density $f_{i,\boldsymbol{\theta}}(x)$ is thrice differentiable with respect to $\boldsymbol{\theta}$ and the third partial derivatives are continuous with respect to $\boldsymbol{\theta}$.

Assumption (A3): For $i = 1, 2, \ldots$ the integrals $\int[f_{i,\boldsymbol{\theta}}(x)B'(f_{i,\boldsymbol{\theta}}(x)) - B(f_{i,\boldsymbol{\theta}}(x))]dx$ and $\int B'(f_{i,\boldsymbol{\theta}}(x))g_i(x)dx$ can be differentiated thrice with respect to $\boldsymbol{\theta}$, and the derivatives can be taken under the integral sign.

Assumption (A4): For each $i = 1, 2, \ldots$ the matrices $J^{(i)}$ are positive definite and $\lambda_0 = \inf_n$ [min eigenvalue of Ψ_n] is positive.

Assumption (A5): There exists a function $M^{(i)}_{jkl}(x)$ such that$\mid \nabla_{jkl}V_{i,\boldsymbol{\theta}}(x) \mid \leq M^{(i)}_{jkl}(x)\ \forall\ \boldsymbol{\theta} \in \Omega$ and $\forall\ i = 1, 2, \ldots$ where $n^{-1}\sum_{i=1}^n E_{g_i}[M^{(i)}_{jkl}(X_i)] = O(1)\ \forall\ j, k, l$.

Assumption (A6): For all j, k we have

$$\lim_{N\to\infty}\sup_n \left\{\frac{1}{n}\sum_{i=1}^{n} E_{g_i}\Big(\mid \nabla_j V_{i,\boldsymbol{\theta}}(X_i) \mid I(\mid \nabla_j V_{i,\boldsymbol{\theta}}(X_i) \mid > N)\Big)\right\} = 0.$$

$$\lim_{N\to\infty}\sup_n \left\{\frac{1}{n}\sum_{i=1}^{n} E_{g_i}\Big(\big(\mid \nabla_{jk} V_{i,\boldsymbol{\theta}}(X_i) - E_{g_i}(\nabla_{jk}V_{i,\boldsymbol{\theta}}(X_i)) \mid \big) \times \right. \\ \left. I\big(\mid \nabla_{jk}V_{i,\boldsymbol{\theta}}(X_i) - E_{g_i}(\nabla_{jk}V_{i,\boldsymbol{\theta}}(X_i)) \mid > N\big)\Big)\right\} = 0.$$

where $I(A)$ denotes the indicator variable associated with the event A.

Assumption (A7): For all $\epsilon > 0$, we have

$$\lim_{n\to\infty}\left\{\sum_{i=1}^{n} E_{g_i}\left(\parallel \Omega_n^{-1/2}\nabla_{\boldsymbol{\theta}} V_{i,\boldsymbol{\theta}}(X_i) \parallel^2 I(\parallel \Omega_n^{-1/2}\nabla_{\boldsymbol{\theta}} V_{i,\boldsymbol{\theta}}(X_i) \parallel > \ \epsilon\sqrt{n}\right)\right\} = 0.$$

Theorem 1 *Under assumptions (A1) - (A7), the following results hold.*

1. *There exists a consistent sequence $\hat{\boldsymbol{\theta}}_n$ of roots satisfying the minimum Bregman divergence estimating equation given by Equation (3).*
2. *The asymptotic distribution of $\Omega_n^{-1/2}\Psi_n[\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_g)]$ is s-dimensional normal with mean vector $\mathbf{0}$ and covariance I_s, the s-dimensional identity matrix.*

In the interest of brevity, we do not present the proof here. Instead, the proof is presented in Section 2 of the supplement.

Remark 1 Setting $f_i = f \ \forall i$, we get the corresponding asymptotic properties of the minimum Bregman divergence estimator for i.i.d. case as given in Basu et al. (1998). In particular, choosing $B(x) = x^{1+\alpha}/\alpha$, we recover the asymptotic distribution of MDPDE(α). Assumptions (A1) - (A5) are similar to those given by Basu et al. (1998) while assumptions (A6) and (A7) are automatically satisfied by the dominated convergence theorem for i.i.d. data.

3.3 Influence function analysis

Let G_i (g_i) be the true distribution (density) for Y_i, $i = 1, 2, \dots, n$ and $T_B(G_1, G_2, \dots, G_n)$ be the minimum Bregman divergence functional obtained, under appropriate regularity conditions, as the solution of the system given by Equation (3). To derive the influence function of the minimum Bregman divergence estimator in the context of non-homogeneous data, we consider the set of contaminated distributions $G_{i,\epsilon} = (1-\epsilon)G_i + \epsilon\Delta(t_i)$, where $\Delta(t_i)$ is the degenerate distribution at the point of contamination t_i ($i = 1, 2, \dots, n$) and $0 < \epsilon < 1$. The associated contaminated density is given by $g_{i,\epsilon}$. Let $\boldsymbol{\theta}_0 = T_B(G_1, G_2, \dots, G_n)$ and let $\boldsymbol{\theta}_\epsilon^i = T_B(G_1, G_2, \dots, G_{i-1}, G_{i,\epsilon}, G_{i+1}, \dots, G_n)$ be the minimum Bregman divergence functional with contamination only in the i-th direction. Substituting $\boldsymbol{\theta}_\epsilon^i$ and $g_{i,\epsilon}$ in the estimating equations (3), differentiating with respect to ϵ and evaluating the derivative at $\epsilon = 0$, we obtain the influence function of the functional which considers contamination only along the i-th direction to be

$$IF_i(t_i, T_B, G_1, G_2, \dots, G_n) = \frac{\Psi_n^{-1}}{n}\Big[u_{i,\boldsymbol{\theta}}(t_i)w(f_{i,\boldsymbol{\theta}}(t_i)) - \xi_i\Big],$$

where Ψ_n is as defined in Equation (10), $w(t) = B''(t) \times t$ and $\xi_i = \int u_{i,\boldsymbol{\theta}}(x) w(f_{i,\boldsymbol{\theta}}(x)) g_i(x) dx$. Letting $\boldsymbol{\theta}_\epsilon = T_B(G_{1,\epsilon}, G_{2,\epsilon}, \ldots, G_{n,\epsilon})$ and proceeding similarly, we obtain the influence function with contamination at all the data points as

$$IF(t_1, \ldots, t_n, T_B, G_1, \ldots, G_n) = \frac{\Psi_n^{-1}}{n} \sum_{i=1}^{n} \Big[u_{i,\boldsymbol{\theta}}(t_i) w(f_{i,\boldsymbol{\theta}}(t_i)) - \xi_i \Big]. \qquad (14)$$

In particular, letting $t_i = t$, $G_i = G$ and $f_i = f$, $\forall i$, we get back the influence function of the minimum Bregman divergence estimator for the i.i.d. case. In particular, for the MEWDE(β), the influence function reduces to

$$IF(t, T_\beta, G) = J^{-1} \Big[u_{\boldsymbol{\theta}}(t) \big[1 - \exp(f_{\boldsymbol{\theta}}(t)/\beta) \big] - \xi \Big], \qquad (15)$$

where $J = \int u_{\boldsymbol{\theta}}(x) u_{\boldsymbol{\theta}}^T(x) [1 - \exp(-f_{\boldsymbol{\theta}}(x)/\beta)] f_{\boldsymbol{\theta}}(x) dx$ and $\xi = \int u_{\boldsymbol{\theta}}(x) [1 - \exp(-f_{\boldsymbol{\theta}}(x)/\beta)] f_{\boldsymbol{\theta}}(x) dx$. Figure 3 describes the theoretical influence function for $\hat{\mu}(\beta)$ – the MEWDE(β) functional for the mean parameter of the contaminated normal distribution $(1-\epsilon)N(\mu, 1) + \epsilon\Delta_t$ for various contamination values t. For all $\beta > 0$ considered here, we note their bounded and redescending nature.

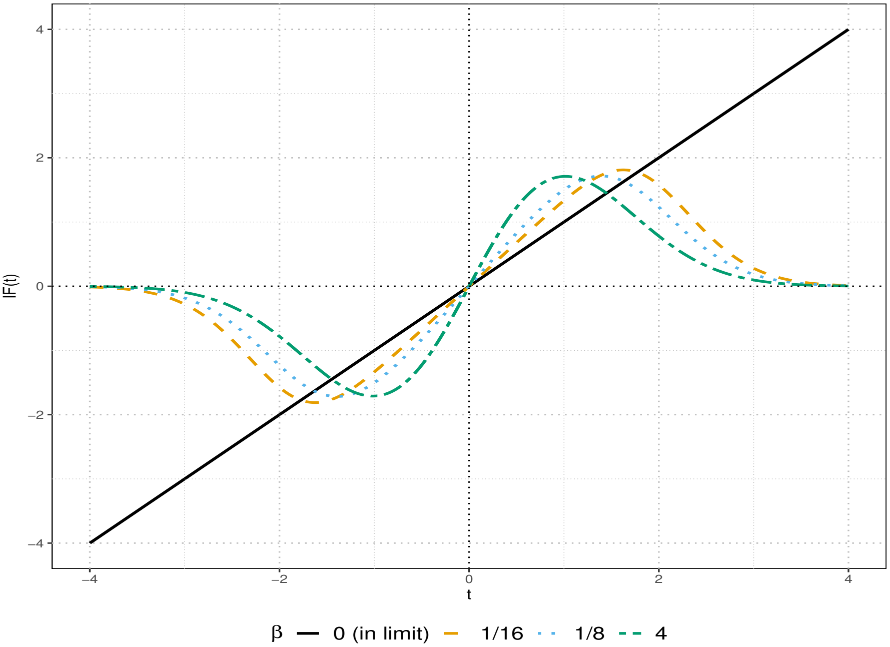

Fig. 3 Influence function for $\hat{\mu}$ from the contaminated normal distribution $(1-\epsilon)N(\mu, 1) + \epsilon\Delta_t$ for some MEWDE(β).

4 Simulation studies for MEWDE(β)

4.1 Introduction

For i.i.d. data, when the true density g belongs to the model, i.e. $g = f_{\boldsymbol{\theta}}$ for some $\boldsymbol{\theta} \in \Omega$, let $\hat{\boldsymbol{\theta}}$ refer to the MEWDE of an unknown parameter $\boldsymbol{\theta}$. For $\beta > 0$, the asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is an s- variate normal distribution with mean vector $\mathbf{0}$ and dispersion matrix given by $J^{-1}KJ^{-1}$, where

$$\begin{aligned} J &= \int u_{\boldsymbol{\theta}}(x)u_{\boldsymbol{\theta}}^T(x)[1 - \exp(-f_{\boldsymbol{\theta}}(x)/\beta)]f_{\boldsymbol{\theta}}(x)dx, \\ K &= \int u_{\boldsymbol{\theta}}(x)u_{\boldsymbol{\theta}}^T(x)[1 - \exp(-f_{\boldsymbol{\theta}}(x)/\beta)]^2 f_{\boldsymbol{\theta}}(x)dx \quad - \quad \xi\xi^T, \\ \xi &= \int u_{\boldsymbol{\theta}}(x)[1 - \exp(-f_{\boldsymbol{\theta}}(x)/\beta)]f_{\boldsymbol{\theta}}(x)dx. \end{aligned} \tag{16}$$

As $\beta \to 0$, both J and K reduce to the Fisher information matrix. We now consider different parametric families and compare the performance of MEWDE's and MDPDE's under different contamination scenarios.

4.2 Simulation scheme

First we compute, for a fixed parametric family of densities $\mathscr{F}_{\boldsymbol{\theta}} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Omega \subseteq \mathrm{R}^s\}$, the empirical mean square errors (MSEs) of the parameter estimates – for several members of both the MDPDE and the MEWDE classes – under pure data generated from the given parametric model. Then we identify several sets of combinations (α, β), the tuning parameters of the two families, for which the empirical MSEs of MDPDE(α) and MEWDE(β) are approximately equal. Subsequently we generate data from contaminated model distributions having densities of the form $g(x) = (1 - \epsilon)f_{\boldsymbol{\theta}_0}(x) + \epsilon v(x)$, where ϵ is the contaminating proportion, $v(x)$ is a suitable contaminating density, but $\boldsymbol{\theta}_0$ is still the target parameter. We compare the MSEs (against $\boldsymbol{\theta}_0$) of MDPDE (α) and MEWDE(β) to determine which one has better outlier stability. Unless otherwise specified, we have used samples of size $n = 200$, with $r = 2000$ replications for each scenario. The finite sample relative efficiency (FSRE) of the MDPDE is defined to be the ratio of MSE(MLE) to MSE(MDPDE); similarly for the MEWDE.

4.3 Results

We consider three separate simulation designs involving (a) estimation of the mean of a univariate normal distribution with known standard deviation, (b) estimation of the standard deviation of a univariate normal distribution with known mean and (c) estimation of the mean parameter of an exponential distribution.

For simulation (a), the true distribution is taken to be $N(0, 1)$ and the contaminating distribution is $N(\mu_c, 1)$. We run simulations for $\mu_c = 3$ and 5 and estimate the mean under the $N(\mu, 1)$ model. Our findings are presented in Table 1. We note that for uncontaminated data, the MLE is the most efficient estimator, as it should be. However, even under a small contamination there is a severe degradation in performance of the MLE, and the other two estimators quickly overtake it. As the proportion of contamination increases, larger tuning parameters give better performance (on account of their stronger downweighting). However, this improvement is not absolute. Generally, with increasing tuning parameter, the performance of the estimators reach a peak at some point, and thereafter drops again. Another point of interest is when the contaminating component is far separated from the target distribution, smaller parameters suffice to provide good outlier stability. However, the most important observation is that in all the identified pairs having comparable pure data MSE, the MEWDE beats the MDPDE, sometimes quite soundly, under contaminated scenarios. In Supplementary Figures 3 and 4 we graphically present the MSEs of different members of the MEWDE(β) class for the indicated pure normal data and contaminated normal data situations over a sequence of sample sizes.

Table 1 FSRE's of MDPDE(denoted D(α)) and MEWDE (denoted EWD(β)). Figures in bold denote best FSRE in that contamination scheme. The minimum L_2 ($D(1)$) case is provided for comparison.

	$\mu_c = 3$				$\mu_c = 5$		
$\hat{\mu}$	$\epsilon = 0$	$\epsilon = 0.05$	$\epsilon = 0.10$	$\epsilon = 0.20$	$\epsilon = 0.05$	$\epsilon = 0.10$	$\epsilon = 0.20$
MLE	**1**	1	1	1	1	1	1
D(0.05)	0.996	1.358	1.358	1.250	2.635	2.568	2.059
E(0.001)	0.996	1.791	1.806	1.495	12.326	31.141	52.727
D(0.1)	0.956	2.567	3.027	2.552	10.966	23.342	29.168
E(0.004)	0.954	3.409	4.863	4.221	**13.509**	**43.633**	**140.125**
D(0.43)	0.871	3.592	6.075	6.213	12.106	38.450	115.779
E(0.063)	0.867	**4.003**	7.947	9.664	12.356	40.769	137.303
D(0.74)	0.749	3.693	8.567	13.500	10.495	34.680	117.038
E(0.25)	0.747	3.763	**9.075**	15.557	10.525	34.861	118.461
D(0.98)	0.666	3.428	8.837	17.716	9.304	30.871	105.076
E(4)	0.666	3.430	8.861	17.852	9.304	30.871	105.129
L_2	0.659	3.401	8.821	**17.977**	9.206	30.553	104.007

Since our findings for simulations (b) and (c) are similar, in the interest of conciseness, we present our findings for (a) in the main paper and for (b) and (c) in Supplementary Tables 1 and 2. In all three cases, we note that the MEWDE class provides convincing improvements over the MDPDE class.

5 Modeling real data

5.1 Shoshoni rectangles

We consider the data on Shoshoni rectangles presented and analyzed by Hettmansperger and McKean (2010) (see Supplementary Table 3). The histogram of the observations (Supplementary Table 5) indicates that three out of the twenty observations (colored in red) are well-separated from the 'main body' of the remaining observations. Supplementary Figure 6 of the full data (suitably centered and scaled) Q-Q plot supports the claim that the 3 largest observations are possible outliers. In contrast, a study of the Q-Q plot generated from the 'outlier deleted' data (again suitably centered and scaled), motivates us to model the dataset using a normal distribution with unknown mean μ and standard deviation σ (i.e., $\boldsymbol{\theta} = (\mu, \sigma)^T$). Indicating the full data maximum likelihood estimates by ML and the outlier deleted ones by ML+D, we get, under the normal model, $\hat{\mu}_{ML} = 0.660$ and $\hat{\sigma}_{ML} = 0.093$; the outlier deleted estimates are $\hat{\mu}_{ML+D} = 0.628$ and $\hat{\sigma}_{ML+D} = 0.043$. Thus the outliers have a moderate efect on the mean, but a substantial effect on the scale parameter. However all the different tuning parameters β for EWD used in this case produce stable estimators and outlier resistant fits to the full data. As the outliers are quite distant from the majority of the data, small values of β appear to sufficient in either case. Figures 4 and 5 describe how the kernel density estimate and normal density curves based on the MEWDEs and the MDPDEs, respectively fit the data.

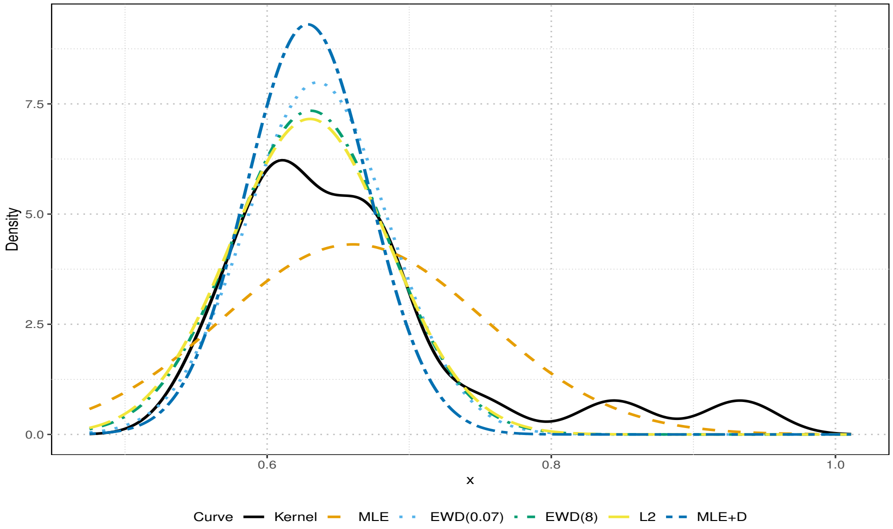

Fig. 4 Density estimates for Shoshoni rectangles using MEWDEs.

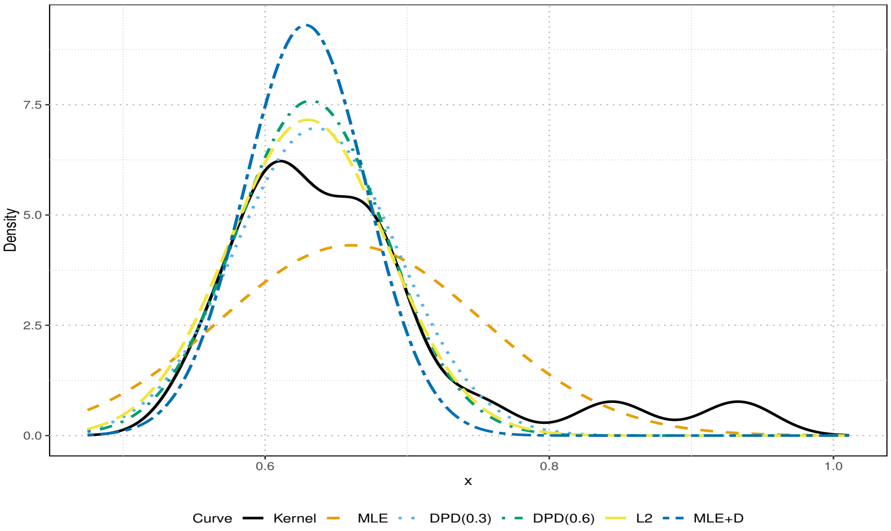

Fig. 5 Density estimates for Shoshoni rectangles using MDPDEs.

5.2 Drosophila offspring counts

We compare the performance of the MEWDE and MDPDE in the context of data on fruit flies (see Woodruff, Mason, Valencia, and Zimmering (1984)). In this experiment male flies were sprayed with a chemical, and then made to mate with unexposed females. The response, for each father fly, was the number of daughter flies having a recessive lethal mutation in the X-chromosome.

Table 2 Fitted frequencies for Drosophila data using MLE, MDPDE(denoted by D(α)) and MEWDE(denoted by E(β)), where α and β are tuning parameters for MDPDE and MEWDE respectively.

Count	0	1	2	3	4	≥ 5	$\hat{\lambda}$
Observed	23	7	3	0	0	1 (91)	–
MLE.	1.596	4.882	7.467	7.613	5.822	6.620	3.059
D(0.10)	22.981	9.002	1.763	0.230	0.023	0.002	0.392
D(0.50)	23.375	8.759	1.641	0.205	0.019	0.002	0.375
D(0.75)	23.549	8.649	1.588	0.194	0.018	0.001	0.367
E(0.001)	22.894	9.055	1.791	0.236	0.023	0.002	0.396
E(0.02)	22.614	9.222	1.880	0.256	0.026	0.002	0.408
E(0.25)	23.712	8.545	1.540	0.185	0.017	0.001	0.360
L_2	23.609	8.611	1.570	0.191	0.017	0.001	0.365
MLE + D	22.93	9.03	1.78	0.23	0.02	0	0.39

The frequencies of these responses (presented in the first row of Table 2) are modeled as Poisson variables, and the estimates of the Poisson mean parameter λ (as well as the estimated frequencies), using several members of the DPD and EWD families, the MLE and the MLE+D (outlier deleted MLE) are presented in Table 2. The single extreme value at 91 is treated as the obvious outlier. Both sets of estimators have comparable (satisfactory) performance, which adequately discount the outlier.

5.3 Homicide from firearm use and GDP

We consider modeling age-standardized national firearm-related homicide rates in 23 Western countries as a function of per-capita gross domestic product as of 2017. Country-specific information on GDP (from The CIA World Factbook (n.d.)) and data on firearm-related homicide rates (from Roser and Ritchie (2020)) were obtained. Figure 6 is a scatter-plot of the data set described, where the independent variable per-capita gross domestic product is plotted on the X-axis, and firearm related homicide rate on the Y-axis. The USA has an abnormally high firearm related homicide rate in relation to its per-capita GDP, and this single outlier forces the least squares regression line to have a positive slope, which clearly contradicts the general configuration of points. In comparison, the two MEWDE fits show a clear reversal in slope, and give more satisfactory descriptions of the rest of the data, sacrificing the large outlier. Table 3 shows how estimated coefficients vary as we change the tuning parameter β. We observe that for very small β, the MEWDE almost mimics the outlier-deleted maximum likelihood estimator (MLE+D), implying that

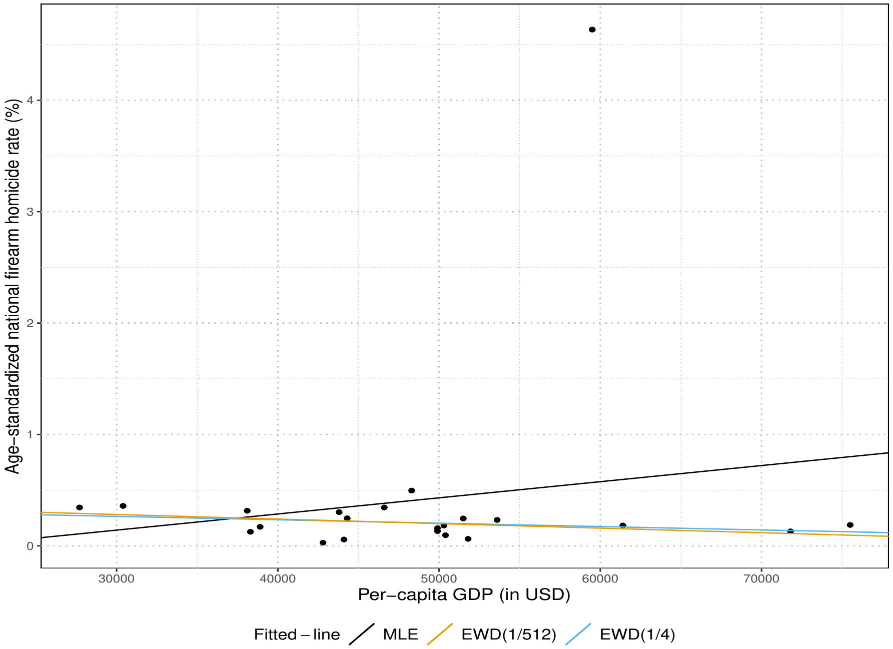

Fig. 6 Modeling firearm-related homicide rates in Western countries as a function of per-capita gross domestic product: fits with MLE and MEWDE.

Table 3 Estimated regression parameters for homicide data.

Estimates	MLE	E(0.002)	E(0.02)	E(0.25)	E(1)	MLE+D
Intercept	−0.293	0.356	0.356	0.404	0.359	0.356
GDP ($\times 10^{-6}$)	14.49	−3.045	−3.042	−4.087	−3.250	−3.042
Error s.d.	0.959	0.110	0.111	0.131	0.106	0.088

the MEWDE fits the data well by automatically downweighting the outlier, even for very small values of β.

5.4 Solubility of alcohols in water

We consider fitting a multiple linear regression model to the dataset concerning alcohol solubility in water (see Maronna, Martin, Yohai, and Salibián-Barrera (2019)). The dataset gives, for 44 aliphatic alcohols, the logarithm of their solubility together with three physicochemical characteristics (namely, solvent accessible surface-bounded molecular volume (SAG), mass and volume). The interest is in predicting the solubility. Following the authors' suggestion of fitting an MM regression-based model to the data, we observe that four data points (roughly 10% of the data set) are assigned much smaller 'robustness weights' as compared to the remaining 40 data points. Treating these four observations as outliers, we obtain the outlier-deleted maximum likelihood estimates (denoted by MLE+D) of the regression coefficients and error standard deviation. We also compute the robust LMS estimate. In order to estimate the error s.d. σ, we compute $\hat{\sigma} = \text{median}|r_i - \text{median}(r_i)|/0.67449$. Finally, we compute minimum EWD(β) regression parameter estimates for various values of β. Our findings are presented in Table 4. As a visual inspection is not possible for the fits in this multiple linear regression model, the coefficients of Table 4 are not sufficient on their own to give a full idea about how stable and outlier-resistant the fits are. By means of Supplementary Figures 7 and 8, we examine the residuals of each of the fits for the data and explore how well they fare in terms of separating out the outliers. Supplementary Figures 7 and 8 demonstrate a particular MEWDE and the LMS estimate are much more successful in making the outliers stand out and giving stable fits compared to the least squares fit provided by the MLE.

Table 4 Estimated regression parameters for alcohol solubility data

Estimates	MLE	LMS	E(0.1)	E(0.4)	E(0.7)	MLE+D
Intercept	8.777	3.617	5.883	3.974	5.444	6.829
SAG	0.014	0.177	0.110	0.163	0.129	0.077
Volume	−0.040	−0.191	−0.133	−0.179	−0.152	−0.102
Mass	0.027	0.248	0.172	0.235	0.206	0.127
Error s.d.	0.504	0.405	0.372	0.145	0.221	0.389

6 Tuning parameter selection

In minimum EWD estimation, small values of β provide greater model efficiency, while large values of β provide greater outlier stability and protection against small model violations. Given any real data set we must choose the 'optimal', data-based tuning parameter β so that the procedure has the right amount of balance for the data set in question. We follow the approach of Warwick and Jones (2005) to derive the optimal estimate of the tuning parameter. This approach constructs an empirical estimate of the mean square error as a function of the tuning parameter β and a pilot estimator $\boldsymbol{\theta}^P$ given by

$$\begin{aligned}\widehat{MSE}_\beta(\boldsymbol{\theta}^P) = \left(\widehat{\boldsymbol{\theta}}_\beta - \boldsymbol{\theta}^P\right)^T \left(\widehat{\boldsymbol{\theta}}_\beta - \boldsymbol{\theta}^P\right) + \\ \frac{1}{n}\mathrm{tr}\left(\Psi_n^{-1}\left(\widehat{\boldsymbol{\theta}}_\beta\right)\Omega_n\left(\widehat{\boldsymbol{\theta}}_\beta\right)\Psi_n^{-1}\left(\widehat{\boldsymbol{\theta}}_\beta\right)\right),\end{aligned} \tag{17}$$

where $\mathrm{tr}(\cdot)$ is the trace of a matrix and Ψ_n and Ω_n are the matrices defined in Equations (10) and (11), respectively, evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_\beta = \mathrm{MEWDE}(\beta)$. Further, $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. By minimizing the objective function given in Equation (17) over $\beta > 0$, we get a data driven 'optimal' estimate of the tuning parameter. Warwick and Jones (2005) propose the minimum L_2 estimator as the pilot estimator in the above calculation. Through this procedure, the optimal tuning parameter for the data on Shoshoni rectangles is found to be $\beta_{\mathrm{OPT}} = 0.43$, and the associated estimated parameters are $\hat{\mu} = 0.63$ and $\hat{\sigma} = 0.05$. The corresponding(sample size-scaled) asymptotic mean-squared error is 5.07×10^{-3}. For the data on Drosophila fruit flies, the optimal tuning parameter is found to be $\beta_{OPT} = 0.08$, and the estimated mean parameter is $\hat{\lambda} = 0.377$. The corresponding (sample size-scaled) asymptotic mean-squared error is 0.46. For the data on firearm-related homicide and GDP, the optimal tuning parameter is $\beta_{OPT} = 1.6$, and the corresponding estimated regression parameters (intercept, GDP, error standard deviation) are $(0.416, -3.932 \times 10^{-6}, 0.066)$. Finally, for the data on alcohol solubility, the optimal tuning parameter is $\beta_{OPT} = 0.66$, and the corresponding estimated regression parameters (intercept, SAG, Volume, Mass, error standard deviation) are $(6.084, 0.112, -0.135, 0.174, 0.062)$.

7 Testing of hypotheses

7.1 Introduction

Here we develop general robust tests of hypothesis based on Bregman divergences. This generalizes the works of Basu et al. (2013, 2018). We establish the asymptotic null distribution of the proposed test statistic and apply the theory developed to a real-life data set. Our focus will remain on EWD(β).

Unlike the previous sections, we will consider the case of i.i.d. data only. We begin with an identifiable parametric family of probability measures $P_{\boldsymbol{\theta}}$

on a measurable space $\{\chi, \mathscr{A}\}$ with an open parameter space $\Omega \subseteq \mathbb{R}^s,\ s \geq 1$. Measures $P_{\boldsymbol{\theta}}$ are described by densities $f_{\boldsymbol{\theta}} = dP_{\boldsymbol{\theta}}/d\mu$, absolutely continuous with respect to a dominating σ-finite measure μ on χ. We have an i.i.d. sample of size n given by $X_1, X_2, \ldots, X_n$ from a density belonging to the family $\mathscr{F}_{\boldsymbol{\theta}} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Omega\}$. We will assume that the support of the distribution is independent of $\boldsymbol{\theta}$. The hypothesis of interest is $H_0 : \boldsymbol{\theta} \in \Omega_0$ against $H_1 : \boldsymbol{\theta} \notin \Omega_0$. We use the common approach where the restricted parameter space specified by H_0 can be rewritten by a set of $r < s$ restrictions of the form

$$\boldsymbol{m(\theta)} = \mathbf{0}_r \tag{18}$$

on Ω, where $\boldsymbol{m} : \mathbb{R}^s \rightarrow \mathbb{R}^r$ is a vector valued function such that the $s \times r$ matrix $M(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{m^T(\theta)}}{\partial \boldsymbol{\theta}}$ exists and is continuous in $\boldsymbol{\theta}$ and $\text{rank}(M(\boldsymbol{\theta})) = r$.

7.2 The test statistic

To perform this test of hypothesis, we first obtain $\hat{\boldsymbol{\theta}}_{B_1}$, the unrestricted minimum Bregman divergence estimator for a given B_1 function and then obtain the restricted estimator $\tilde{\boldsymbol{\theta}}_{B_1}$, subject to the constraints of Equation (18). We then examine the family of Bregman divergence test statistics (BDTS)

$$T_{B_2}(\hat{\boldsymbol{\theta}}_{B_1}, \tilde{\boldsymbol{\theta}}_{B_1}) = 2n \times D_{B_2}(f_{\hat{\boldsymbol{\theta}}_{B_1}}, f_{\tilde{\boldsymbol{\theta}}_{B_1}}), \tag{19}$$

where $D_{B_2}(g, f)$ is the Bregman divergence between two densities g and f, defined in Equation (1) with B_2 as the B function. We will consider the functions B_1 and B_2 to have the same functional form (eg., as given by the exponentially weighted divergences) only differing, if at all, in the values of their tuning parameters. Thus, in the derivation of the asymptotic distribution of the test statistics, the functions B_1 and B_2 are allowed to be different. In practice, however, a single, suitably chosen common function B will generally work well in most cases.

In addition to assumptions (B1) to (B5) presented in Section 3.4 of the supplement, we make the following assumption

Assumption (B6): For all $\boldsymbol{\theta} \in \omega$, the partial derivatives $\partial^2 m_l(\boldsymbol{\theta})/\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_k$ are bounded for all j, k and l, where $m_l(\cdot)$ is the l-th element of $\boldsymbol{m}(\cdot)$.

Theorem 2 *Under assumptions (B1) - (B6), and assuming that the true distribution belongs to the model, i.e., $G = F_{\boldsymbol{\theta}_g}$ for some $\boldsymbol{\theta}_g \in \Omega$ which satisfies the set of constraints given by Equation (18), the constrained minimum Bregman divergence estimator $\tilde{\boldsymbol{\theta}}_{n,B_1}$ has the following properties: the underlying B function is denoted by B_1.*

1. *Consistency: $\tilde{\boldsymbol{\theta}}_{n,B_1} \xrightarrow{P} \boldsymbol{\theta_g}$ as $n \to \infty$.*
2. *Asymptotic normality: The asymptotic null distribution of $\sqrt{n}(\tilde{\boldsymbol{\theta}}_{n,B_1} - \boldsymbol{\theta}_g)$ is s-dimensional multivariate normal with the zero mean vector and an $s \times s$ dispersion matrix $\Sigma_{B_1} = P_{B_1} K_{B_1} P_{B_1}$, where K_{B_1} is the i.i.d. analogue of*

the Ω_n matrix defined by Equation (11) with B_1 serving as the B function. The P_{B_1} matrix is defined as $P_{B_1} = J_{B_1}^{-1} - Q_{B_1} M J_{B_1}^{-1}$, where J_{B_1} is the i.i.d. analogue of the Ψ_n matrix defined by Equation (10). Further, $Q = J_{B_1}^{-1} M [M^T J_{B_1}^{-1} M]^{-1}$; $M = M(\boldsymbol{\theta})$ is as defined in Section (7.1).

An outline for the proof is presented in Section 4 of the supplement. Theorem 3 presents the asymptotic distribution of the test statistic defined in Equation (19).

Remark 2 Theorem 2 extends Theorem 1 in the context of i.i.d. data. Under the setup of Theorem 1, M becomes a null matrix and consequently, $P_{B_1} = J_{B_1}^{-1}$ and the asymptotic dispersion matrix of the unrestricted minimum Bregman divergence estimator assumes the form specified by Theorem 1.

Theorem 3 *Under assumptions (B1) - (B6), the asymptotic distribution of the test statistic defined in Equation (19) is identical with, under the null hypothesis specified in Equation (18), the distribution of the random variable*

$$\sum_{i=1}^{k} \lambda_i(B_1, B_2, \boldsymbol{\theta}) Z_i^2, \tag{20}$$

where $Z_1, \ldots, Z_k$ are independent standard normal variables and $\lambda_i(B_1, B_2, \boldsymbol{\theta})$ for $i = 1, \ldots, k$ are the nonzero eigenvalues of the matrix $A_{B_2} B_{B_1} K_{B_1} B_{B_1}$, and k is the rank of the matrix $B_{B_1} K_{B_1} B_{B_1} A_{B_2} B_{B_1} K_{B_1} B_{B_1}$. The (i,j)-th element of A_{B_2} is defined as follows

$$A_{B_2}(i,j) = \int B_2''(f_{\boldsymbol{\theta_g}}(x)) \left[\frac{\partial f_{\boldsymbol{\theta_g}}(x)}{\partial \boldsymbol{\theta}_i} \frac{\partial f_{\boldsymbol{\theta_g}}(x)}{\partial \boldsymbol{\theta}_j} \right] dx$$

and the matrix B_{B_1} is equal to

$$J_{B_1}^{-1} M [M^T J_{B_1}^{-1} M]^{-1} M^T J_{B_1}^{-1}.$$

An outline of the proof is presented in Section 5 of the supplement.

Remark 3 We note that the ranks of $B_{B_1} K_{B_1} B_{B_1} A_{B_2} B_{B_1} K_{B_1} B_{B_1}$, $B_{B_1} K_{B_1} B_{B_1}$ and M are simultaneously equal to r.

Remark 4 An easy way to approximate the required critical region of the above test is outlined here. From Theorem 3 it is obvious that the k eigenvalues described are functions of $\boldsymbol{\theta}_g$. Under the null, they can be consistently estimated by plugging in $\tilde{\boldsymbol{\theta}}_{B_1}$ in place of $\boldsymbol{\theta}_g$. Let these estimated eigenvalues be $\hat{\lambda}_1, \cdots, \hat{\lambda}_k$. Generating k independent observations $Z_1, \cdots, Z_k$ from the $N(0,1)$ distribution, one can approximate the quantiles of $\sum_i \hat{\lambda}_i Z_i^2$ by replicating this procedure a large number of times, which can then serve as consistent estimates of the quantiles of the limiting variable in Equation (20).

Remark 5 An approximate form of the power function of the test statistic can be obtained by following the steps outlined by Theorem 7 of Basu et al. (2018).

7.3 An Example: Testing of Hypothesis for the Shoshoni Data

The Shoshoni data, considered earlier in Section 5.1 are assumed to come from a $N(\mu, \sigma^2)$ distribution with both location and scale parameters being unknown. Hettmansperger and McKean (2010) note that if we were to implement the outlier-sensitive likelihood ratio test for the hypothesis

$$H_0 : \mu = 0.618 \text{ versus } H_1 : \mu \neq 0.618, \tag{21}$$

we would get a p-value of 0.053, which is at the borderline of significance at 5% level. On the other hand, the non-parametric one-sample sign test returns an entirely insignificant p value of 0.823. It would be interesting to investigate the performance of the robust test statistic presented in Section 7.2 in this context. In particular, we focus on the test statistic obtained by setting $B_1 = B_2 = B$, where B is the function corresponding to the exponentially weighted divergence, as defined in Equation (5). We term the test statistic so obtained as the exponentially weighted divergence test statistic (denoted by EWDTS(β)) and compute it for varying values of $\beta > 0$. For $\beta \to 0$, the test resembles the standard likelihood ratio test and returns a p-value slightly exceeding the 0.05 threshold, as reported by Hettmansperger and McKean (2010). As β increases, the p-value quickly becomes highly insignificant, indicating that the borderline significance under the likelihood based methods is driven by the outliers. Note that the p-value for the t-test statistic for the data with three outliers removed is 0.329, which conforms to the p-values of to the EWDTS for moderately large values of β. The graph of the p-values (Figure 7) demonstrate the outlier stability of EWDTS for large values of β.

8 Conclusion

In this paper, we have presented an estimator based on a sub-class of density-based Bregman divergences, which is seen to be outperforming the existing standard (i.e., the DPD based estimator). We have shown several asymptotic and distributional properties of the proposed estimator, both in the context of i.i.d data as well as independent and non-homogenous data. A special case of linear regression (both simple and multiple) has been explored in the context of real data. We have also discussed 'judicial' choice(s) of the tuning parameter which, when chosen properly, yields highly robust and efficient estimators which can often dominate the MDPDE. We have also considered an hypothesis testing strategy for parametric models which may serve as robust alternatives to the classical likelihood ratio and other likelihood based tests. As we have noted, the weight function generated by EWD converges to 1 as its argument, the value of the density function, increases. We feel that this is the more balanced way for weighting the observations, rather than the weighing provided by the DPD, where the weights increase indefinitely with increase in the value of the density. It may also be mentioned that the proposal based on the EWD

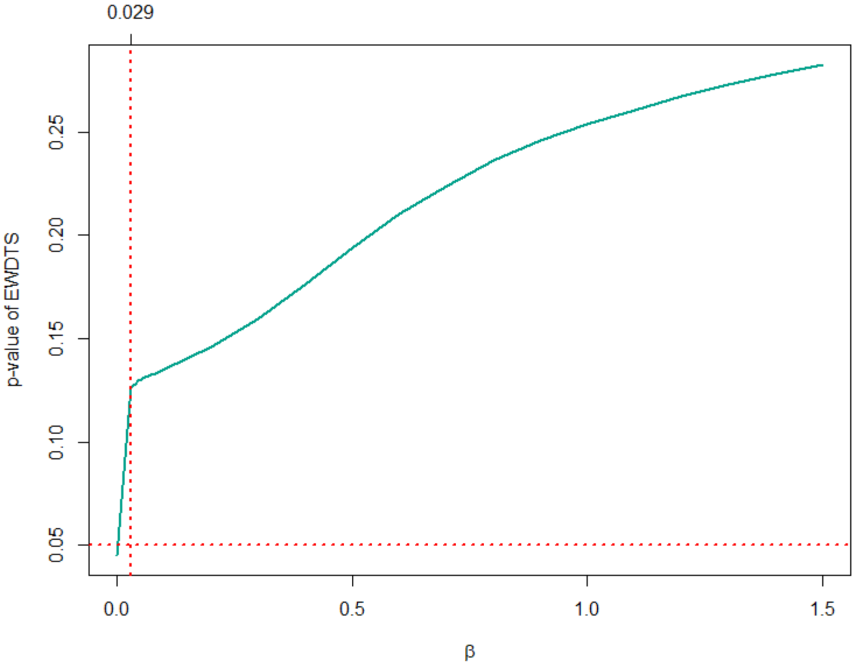

Fig. 7 p-value of EWDTS(β) for $\beta > 0$ for the Shoshoni rectangles dataset.

has the potential to be useful in all the situations where the DPD has been successfully applied, such as generalized linear models, survival analysis and Bayesian inference, to name a few. We hope to pursue all of these in our future research.

In case of hypothesis testing, we have only investigated the analogues of the likelihood-ratio tests. Other Wald-type tests based on the EWD should also be studied which are likely to have simpler asymptotic null distributions compared to that in Equation (20). The DPD based Wald-type test has been extensively used in the literature, and comparisons with EWD based tests will be interesting. We also hope to refine the tuning parameter selection strategy using the recently developed method of Basak, Basu, and Jones (2020).

References

Basak, S., Basu, A., Jones, M. (2020). On the 'optimal' density power divergence tuning parameter. *Journal of Applied Statistics*. Retrieved from https://doi.org/10.1080/02664763.2020.1736524

Basu, A., Harris, I.R., Hjort, N.L., Jones, M. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, *85*(3), 549–559.

Basu, A., Mandal, A., Martin, N., Pardo, L. (2013). Testing statistical hypotheses based on the density power divergence. *Annals of the Institute of Statistical Mathematics*, *65*(2), 319–348.

Basu, A., Mandal, A., Martin, N., Pardo, L. (2018). Testing composite hypothesis based on the density power divergence. *Sankhya B*, *80*(2), 222–262.

Basu, A., Shioya, H., Park, C. (2011). *Statistical Inference: The Minimum Distance Approach.* Chapman and Hall/CRC.

Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *The Annals of Statistics*, *5*(3), 445–463.

Broniatowski, M., Toma, A., Vajda, I. (2012). Decomposable pseudodistances and applications in statistical estimation. *Journal of Statistical Planning and Inference*, *142*(9), 2574–2585.

Csiszár, 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.

Csiszár, I., et al. (1991). Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *The Annals of Statistics*, *19*(4), 2032–2066.

Hettmansperger, T.P., & McKean, J.W. (2010). *Robust nonparametric statistical methods.* CRC Press.

Jana, S., & Basu, A. (2019). A characterization of all single-integral, non-kernel divergence estimators. *IEEE Transactions on Information Theory*, *65*(12), 7976–7984.

Lindsay, B.G. (1994). Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *The Annals of Statistics*, *22*(2), 1081–1114.

Maronna, R.A., Martin, R.D., Yohai, V.J., Salibián-Barrera, M. (2019). *Robust Statistics: Theory and Methods (with R).* John Wiley & Sons.

Pardo, L. (2006). *Statistical inference based on divergence measures.* CRC press.

Roser, M., & Ritchie, H. (2020). Homicides. *Our World in Data.* (https://ourworldindata.org/homicides)

Simpson, D.G. (1989). Hellinger deviance tests: efficiency, breakdown points, and examples. *Journal of the American Statistical Association*, *84*(405), 107–113.

The CIA World Factbook, C.I.A. (n.d.). Country comparison :: Gdp - per capita (ppp). *Central Intelligence Agency.* Retrieved from https://www.cia.gov/library/publications/the-world-factbook/rankorder/2004rank.html

Warwick, J., & Jones, M. (2005). Choosing a robustness tuning parameter. *Journal of Statistical Computation and Simulation*, *75*(7), 581–588.

Woodruff, R., Mason, J., Valencia, R., Zimmering, S. (1984). Chemical mutagenesis testing in drosophila: I. comparison of positive and negative control data for sex-linked recessive lethal mutations and reciprocal translocations in three laboratories. *Environmental mutagenesis*, *6*(2), 189–202.