Electronic Journal of Statistics
Vol. 7 (2013) 2420–2456
ISSN: 1935-7524
DOI: 10.1214/13-EJS847

Robust estimation for independent non-homogeneous observations using density power divergence with applications to linear regression

Abhik Ghosh* **and Ayanendranath Basu**

Bayesian and Interdisciplinary Research Unit
Indian Statistical Institute
203 B. T. Road
Kolkata 700 018, India
e-mail: abhianik@gmail.com; ayanbasu@isical.ac.in

Abstract: In real life we often have to deal with situations where the sampled observations are independent and share common parameters in their distribution but are not identically distributed. While the methods based on maximum likelihood provide canonical approaches for doing statistical inference in such contexts, it carries with it the usual baggage of lack of robustness to small deviations from the assumed conditions. In the present paper we develop a general estimation method for handling such situations based on a minimum distance approach which exploits the robustness properties of the density power divergence measure (Basu et al. 1998 [2]). We establish the asymptotic properties of the proposed estimators, and illustrate the benefits of our method in case of linear regression.

AMS 2000 subject classifications: Primary 62F35; secondary 62J05.
Keywords and phrases: Density power divergence, robustness, linear regression.

Received April 2013.

Contents

*This is part of the Ph.D. research work of the first author which is ongoing at the Indian Statistical Institute

1. Introduction

The standard and basic problem of statistical inference provides the experimenter with a suitably chosen random sample from a distribution of interest which is appropriately modeled by a parametric family and the experimenter has to estimate the unknown parameters and/or perform tests of hypothesis about them. However more complex cases are quite frequent in real life, and often the experimenter is faced with the situation where the observations, although independent, do not have the same distribution. Yet the associated random variables may share a common parameter which might be of interest to us. Depending on the situation, the problem in this case can be quite non-routine and useful methods that can deal with such situations may be of great practical value. Our aim in this paper is to develop a general method of estimation for such problems with particular attention on the robustness issue. We plan to exploit the robustness and the other desirable properties of the density power divergence measure (Basu et al. 1998 [2]) to develop estimators with good robustness properties in this general scenario.

In some cases, of course, such problems have been extensively studied. A simple example in this connection is the linear regression problem with normal errors. The least squares method has long been in use to handle this problem, and the robustness problem of this method has also been recognized for a while. Yet other situations, such the Poisson regression problem, have not been explored nearly to that extent, and there are few, if any, robust techniques available to deal with such problems. We trust that our approach will provide a general technique to handle this and other situations involving independent but non-homogeneous data.

The rest of the paper is organized as follows. In Section 2 we present the proposed approach where we define the minimum density power divergence estimator in case of independent but non-homogeneous observations and the corresponding estimating equation. Section 3 presents the asymptotic properties of the proposed minimum density power divergence estimators. The robustness issue of these estimators is considered in Sections 4 and 5 through the influence function analysis and breakdown results respectively. Section 6 provides the application of the proposed approach in case of linear regression models and some real data regression examples are presented in Section 7. Concluding remarks are presented in Section 8. To avoid interrupting the flow of the article, the proofs of all the results derived in this paper are presented in the Appendix.

2. The minimum density power divergence (DPD) estimator for independent non-homogeneous observations

Basu et al. (1998) [2] introduced the density power divergence family as a measure of discrepancy between two probability density functions and used this family for robustly estimating the model parameter under the usual set up of independent and identically distributed data. The density power divergence measure $d_\alpha(g, f)$ between the densities g and f is defined, as the function of a single tuning parameter α (≥ 0), as

$$d_\alpha(g,f) = \int \left\{ f^{1+\alpha} - \left(1+\frac{1}{\alpha}\right) f^\alpha g + \frac{1}{\alpha} g^{1+\alpha} \right\}, \quad \alpha > 0 \tag{2.1}$$

$$d_0(g,f) = \int g \ln\left(\frac{g}{f}\right). \tag{2.2}$$

Here ln represents the natural logarithm. Basu et al. (1998) [2] demonstrated that the parameter α controls the trade-off between efficiency and robustness of the minimum density power divergence estimator. While the divergence is not defined for $\alpha = 0$, $d_0(\cdot,\cdot)$ represents the divergence obtained in the limit as $\alpha \to 0$; presented in Equation (2.2), this measure is a version of the Kullback-Leibler divergence. On the other hand $\alpha = 1$ generates the squared L_2 distance.

Let G represent the true, data generating distribution, and let g be the corresponding density function. We model the true unknown density function g with the family of densities $\mathcal{F}_\theta = \{f_\theta(x) : \theta \in \Theta \in \mathbb{R}^p\}$; the minimizer of $d_\alpha(g, f_\theta)$ over $\theta \in \Theta$ is the minimum DPD functional at the distribution point G. Notice that the third term of the divergence $d_\alpha(g, f_\theta)$ is independent of θ and hence does not figure in the minimization process; the relevant objective function therefore includes the first two terms only. Suppose now that an independent and identically distributed (i.i.d.) sample $X_1, \ldots, X_n$ is available from the true distribution. The minimum DPD estimator of θ can then be obtained by minimizing

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

over $\theta \in \Theta$. In the above expression the empirical G_n has been used to approximate the relevant theoretical quantity; this allows the experimenter to avoid kernel density estimation and related bandwidth selection issues. This approximation works only because the density g shows up linearly in the middle term of $d_\alpha(g, f)$ in Equation (2.1).

Here we generalize the above concept of robust minimum density power divergence estimation to the case of independent but not identically distributed observations. Let us assume that our observed data $Y_1, \ldots, Y_n$ are independent but for each i, $Y_i \sim g_i$ where $g_1, \ldots, g_n$ are possibly different densities with respect to some common dominating measure. We want to model g_i by the family $\mathcal{F}_{i,\theta} = \{f_i(\cdot;\theta)|\theta \in \Theta\}$ for all $i = 1, 2, \ldots$. Thus while the distributions are possibly different, they all share the same parameter θ. We want to estimate θ by minimizing the density power divergence between the data and the model. However, here the model density is different for each Y_i, and hence we need to calculate the divergence between data and model separately for each data point. Considering all the data points it is intuitive to minimize the average divergence between the data points and the models. Therefore if $d_\alpha(\hat{g}_i, f_i(.;\theta))$ denotes the density power divergence between the density estimate corresponding to the i-th data point and the associated model density, we minimize

$$\frac{1}{n}\sum_{I=1}^{n} d_\alpha(\hat{g}_i, f_i(.;\theta))$$

with respect to $\theta \in \Theta$. In the presence of only one data point Y_i from density g_i, the best possible density estimate of g_i is the (degenerate) density which puts the entire mass on Y_i so that we have

$$d_\alpha(\hat{g}_i, f_i(\cdot;\theta)) = \int f_i(y;\theta)^{1+\alpha} dy - \left(1 + \frac{1}{\alpha}\right) f_i(Y_i;\theta)^\alpha + K$$

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

$$H_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left[\int f_i(y;\theta)^{1+\alpha} dy - \left(1 + \frac{1}{\alpha}\right) f_i(Y_i;\theta)^\alpha\right] = \frac{1}{n}\sum_{i=1}^{n} V_i(Y_i;\theta) \tag{2.3}$$

where $V_i(\cdot;\theta)$ is the indicated term within the square brackets in the above equation. Differentiating the above with respect to θ we get the estimating equation of the minimum density power divergence estimator for non-homogeneous observations as

$$\nabla \sum_{i=1}^{n} V_i(Y_i;\theta) = 0$$

which on simplification leads to the equation

$$\sum_{i=1}^{n}\left[f_i(Y_i;\theta)^\alpha u_i(Y_i;\theta) - \int f_i(y;\theta)^{1+\alpha} u_i(y;\theta) dy\right] = 0. \tag{2.4}$$

Here ∇ represents the gradient with respect to θ, and $u_i(y;\theta) = \nabla \ln f_i(y;\theta)$ is the score function for the model of the i-th density. Note that the above estimating equation is unbiased when each data generating density g_i belongs to the corresponding model family $\mathcal{F}_{i,\theta}$. Further note that like the minimum density power divergence estimator for the i.i.d. case, here also we do not require the kernel density estimator and hence we can avoid the problem of bandwidth selection and other associated difficulties.

Note that in the limit where $\alpha \to 0$, the corresponding objective function to be minimized is given by

$$\frac{1}{n}\sum_{i=1}^{n}\left[-\ln(f_i(Y_i;\theta))\right].$$

The minimizer of the above also maximizes $\sum_{i=1}^{n}\ln(f_i(Y_i;\theta))$, and hence $\prod_{i=1}^{n} f_i(Y_i;\theta)$, with respect to θ. Thus the minimum density power estimator with $\alpha = 0$ is nothing but the maximum likelihood estimator for which the estimating equation has the form

$$\sum_{i=1}^{n} u_i(Y_i,\theta) = 0.$$

Therefore the estimating equation in (2.4) is a simple generalization of the maximum likelihood score equation for independently and identically distributed data.

In terms of statistical functionals, the minimum density power divergence functional $T_\alpha(G_1,\ldots,G_n)$ for non-homogeneous observations is given by the relation

$$\frac{1}{n}\sum_{i=1}^{n} d_\alpha(g_i(.), f_i(.;T_\alpha(G_1,\ldots,G_n))) = \min_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n} d_\alpha(g_i(.), f_i(.;\theta)).$$

Since the density power divergence is a genuine divergence in the sense that it is nonnegative and attains its minimum if and only if the two arguments are identical, it follows that the functional $T_\alpha(G_1,\ldots,G_n)$ is Fisher consistent under the assumption of identifiability of the model.

3. Asymptotic properties

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

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

provided such a minimum exists. Let us first present the necessary set up and conditions. Let the parametric model $\mathcal{F}_{i,\theta}$ be as defined above in Section 2. We also assume that there exists a best fitting parameter of θ which is independent

of the index i of the different densities and let us denote it by θ^g. All the results of this section will be derived under these assumptions. The assumption holds if all the true densities g_i belong to the model family so that $g_i = f_i(\cdot;\theta)$ for some common θ, and in that case the best fitting parameter is nothing but the true parameter θ.

Next, recall that the minimum DPD estimator $\hat{\theta}_n$ is obtained as a solution of the estimating equation (2.4). This equation is satisfied by the minimizer of $H_n(\theta)$ in (2.3). Similarly, we also define, for $i = 1, 2, \ldots$,

$$H^{(i)}(\theta) = \int f_i(y;\theta)^{1+\alpha} dy - \left(1 + \frac{1}{\alpha}\right) \int f_i(y;\theta)^{\alpha} g_i(y) dy \tag{3.1}$$

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

$$\nabla H^{(i)}(\theta^g) = 0, \quad i = 1, 2, \ldots.$$

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

$$J^{(i)}_{kl} = \frac{1}{1+\alpha} E_{g_i} \left[\nabla_{kl} V_i(Y_i;\theta)\right], \tag{3.2}$$

where ∇_{kl} represents the partical derivative with respect to the indicated components of θ. We further define the quantities

$$\Psi_n = \frac{1}{n} \sum_{i=1}^{n} J^{(i)}, \tag{3.3}$$

$$\Omega_n = \frac{1}{n} \sum_{i=1}^{n} Var_{g_i} \left[\nabla V_i(Y_i;\theta)\right], \tag{3.4}$$

where Var represents the variance of the random variable. A simple calculation shows that,

$$\begin{aligned} J^{(i)} = & \int u_i(y;\theta^g) u_i^T(y;\theta^g) f_i^{1+\alpha}(y;\theta^g) dy \\ & - \int \{\nabla u_i(y;\theta^g) + \alpha u_i(y;\theta^g) u_i^T(y;\theta^g)\} \{g_i(y) - f_i(y;\theta^g)\} f_i(y;\theta^g)^{\alpha} dy \end{aligned} \tag{3.5}$$

and

$$\Omega_n = \frac{1}{n} \sum_{i=1}^{n} \left[\int u_i(y;\theta^g) u_i^T(y;\theta^g) f_i(y;\theta^g)^{2\alpha} g_i(y) dy - \xi_i \xi_i^T \right], \tag{3.6}$$

where

$$\xi_i = \int u_i(y;\theta^g) f_i(y;\theta^g)^{\alpha} g_i(y) dy. \tag{3.7}$$

We will make the following assumptions to establish the asymptotic properties of the minimum DPD estimators:

(A1) The support $\chi = \{y | f_i(y;\theta) > 0\}$ is independent of i and θ for all i; the true distributions G_i are also supported on χ for all i.

(A2) There is an open subset of ω of the parameter space Θ, containing the best fitting parameter θ^g such that for almost all $y \in \chi$, and all $\theta \in \Theta$, all $i = 1, 2, \ldots$, the density $f_i(y;\theta)$ is thrice differentiable with respect to θ and the third partial derivatives are continuous with respect to θ.

(A3) For $i = 1, 2, \ldots$, the integrals $\int f_i(y;\theta)^{1+\alpha} dy$ and $\int f_i(y;\theta)^{\alpha} g_i(y) dy$ can be differentiated thrice with respect to θ, and the derivatives can be taken under the integral sign.

(A4) For each $i = 1, 2, \ldots$, the matrices $J^{(i)}$ are positive definite and

$$\lambda_0 = \inf_n \quad [\min \text{ eigenvalue of } \Psi_n] > 0$$

(A5) There exists a function $M_{jkl}^{(i)}(y)$ such that

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

where

$$\frac{1}{n} \sum_{i=1}^{n} E_{g_i} \left[M_{jkl}^{(i)}(Y) \right] = O(1) \qquad \forall j, k, l.$$

(A6) For all j, k, we have

$$\lim_{N \to \infty} \sup_{n > 1} \left\{ \frac{1}{n} \sum_{i=1}^{n} E_{g_i} \left[|\nabla_j V_i(Y;\theta)| I(|\nabla_j V_i(Y;\theta)| > N) \right] \right\} = 0 \tag{3.8}$$

$$\lim_{N \to \infty} \sup_{n > 1} \left\{ \frac{1}{n} \sum_{i=1}^{n} E_{g_i} \left[|\nabla_{jk} V_i(Y;\theta) - E_{g_i}(\nabla_{jk} V_i(Y;\theta))| \times I(|\nabla_{jk} V_i(Y;\theta) - E_{g_i}(\nabla_{jk} V_i(Y;\theta))| > N) \right] \right\} = 0 \tag{3.9}$$

where $I(B)$ denotes the indicator variable of the event B.

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

$$\lim_{n \to \infty} \left\{ \frac{1}{n} \sum_{i=1}^{n} E_{g_i} \left[||\Omega_n^{-1/2} \nabla V_i(Y;\theta)||^2 I(||\Omega_n^{-1/2} \nabla V_i(Y;\theta)|| > \epsilon \sqrt{n}) \right] \right\} = 0 \tag{3.10}$$

Theorem 3.1. *Under Assumptions (A1)–(A7), the following results hold:*

(i) There exists a consistent sequence θ_n of roots to the minimum DPD estimating equation (2.4).

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

Remark 3.1. [i.i.d. Case] Note that, setting $f_i = f$ for all i, we get back the corresponding asymptotic properties of the minimum density power divergence estimator for the i.i.d. case as given in Basu et al. (1998) [2]. If $f_i = f$, $i = 1, 2, \ldots$, we get $J^{(i)} = J$, $\xi_i = \xi$ for all i; thus $\Psi_n = J$ and $\Omega_n = K$. Here J, K and ξ are as defined in Basu et al. (1998 [2], Section 3.2). In this case assumptions (A1)–(A5) are exactly the same as the assumptions given in Basu et al. [2], while assumptions (A6) and (A7) are automatically satisfied by the dominated convergence theorem. Thus the Basu et al. (1998) [2] result, which establishes the consistency and asymptotic normality of the minimum density power divergence estimator $\hat{\theta}$ with $n^{1/2}(\hat{\theta}-\theta^g)$ having the asymptotic covariance matrix $\Psi_n^{-1}\Omega_n\Psi_n^{-1} = J^{-1}KJ^{-1}$, emerges as a special case of Theorem 3.1.

Remark 3.2. The assumptions (A1)–(A5) are simple generalizations of the assumptions of Basu et al. (1998) [2] for proving the asymptotic normality of the minimum density power divergence estimator in the i.i.d. case. The assumptions (A6) and (A7) are similar in spirit to the corresponding assumptions required in the case of the maximum likelihood estimators under the similar independent non-homogeneous set-up [Ibragimov and Has'minskii (1981 [10], p. 191)]. These assumptions hold automatically for the minimum density power divergence estimators in the i.i.d. case as observed in Remark 3.1. In subsequent sections we will see that these assumptions hold, for example, for the normal linear regression models under some mild conditions on the independent variables.

4. Influence function analysis

We will now derive the influence function of the minimum density power divergence functional for the non-homogeneous data case. Let G_i denote the true distribution for the datum Y_i, and $T(G_1, \ldots, G_n)$ be the minimum density power divergence functional defined as the minimizer of

$$\sum_{i=1}^{n} H^{(i)}(\theta) = \sum_{i=1}^{n} \left[\int f_i(y;\theta)^{1+\alpha} dy - \frac{1+\alpha}{\alpha}\int f_i(y;\theta)^{\alpha} dG_i(y)\right], \qquad (4.1)$$

or, under appropriate differentiability conditions, as the solution of the estimating equation

$$\textstyle\sum_{i=1}^{n} \nabla H^{(i)}(\theta) = 0, \quad \text{i.e.,}$$

$$(1+\alpha)\textstyle\sum_{i=1}^{n} \left[\int f_i(y;\theta)^{1+\alpha} u_i(y;\theta) dy - \int f_i(y;\theta)^{\alpha} u_i(y;\theta) g_i(y) dy\right] = 0. \qquad (4.2)$$

To derive the Influence function for our special non i.i.d. set-up, we will follow the approach used by Huber (1983) [9] in the context of the influence function for the non i.i.d. fixed-carriers linear models. We consider the contaminate density $g_{i,\epsilon} = (1-\epsilon)g_i + \epsilon\wedge_{t_i}$ where $\wedge_{t_i}$ is the degenerate distribution at the the point of

contamination t_i and G_i denotes corresponding distribution function for all $i = 1, \ldots, n$. Let $\theta = T_\alpha(G_1, \ldots, G_n)$, and let $\theta_\epsilon^{i_0} = T_\alpha(G_1, \ldots, G_{i_0-1}, G_{i_0,\epsilon}, \ldots, G_n)$ be the minimum density power divergence functional with contamination only in the i_0-th direction. Now substitute $\theta_\epsilon^{i_0}$ and $g_{i_0,\epsilon}$ in place of θ and g_{i_0} respectively in the estimating equation (4.2); differentiating with respect to ϵ and evaluating at $\epsilon = 0$, we then get the influence function of the functional which considers contamination only along the i_0-th direction to be

$$IF_{i_0}(t_{i_0}, T_\alpha, G_1, \ldots, G_n) = \Psi_n^{-1} \frac{1}{n} \left[f_{i_0}(t_{i_0}; \theta)^\alpha u_{i_0}(t_{i_0}; \theta) - \xi_{i_0} \right]. \tag{4.3}$$

where $\xi_i = \int u_i(y; \theta) f_i(y; \theta) g_i(y) dy$. Similarly, letting $\theta_\epsilon = T_\alpha(G_{1,\epsilon}, \ldots, G_{n,\epsilon})$ and proceeding similarly, we get the influence function with contamination at all the data-points as

$$IF(t_1, \ldots, t_n, T_\alpha, G_1, \ldots, G_n) = \Psi_n^{-1} \frac{1}{n} \sum_{i=1}^{n} \left[f_i(t_i; \theta)^\alpha u_i(t_i; \theta) - \xi_{i_0} \right]. \tag{4.4}$$

In particular, letting $t_i = t$, $G_i = G$ and $f_i = f$ in above, we get back the influence function of the minimum density power divergence estimator for the i.i.d. case given by

$$IF(t, T_\alpha, G) = J^{-1} \left[f(t; \theta)^\alpha u(t; \theta) - \xi \right] \tag{4.5}$$

where J and ξ are as given in Section 3.2 of Basu et al.(1998) [2].

Hampel (1968 [6], 1974 [7]) defined several summary measure of robustness based on the influence function in case of i.i.d. data; see Hampel et al. (1986) [8] for details. Following these approaches, we will define some influence function based gross summary measures for our non-homogeneous set up. The simplest one is the (unstandardized) *gross-error sensitivity* of the functional T_α at the true distributions $G_1, \ldots, G_n$ considering contamination only in the i_0-th direction, which is defined as

$$\gamma_{i_0}^u(T_\alpha, G_1, \ldots, G_n) = \sup_t \{ || IF_{i_0}(t, T_\alpha, G_1, \ldots, G_n) || \} \tag{4.6}$$

$$= \frac{1}{n} \sup_t \{ \left[f_{i_0}(t; \theta)^\alpha u_{i_0}(t; \theta) - \xi_{i_0} \right]^T \Psi_n^{-2} \left[f_{i_0}(t; \theta)^\alpha u_{i_0}(t; \theta) - \xi_{i_0} \right] \}^{\frac{1}{2}}. \tag{4.7}$$

However it is not invariant to scale transformation of the individual parameter components. Whenever, the asymptotic variance of the corresponding MDPE exists, we can overcome this problem by considering the *Self-Standardized Sensitivity*. For contamination along the i_0-th direction only, this is defined as

$$\gamma_{i_0}^s(T_\alpha, G_1, \ldots, G_n)$$
$$= \sup_t \{ IF_{i_0}(t, T_\alpha, G_1, \ldots, G_n)^T (\Psi_n^{-1} \Omega_n \Psi_n^{-1})^{-1} IF_{i_0}(t, T_\alpha, G_1, \ldots, G_n) \}^{\frac{1}{2}} \tag{4.8}$$

$$= \frac{1}{n} \sup_t \{ \left[f_{i_0}(t; \theta)^\alpha u_{i_0}(t; \theta) - \xi_{i_0} \right]^T \Omega_n^{-1} \left[f_{i_0}(t; \theta)^\alpha u_{i_0}(t; \theta) - \xi_{i_0} \right] \}^{\frac{1}{2}}. \tag{4.9}$$

When we have contamination in all the directions, we can define the (unstandardized) gross-error sensitivity $\gamma^u(T_\alpha, G_1, \ldots, G_n)$ and the self-standardized sensitivity $\gamma^s(T_\alpha, G_1, \ldots, G_n)$ using equation (4.6) and (4.8) respectively with $IF_{i_0}(t, T_\alpha, G_1, \ldots, G_n)$ replaced by $IF(t_1, \ldots, t_n, T_\alpha, G_1, \ldots, G_n)$ and taking supremum over all possible $t_1, \ldots, t_n$.

5. Breakdown point of the location parameter in a location-scale type model

In this section we will derive the breakdown point of the minimum density power divergence estimator in above set-up of non-homogeneous observations for the location parameter in a special class of models. We will consider the above set-up and assume that

$$f_i(y;\theta) \in \mathcal{F}_{i,\theta} = \left\{ \frac{1}{\sigma} f\left(\frac{y - l_i(\mu)}{\sigma} \right) : \theta = (\mu, \sigma) \in \Theta \right\} \tag{5.1}$$

where $l_i(\cdot)$ is some one-to-one function for each $i = 1, \ldots, n$. We consider the breakdown point of the estimator of the location parameter μ at the model and will assume the scale-parameter σ to be fixed (for example, σ can be substituted with any suitable robust scale estimator) and the true data generating densities g_i to belongs to the model family $\mathcal{F}_{i,\theta}$; thus, for each i, $g_i(y) = \frac{1}{\sigma} f(\frac{y - l_i(\mu_g)}{\sigma})$, where μ_g is the true value of the location parameter μ. For given σ, the minimum density power divergence estimator of μ is defined as

$$T^\mu_\alpha(G_1, \ldots, G_n) = \arg\min_\mu \frac{1}{n} \sum_{i=1}^n d_\alpha(g_i(\cdot), f_i(\cdot;\theta)).$$

Assume n to be fixed and consider the contamination models

$$H_{i,\epsilon,m} = (1-\epsilon)G_i + \epsilon K_{i,m},$$

for each i where $\{K_{i,m}\}$ is a sequence of contaminating distributions. Denote the corresponding densities by $h_{i,\epsilon,m}$, g_i and $k_{i,m}$. Following Simpson (1987) [14], we say that there is breakdown in T^μ_α for ϵ level contamination if there exists sequences $K_{i,m}$ such that

$$|T^\mu_\alpha(H_{1,\epsilon,m}, \ldots, H_{n,\epsilon,m}) - T^\mu_\alpha(G_1, \ldots, G_n)| \to \infty \quad \text{as} \quad m \to \infty.$$

Here we will use a generalization of the argument used by Park and Basu (2004) [11] to derive the breakdown of the minimum disparity estimators. Recall that we can also write the density power divergence in equation (2.1) as

$$d_\alpha(g, f) = \int f^{1+\alpha} C_\alpha(g/f) = \int f^{1+\alpha} C_\alpha(\delta + 1)$$

where $\delta = g/f - 1$ and

$$C_\alpha(\delta + 1) = \frac{1}{\alpha} \left[\alpha - (1+\alpha)(\delta+1) + (\delta+1)^{1+\alpha} \right]. \tag{5.2}$$

In the above integral we have suppressed the dummy variable and the differential for simplicity of notation. Note that $C_\alpha(0) = 1$. Define $D_\alpha(g, f) = f^{1+\alpha} C_\alpha(g/f)$. Whenever $\alpha > 0$, we have

$$D_\alpha(g, 0) = \lim_{f \to 0} D_\alpha(g, f) = \lim_{f \to 0} \left[f^{1+\alpha} - \frac{1+\alpha}{\alpha} f^\alpha g + \frac{1}{\alpha} g^{1+\alpha} \right] = \frac{1}{\alpha} g^{1+\alpha}. \quad (5.3)$$

We also utilize useful results based on the special structure of the location-scale type model considered here. For example, note that

$$\int \left\{ \frac{1}{\sigma} f\left(\frac{y - l_i(\mu)}{\sigma} \right) \right\}^{1+\alpha} dy = \frac{1}{\sigma^\alpha} \int \{f(x)\}^{1+\alpha} dx = \frac{1}{\sigma^\alpha} M_f^\alpha, \quad \text{say}$$

which is independent of the location parameter μ and the index i. In addition, we have the crucial lemma given below.

Lemma 5.1. *Assume that $\alpha > 0$ and fix any i. Then for any two densities g_i, h_i in the location-scale model $\mathcal{F}_{i,\theta}$ in equation (5.1) with fixed $\sigma > 0$ and any $\epsilon \in (0, 1)$, the integral $\int D_\alpha(\epsilon g_i, h_i)$ is minimized when $g_i = h_i$.*

We are now in a position to state and prove our main result on breakdown. First we provide the necessary set of assumptions.

(BP1) For each $i = 1, \ldots, n$, $\int \min\{f_i(y; (\mu, \sigma)), k_{i,m}(y)\} \to 0$ as $m \to \infty$ uniformly for $|\mu| \le c$ for any fixed c. That is, the contamination distribution is asymptotically singular to the true distribution and to specified models within the parametric family.

(BP2) For each $i = 1, \ldots, n$, $\int \min\{f_i(y; (\mu_g, \sigma)), f_i(y; (\mu_m, \sigma))\} \to 0$ as $m \to \infty$ if $|\mu_m| \to \infty$ as $m \to \infty$. That is, large values of the parameter μ give distributions which become asymptotically singular to the true distribution.

(BP3) Let $C_\alpha(\cdot)$ be as in equation (5.2). For each $i = 1, \ldots, n$, the contaminating sequence $\{k_{i,m}\}$ is such that

$$d_\alpha(\epsilon k_{i,m}(.), f_i(.; \theta)) \ge d_\alpha(\epsilon f_i(.; \theta), f_i(.; \theta)) = \frac{C_\alpha(\epsilon)}{\sigma^\alpha} M_f^\alpha$$

for any $\theta \in \Theta$ and $0 < \epsilon < 1$ and

$$\limsup_{m \to \infty} \frac{1}{n} \sum_{i=1}^{n} \int k_{i,m}^{1+\alpha} \le \frac{M_f^\alpha}{\sigma^\alpha}.$$

We then have the following theorem.

Theorem 5.2. *Assume that $\alpha > 0$. Then under the assumptions (BP1)–(BP3) above, the asymptotic breakdown point ϵ^* of the minimum DPD functional T_α^μ of the location parameter μ is at least $\frac{1}{2}$ at the location scale set up of (5.1) for fixed scale parameters.*

The above theorem establishes that the minimum DPD procedure generates estimators with high breakdown points for all $\alpha > 0$. For the i.i.d. set up obtained by letting $f_i(y; \theta) = f_\theta(y)$, Theorem 5.2 directly yields the following Corollary.

Corollary 5.3. *Suppose independent and identically distribution data are obtained from the location scale model* $\mathcal{F}_\theta = \{\frac{1}{\sigma} f(\frac{y-\mu}{\sigma}) : \theta = (\mu, \sigma) \in \Theta\}$ *and assume that the scale parameter* σ *is fixed. Then under assumptions (BP1)–(BP3) and for all* $\alpha > 0$*, the minimum density power divergence estimator of the location parameter* μ *has asymptotic breakdown point of at least* $\frac{1}{2}$ *at the model.*

The above result may be contrasted with the Basu et al. (1998 [2], Section 4.3) result which gives the simultaneous location and scale breakdown point of the minimum DPD estimator to be $\alpha/(1+\alpha)^{3/2}$. The remarks give us some justification for assumptions (BP1)–(BP3).

Remark 5.1. Suppose that the contaminating densities $\{k_{i,m}\}$ belongs to the model presented in equation (5.1), and satisfies the set up of this section for all i and all m. Then the following results are seen to be true.

1 Assumption (BP3) holds. The second part of the assumption holds trivially and the first part of the assumption holds by Lemma 5.1.
2 Let $k_{i,m} = f_i(y; (\mu_m, \sigma))$ and suppose that $|\mu_m| \to \infty$ as $m \to \infty$. Then assumption (BP2) implies assumption (BP1).
3 If we assume that $f(\cdot) = \phi(\cdot)$ in the model represented by equation (5.1) where $\phi(\cdot)$ is the univariate normal density, assumption (BP2) also holds trivially.

We expect that it will be possible to prove the breakdown result in Theorem 5.2 under conditions where a weaker version of (BP3) will suffice but we do not have a proof at this point.

6. Application: Normal linear regression

A natural situation where the theory proposed above would be immediately applicable is the case of linear regression. In particular all the machinery will immediately fall into place for the case of linear regression set up with normal errors where the conditional approach to inference given fixed values of the explanatory variable is adopted. In the rest of the paper we will provide applications of the proposed method in case of linear regression. Other application domains of the proposed theory will be considered in a sequel paper.

A qualification of the linear regression set up employed in the following is necessary in this context. To describe this, consider the simple linear regression set up with a single independent variable. If one considered the joint distribution of the entire data considering the explanatory variable X to be random (together with the response variable Y), one could write down a single "multivariate" DPD measure between the data and an appropriate multivariate model such as the bivariate normal. In such a case one could obtain the estimates of the model parameters by using the i.i.d. formulation of the DPD based on the bivariate data vectors (X, Y), and the estimators of the bivariate model may then be presumably used to determine the estimates of the regression parameters of interest. However the elegance of the linear regression model stems from

the fact that one does not need to model the explanatory variable and can concentrate entirely on the parameters of the conditional distribution alone. Indeed this is how linear regression is performed in most applications and this is the approach we follow. Unlike the multivariate DPD formulation, this requires the treatment of nonhomogeneous observations where the theory developed in this paper becomes immediately applicable. Apart from the fact that this avoids the modelling of unnecessary parameters and keeps the method simple in implementation, this has a theoretical advantage over the multivariate DPD as well. Basu et al. (1998) [2] pointed out the DPD method typically loses efficiency for fixed α as the dimension of the data vector increases. Under the present theory, however, it follows from Theorem 6.2 that the asymptotic relative efficiency of the minimum DPD estimators of the regression parameters (relative to the least squares estimators under the normal model) is exactly the same as the aysmptotic relative efficiency of the minimum DPD estimator of the normal mean in the univariate normal model (relative to maximum likelihood). The efficiency gain for the non-homogeneous approach over the multivariate DPD approach will be further pronounced in case of multiple linear regression.

Basu et al. (1998 [2], Section 3.5) briefly suggested linear regression as being among the likely scenarios where the methods based on density power divergence could be extended. A technical report by the same authors, with the same title, gives some more details of the possibilities of this method for the regression case (Statistical Report Number 7, Department of Mathematics, University of Oslo, 1998). Durio and Isaia (2011) [4] followed up on this method and provided some simulation results to indicate the superior robust behavior of the minimum DPD estimators of the regression parameters. Scott (2001) considered the special case of this for the minimum L_2 distance estimator ($\alpha = 1$). However the asymptotic properties of the minimum DPD estimators in this context have not yet been rigorously studied in the literature. As the observations are no longer identically distributed, the theory needs to be suitably extended, both for the efficiency and the robustness results, without which the results remain ad-hoc. Here we will provide the theoretical background and fill in this gap in the literature using the set up of independent but non-homogeneous observations, and establish the asymptotic properties and robustness credentials of the minimum DPD estimators.

Consider the linear regression model:

$$y_i = x_i^T \beta + \epsilon_i, \quad i = 1, \ldots, n, \tag{6.1}$$

where the error ϵ_i's are i.i.d. normal variables with mean zero and variance σ^2, $x_i^T = (x_{i1}, \ldots, x_{i,p})$ is the vector of the independent variables corresponding to the i-th observation and $\beta = (\beta_1, \ldots, \beta_p)^T$ represents the regression coefficients. We will assume that x_i's are fixed. Then $y_i \sim N(x_i^T \beta, \sigma^2)$, and hence the y_i's are independent but not identically distributed. Thus y_i's satisfy our above set-up and hence the minimum density power divergence of the parameter $\theta = (\beta^T, \sigma^2)^T$ can be obtained by minimizing the expression in equation (2.3) with

$f_i \equiv N(x_i^T\beta, \sigma^2)$. Under the notation of equation (2.3), we then have

$$V_i(y_i;\theta,x_i) = \frac{1}{(2\pi)^{\alpha/2}\sigma^\alpha\sqrt{1+\alpha}} - \frac{1+\alpha}{\alpha}\frac{1}{(2\pi)^{\alpha/2}\sigma^\alpha}e^{-\alpha(y_i-x_i^T\beta)^2/(2\sigma^2)}$$

Thus, our objective function to be minimized becomes

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n V_i(y_i;\theta,x_i) &= \frac{1}{(2\pi)^{\alpha/2}\sigma^\alpha\sqrt{1+\alpha}} \\ &\quad - \frac{1+\alpha}{\alpha}\frac{1}{(2\pi)^{\alpha/2}\sigma^\alpha}\frac{1}{n}\sum_{i=1}^n e^{-\alpha(y_i-x_i^T\beta)^2/(2\sigma^2)}, \quad (6.2)\end{aligned}$$

which is exactly the same as the one suggested by Basu et al. (1998) [2] for linear regression and the equation considered by Durio and Isaia (2011 [4], equation (3)). Letting $\nabla_j, j = 1,\ldots,p$ represent the partial derivative with respect to β_j we get

$$\nabla_j V_i(y_i;\theta,x_i) = -\frac{1+\alpha}{(2\pi)^{\alpha/2}\sigma^{\alpha+2}}e^{-\frac{\alpha(y_i-x_i^T\beta)^2}{2\sigma^2}}(y_i - x_i^T\beta)x_{ij} \quad \forall j = 1,\ldots,p$$

and the partial derivative with respect to σ^2 is then

$$\begin{aligned}&\nabla_{p+1}V_i(y_i;\theta,x_i) \\ = \quad &-\frac{1}{(2\pi)^{\alpha/2}}\left[\frac{\alpha}{2\sigma^{\alpha+2}\sqrt{1+\alpha}} - \frac{(1+\alpha)}{2\sigma^{\alpha+2}}e^{-\frac{\alpha(y_i-x_i^T\beta)^2}{2\sigma^2}}\left\{1 - \frac{(y_i-x_i^T\beta)^2}{\sigma^2}\right\}\right].\end{aligned}$$

Thus we get the estimating equation to be

$$\sum_{i=1}^n x_{ij}(y_i - x_i^T\beta)e^{-\frac{\alpha(y_i-x_i^T\beta)^2}{2\sigma^2}} = 0 \quad \forall j = 1,\ldots,p \qquad (6.3)$$

$$\sum_{i=1}^n \left[1 - \frac{(y_i-x_i^T\beta)^2}{\sigma^2}\right]e^{-\frac{\alpha(y_i-x_i^T\beta)^2}{2\sigma^2}} = \frac{\alpha}{(1+\alpha)^{\frac{3}{2}}}. \qquad (6.4)$$

We can then solve these estimating equations numerically to obtain the estimates of θ. Let us denote these estimators by $\hat\theta^T = (\hat\beta^T, \hat\sigma^2)$.

6.1. Asymptotic efficiency

Following Theorem 3.1, we can now obtain the asymptotic distribution of the estimator $\hat\beta$ and $\hat\sigma^2$. For simplicity, we will assume that the true data generating density g_i also belongs to the model family of distributions, i.e., $g_i(y) = f_i(y;\theta)$. Then we can derive the simplified form of the matrices Ψ_n and Ω_n. Note that for this regression model, we have

$$u_i(y_i;\theta) = \begin{pmatrix} \frac{(y_i-x_i^T\beta)}{\sigma^2}x_i \\ \\ \frac{(y_i-x_i^T\beta)^2}{2\sigma^4} - \frac{1}{2\sigma^2} \end{pmatrix}. \qquad (6.5)$$

Thus a routine calculation shows that the matrix $J^{(i)}$ is given by

$$J^{(i)} = \int u_i(y;\theta)u_i(y;\theta)^T f_i(y;\theta)^{1+\alpha}dy = \begin{pmatrix} \zeta_\alpha x_i x_i^T & 0 \\ 0 & \varsigma_\alpha \end{pmatrix} \tag{6.6}$$

where

$$\begin{aligned} \zeta_\alpha &= (2\pi)^{-\frac{\alpha}{2}}\sigma^{-(\alpha+2)}(1+\alpha)^{-\frac{3}{2}} \\ \varsigma_\alpha &= (2\pi)^{-\frac{\alpha}{2}}\sigma^{-(\alpha+4)}\frac{1}{4}\left(\frac{2+\alpha^2}{(1+\alpha)^{\frac{5}{2}}}\right). \end{aligned}$$

Therefore, we have a simplified form for the matrix Ψ_n as

$$\Psi_n = \begin{pmatrix} \frac{\zeta_\alpha}{n}(X^TX) & 0 \\ 0 & \varsigma_\alpha \end{pmatrix} \tag{6.7}$$

where $X^T = (x_1, \ldots, x_n)_{p\times n}$. Similarly, we get

$$\xi_i = \int u_i(y;\theta)f_i(y;\theta)^{1+\alpha}dy = \begin{pmatrix} 0 \\ -\frac{\alpha}{2}\zeta_\alpha \end{pmatrix} \tag{6.8}$$

and hence

$$\Omega_n = \begin{pmatrix} \frac{\zeta_{2\alpha}}{n}(X^TX) & 0 \\ 0 & \varsigma_{2\alpha} - \frac{\alpha^2}{4}\zeta_\alpha^2 \end{pmatrix}. \tag{6.9}$$

Using the above, we are now in a position to derive the asymptotic distributions of the minimum DPD estimator of the regression coefficients and error variances under the assumptions (A1)–(A7). We first present some mild conditions on the given values of the independent variables, under which these assumptions may be shown to hold.

(R1) The values of x_i's are such that for all j, k, and l

$$\sup_{n>1}\max_{1\le i\le n}|x_{ij}| = O(1), \qquad \sup_{n>1}\max_{1\le i\le n}|x_{ij}x_{ik}| = O(1), \tag{6.10}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}|x_{ij}x_{ik}x_{il}| = O(1). \tag{6.11}$$

(R2) The matrix $X^T = (x_1, \ldots, x_n)_{p\times n}$ satisfies

$$\inf_n \left[\text{min eigenvalue of } \frac{(X^TX)}{n}\right] > 0, \tag{6.12}$$

which also implies that the matrix X has full column rank, and

$$n\max_{1\le i\le n}[x_i^T(X^TX)^{-1}x_i] = O(1). \tag{6.13}$$

Then the following lemma is easily seen to be true.

Lemma 6.1. *Consider the set-up of the normal linear regression model and assume that the true data generating density belongs to the model family. Then the conditions (R1) and (R2) imply assumptions (A1)–(A7).*

Note that the conditions (R1) and (R2) on the x_i's mainly says that their values remain bounded in large samples and the spectrum of the corresponding sum-product matrix (X^TX) remains bounded away from zero. With these conditions, the asymptotic distribution of the minimum density power divergence estimators of the parameters of the linear regression model are derived in the following theorem:

Theorem 6.2. *Under the set-up of the normal linear regression model considered here, assume that the true data generating density belongs to the model family and the given values of the independent variables satisfies assumptions (R1) and (R2). Then,*

(i) There exists a consistent sequence as $\hat{\theta}^T = (\hat{\beta}^T, \hat{\sigma}^2)$ of roots to the minimum DPD estimating equations (6.3) and (6.4).

(ii) The asymptotic distributions of $\hat{\beta}$ and $\hat{\sigma}^2$ are independent.

(iii) The asymptotic distribution of $(X^TX)^{\frac{1}{2}}(\hat{\beta}-\beta)$ is a p-dimensional normal with mean (vector) and covariance matrix $v_\alpha^\beta I_p$ and $\sqrt{n}(\hat{\sigma}^2-\sigma^2)$ follows a normal distribution with mean 0 and variance v_α^e, where

$$
\begin{aligned}
v_\alpha^\beta &= \frac{\zeta_{2\alpha}}{\zeta_\alpha^2} = \sigma^2\left(1+\frac{\alpha^2}{1+2\alpha}\right)^{\frac{3}{2}} \\
v_\alpha^e &= \frac{\varsigma_{2\alpha}-\frac{\alpha^2}{4}\zeta_\alpha^2}{\varsigma_\alpha^2} = \frac{4\sigma^4}{(2+\alpha^2)^2}\left[2(1+2\alpha^2)\left(1+\frac{\alpha^2}{1+2\alpha}\right)^{\frac{5}{2}}-\alpha^2(1+\alpha)^2\right].
\end{aligned}
$$

Note that substituting $\alpha = 0$ in the expression of the asymptotic variances of the estimators $\hat{\beta}$ and $\hat{\sigma}^2$ in the above theorem, we will get the exactly the same results as obtained for their maximum likelihood estimates.

We will now look at the *Asymptotic Relative Efficiency* (ARE) of the minimum density power divergence estimators with respect to the (fully efficient) maximum likelihood estimator. The ARE of the estimator $\hat{\beta}$ of the regression coefficient $\beta = (\beta_1, \ldots, \beta_p)$ is the same for all the β_i's and is given by

$$
\frac{v_0^\beta}{v_\alpha^\beta} \times 100 = \left(1+\frac{\alpha^2}{1+2\alpha}\right)^{-\frac{3}{2}} \times 100
$$

Similarly the asymptotic relative efficiency of the estimator $\hat{\sigma}^2$ of the error variance is given by

$$
\frac{v_0^e}{v_\alpha^e} \times 100 = \frac{(2+\alpha^2)^2}{2}\left[2(1+2\alpha^2)\left(1+\frac{\alpha^2}{1+2\alpha}\right)^{\frac{5}{2}}-\alpha^2(1+\alpha)^2\right]^{-1} \times 100.
$$

Table 1 presents the asymptotic relative efficiencies of these estimators for various values of α. From the table it is easy to see that the loss of efficiency is quite small for small values of α. It is also interesting to note that the ARE of

TABLE 1
The Asymptotic Relative Efficiencies of the estimators $\hat{\beta}$ and $\hat{\sigma}^2$ for various values of the tuning parameter α

α	0	0.01	0.02	0.05	0.10	0.15	0.25	0.50	0.75	1.00
$ARE(\hat{\beta})$	100	99.99	99.94	99.66	98.76	97.46	94.06	83.81	73.76	64.95
$ARE(\hat{\sigma}^2)$	100	99.97	99.88	99.32	97.56	95.05	88.84	73.06	61.53	54.11

the minimum DPD estimator of the regression coefficient β is the same as the ARE of the minimum DPD estimator of the normal mean parameter and ARE of the minimum DPD estimator of error variance σ^2 is the same as the ARE of the normal variance are reported in Basu et al. (1998) [2].

6.2. Equivariance of the regression coefficient estimators

Standard regression literature considers three types of equivariance of the estimators of regression coefficients – regression, scale and affine equivariance. It is known that the maximum likelihood estimator satisfies all the three properties. We will now show that our minimum density power divergence estimator of the regression coefficient $\hat{\beta}$ also satisfies all the three equivariance properties also for all $\alpha > 0$.

Theorem 6.3. *The minimum density power divergence estimator $\hat{\beta}$ of the regression coefficient β is regression equivariant, scale equivariant and affine equivariant.*

The regression equivariance of the estimator allows us to assume, without loss of generality, any suitable value for the parameter β while proving any asymptotic properties or describing the Monte Carlo studies. It also implies that no linear structure should remain while regressing the residuals on the explanatory variable x. The scale and affine equivariance of the estimator $\hat{\beta}$ ensures that it does not depend on the choice of measurement unit for the response variable y and on the choice of coordinate system for the explanatory variables x.

Further note that the objective function here depends on the y_i and x_i's through only the summation

$$\sum_{i=1}^{n} e^{-\frac{\alpha(y_i - x_i^T(A\beta))^2}{2\sigma^2}}$$

which is permutation invariant. Thus the corresponding estimators of the regression coefficient β and the error variance σ are both *Permutation Equivariant* so that the ordering of the data does not affect the estimators.

6.3. Influence function and sensitivities

To see the robustness properties of the estimators of the regression coefficients and the error variance, we will now derive the influence function of these estimators following the notations of Section 4. Let us denote the minimum density

power divergence functional of β and σ^2 by T_α^β and T_α^σ respectively so that the corresponding functional for $\theta^T = (\beta^T, \hat{\sigma}^2)$ is given by $T_\alpha = (T_\alpha^\beta, T_\alpha^\sigma)$. Now using the formula derived in Section 4, we can compute the influence function of the functional T_α.

Note that from the expression of Ψ_n given in equation (6.7), we get

$$\Psi_n^{-1} = \begin{pmatrix} \frac{n}{\zeta_\alpha}(X^TX)^{-1} & 0 \\ 0 & \frac{1}{\zeta_\alpha} \end{pmatrix}$$

Then using the expression of u_i and ξ_i from equation (6.5) and (6.8) respectively, we get the influence function of the estimator T_α with contamination at the direction i_0 which is given by

$$IF_{i_0}(t_{i_0}, T_\alpha, G_1, \ldots, G_n) = \begin{pmatrix} \frac{1}{\zeta_\alpha}(X^TX)^{-1}\frac{(t_{i_0}-x_{i_0}^T\beta)}{\sigma^2}x_{i_0}f_{i_0}(t_{i_0};\theta)^\alpha \\ \frac{1}{n\varsigma_\alpha}\left[\left\{\frac{(t_{i_0}-x_{i_0}^T\beta)^2}{2\sigma^4} - \frac{1}{2\sigma^2}\right\}f_{i_0}(t_{i_0};\theta)^\alpha + \frac{\alpha}{2}\zeta_\alpha\right] \end{pmatrix}. \tag{6.14}$$

Simplifying, the influence function for the estimator T_α^β of the regression coefficients with contamination only in i_0-th data-point only becomes

$$IF_{i_0}(t_{i_0}, T_\alpha^\beta, G_1, \ldots, G_n) = (1+\alpha)^{\frac{3}{2}}(X^TX)^{-1}x_{i_0}(t_{i_0} - x_{i_0}^T\beta)e^{-\frac{\alpha(t_{i_0}-x_{i_0}^T\beta)^2}{2\sigma^2}} \tag{6.15}$$

and the influence function for the estimator T_α^σ of the error variance with contamination in the i_0-th data-point only becomes

$$\begin{aligned} IF_{i_0}(t_{i_0}, T_\alpha^\sigma, G_1, \ldots, G_n) &= \frac{2(1+\alpha)^{\frac{5}{2}}}{n(2+\alpha^2)}\left\{(t_{i_0} - x_{i_0}^T\beta)^2 - \sigma^2\right\}e^{-\frac{\alpha(t_{i_0}-x_{i_0}^T\beta)^2}{2\sigma^2}} \\ &\quad + \frac{2\alpha(1+\alpha)^2}{n(2+\alpha^2)}. \end{aligned} \tag{6.16}$$

Since the functions se^{-s^2} and $s^2e^{-s^2}$ are both bounded in $s \in \mathbb{R}$, both the influence functions in (6.15) and (6.16) are bounded in t_{i_0} for all $\alpha > 0$ and for any i_0. This implies that the minimum density power divergence estimators with $\alpha > 0$ will be robust with respect to the outliers in any data-point. However the influence functions are clearly unbounded for $\alpha = 0$ which corresponds to the non-robust maximum likelihood estimators.

Similarly the influence function for the estimator T_α^β of the regression coefficients with contamination in all data-point can be shown to be

$$IF(t_1, \ldots, t_n, T_\alpha^\beta, G_1, \ldots, G_n) = (1+\alpha)^{\frac{3}{2}}(X^TX)^{-1}\sum_{i=1}^n x_i(t_i - x_i^T\beta)e^{-\frac{\alpha(t_i-x_i^T\beta)^2}{2\sigma^2}} \tag{6.17}$$

and the influence function for the estimator T_α^σ of the error variance with contamination in all data-point will be

$$IF(t_1,\cdot,t_n,T_\alpha^\sigma,G_1,\cdot,G_n) = \frac{2(1+\alpha)^{\frac{5}{2}}}{n(2+\alpha^2)}\sum_{i=1}^{n}\left\{(t_i-x_i^T\beta)^2-\sigma^2\right\}e^{-\frac{\alpha(t_i-x_i^T\beta)^2}{2\sigma^2}} + \frac{2\alpha(1+\alpha)^2}{(2+\alpha^2)}. \tag{6.18}$$

Here also the influence functions (6.17) and (6.18) are bounded in t_i's for all $\alpha>0$ and unbounded for $\alpha=0$.

Now let us derive the sensitivities of the estimator of the regression coefficient β to explore the extent of robustness of the estimator with respect to the value of α. Using the form given in Section 4, the gross-error sensitivity of the estimator T_α^β of β in the case of contamination only in i_0^{th} direction can be found to be

$$\begin{aligned}\gamma_{i_0}^u(T_\alpha^\beta,G_1,\ldots,G_n) &= \frac{(1+\alpha)^{\frac{3}{2}}}{\sqrt{\alpha}}\sigma e^{-\frac{1}{2}}||(X^TX)^{-1}x_{i_0}|| \quad \text{if } \alpha>0 \\ &= \infty \qquad\qquad\qquad\qquad\qquad\qquad \text{if } \alpha=0\end{aligned} \tag{6.19}$$

And the self-standardized sensitivity of the estimator T_α^β in the case of contamination only in i_0^{th} direction is given by

$$\begin{aligned}\gamma_{i_0}^s(T_\alpha^\beta,G_1,\ldots,G_n) &= \frac{(1+\alpha)^{\frac{3}{2}}}{\sqrt{\alpha v_\alpha^\beta}}\sigma e^{-\frac{1}{2}}\sqrt{x_{i_0}^T(X^TX)^{-1}x_{i_0}} \quad \text{if } \alpha>0 \\ &= \infty \qquad\qquad\qquad\qquad\qquad\qquad \text{if } \alpha=0\end{aligned} \tag{6.20}$$

or,

$$\begin{aligned}\gamma_{i_0}^s(T_\alpha^\beta,G_1,\ldots,G_n) &= \frac{(1+2\alpha)^{\frac{3}{4}}}{\sqrt{\alpha}}e^{-\frac{1}{2}}\sqrt{x_{i_0}^T(X^TX)^{-1}x_{i_0}} \quad \text{if } \alpha>0 \\ &= \infty \qquad\qquad\qquad\qquad\qquad\qquad \text{if } \alpha=0\end{aligned} \tag{6.21}$$

It is easy to see that both the sensitivities $\gamma_{i_0}^u$ and $\gamma_{i_0}^s$ are decreasing function of $\alpha>0$ for any given x_i's. This implies that in the presence of the outliers in only one direction the robustness of the estimator T_α^β increases as α increases.

However, we have seen in Section 6.1 that the asymptotic relative efficiency of the estimator T_α^β of the regression coefficient β decreases as the tuning parameter α increases. Thus the parameter α gives a trade-off between the efficiency and the robustness of the estimator of regression coefficients.

It is interesting to note that besides the tuning parameter α, the sensitivities also depend on the values of the explanatory variable x_is. Thus the robustness of the estimator T_α^β also depends on the values of x_i's. Moreover, from the expres-

sion of sensitivities, whenever the value of a x_{i_0} becomes far from the center of the data-cloud, the value of both the gross-error sensitivity and self-standardized sensitivity increases implying that the robustness decreases. This fact is quite intuitive from the basic concept of outliers in the explanatory variable.

6.4. Breakdown point of the estimator of regression coefficient

Now we consider the breakdown point of the estimator of the regression coefficients β using the theory developed in Section 5. Note that the regression set-up exactly matches with the set-up considered in Section 5 with $f(\cdot) = \phi(\cdot)$, the standard normal density, $\mu = \beta$, and $l_i(\mu) = l_i(\beta) = x_i^T\beta$ for all i. Since here we are mainly interested about the breakdown of the estimator of β, as in Section 5, we will assume that the error variance σ^2 to be fixed. In practice, we may replace it by any robust estimator that may be assumed to be the same as the true value of σ^2 asymptotically.

Note that by Remark 5.1.3, the assumption (BP2) required for the breakdown result in Theorem 5.2 is satisfied trivially. Thus from Theorem 5.2 it follows that under assumption (BP1) and (BP3) about the contaminating densities, the asymptotic breakdown point of the minimum density power divergence estimator of β is at least $\frac{1}{2}$ at the model for all $\alpha > 0$.

Further it was proved in Theorem 6.2 that the minimum DPD estimator of β is regression equivariant. And it follows from Rousseeuw and Leroy (1987 [12], Theorem 4, page 125) that the finite sample breakdown point of any regression equivariant estimator of β is at most

$$\frac{[(n-p)/2]+1}{n}$$

at all sample Z of size n. Hence the asymptotic breakdown point of β can be at most $\frac{1}{2}$. Thus we get the following theorem giving the maximum asymptotic breakdown of the minimum density power divergence estimator of the regression coefficient β at the model.

Theorem 6.4. *Assume the contaminating densities are such that (BP1) and (BP3) hold. Then for any $\alpha > 0$, the asymptotic breakdown point of the minimum density power divergence estimator of the regression coefficient β is exactly $\frac{1}{2}$ at the model.*

Also if we assume that the contamination densities also belongs to the model family, i.e., $k_{i,m}$ is the $N(x_i^T\beta_m, \sigma^2)$ density with $|\beta_m| \to \infty$, then by Remark 5.1 the assumptions (BP1) and (BP3) again hold true and the above breakdown result follows.

6.5. Comparison with other methods

We believe that in many ways the method described in this paper represents a significant addition to the literature of density-based minimum distance inference. Minimum distance type methods often have a natural robustness property

which we have tried to exploit in this paper. In particular, the property that the implementation of this technique requires no nonparametric smoothing and numerical integration over the range of the variable makes this method particularly appealing. We are not aware of another density-based minimum distance technique which, on the whole, combines the property of robustness with ease of implementation in such a natural way as the present technique.

While the effort we have made has been for developing a simple method useful for non-homogeneous data, our primary illustration in this paper has been in the area of robust regression. We briefly provide a comparison of our method with the leading robust regression techniques available at present. Andersen (2008) [1] provides a useful discussion of the existing robust linear regression methods including Least Absolute Deviation (LAD), Least Median of Squares (LMS), Least Trimmed Squares (LTS), Bounded influence R-estimates, M-estimates, S-estimates, Generalized M-estimates, generalized S-estimates, MM-estimates etc. Except the M-estimates all other mentioned robust regression estimates have bounded influence functions like the minimum DPD estimates with $\alpha > 0$. However, among all the above estimators LAD and M-estimators have breakdown point 0 and the bounded influence R-estimates has a breakdown point less than 0.2 and hence their robustness performances are not alway satisfactory. Some of the high breakdown regression estimates like LMS, LTS and S-estimates (each having an asymptotic breakdown point of $\frac{1}{2}$) have very low asymptotic efficiencies compared to the OLS (ordinary least squares) method. Indeed, LMS and LTS are not $\sqrt{n}$-consistent. The Generalized M-estimates, generalized S-estimates and MM-estimates have high asymptotic breakdown point of $\frac{1}{2}$ and also relatively higher efficiency of 95%, 67% and 95%, although their desirable properties are partially tempered by their complicated computational structure. Considering all these factors we believe, on the basis of the properties of the minimum DPD estimators developed in this paper, that this technique is competitive with most of the existing techniques in the robust regression literature.

7. Real data examples

Durio and Isaia (2011) [4] illustrated the robustness performance of the minimum density power divergence estimators of the regression coefficients through simulation studies. However, they did not compare it with any of the other robust regression methods. Besides the theoretical comparison presented earlier, in this section we will present some real data examples of linear regression and compare the robustness performance of the minimum DPD estimators with the popular LMS estimators. All the real data sets examined below are taken from Rousseeuw and Leroy (1987) [12].

In a real situation it is important to have a technique which leads to the selection of an "optimum" value of α that applies to the given data set. Durio and Isaia (2011) [4] has provided an algorithm to find such an optimum value of α. However, their criterion is complicated, and involves developing a test of hypothesis about the regression parameters. As we restrict ourselves to the

estimation problem in this case, we consider an adaptation of the Warwick and Jones (2005) [15] algorithm for the selection of the DPD tuning parameter for the independent and identically distributed data. The details of this method are given in Ghosh and Basu (2013) [5]. The selection depends on the choice of an initial pilot estimator; here we consider the minimum density power divergence estimators corresponding to $\alpha = 0.3$ and $\alpha = 0.5$ as our pilot estimators.

7.1. Hertzsprung-Russell data of the star cluster

As our first example, we consider the data for the Hertzsprung-Russell diagram of the star cluster CYG OB1 containing 47 stars in the direction of Cygnus (Table 3, Chapter 2, Rousseeuw and Leroy, 1987 [12]). For this data the independent variable x is the logarithm of the effective temperature at the surface of the star (T_e), and the dependent variable y is the logarithm of its light intensity (L/L_0). The data were thoroughly studied by Rousseeuw and Leroy (1987) [12] who inferred that there are two groups of data-points — the four stars in the upper right corner of the scatter plot (Figure 1) clearly form a separate group from the rest of the data-points. In fact these four stars (with indices 11, 20, 30 and 34) are known as giants in astronomy. So, these outliers are actually not

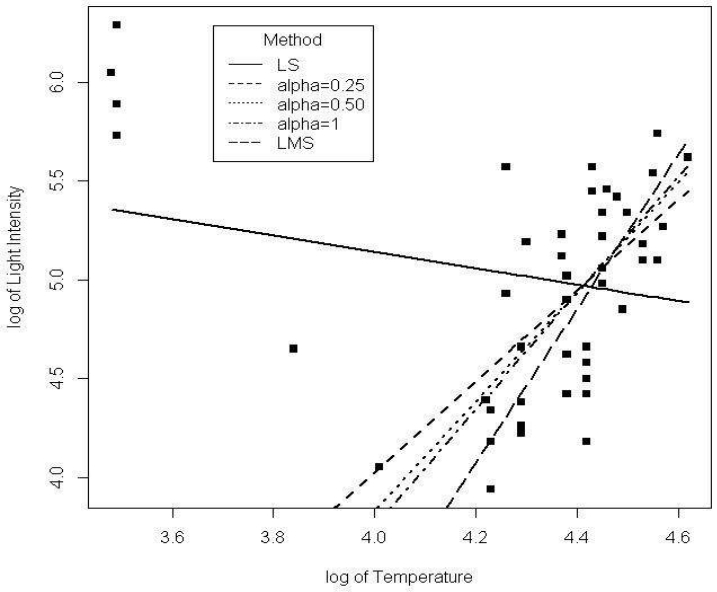

FIG 1. *Plots of the data-points and fitted regression lines for Hertzsprung-Russell Data of the Star Cluster using Least Square (LS), Least Median Square (LMS) and minimum DPD estimators for several* α.

TABLE 2
The parameter estimates of the linear regression model for the Hertzsprung-Russell data using several minimum density power divergence methods and the LMS method

	α									
	0	0.05	0.1	0.25	0.4	0.5	0.6	0.8	1	LMS
β_1	6.69	6.74	6.78	−5.16	−6.57	−7.22	−7.57	−7.89	−8.03	−12.30
β_2	−0.39	−0.40	−0.41	2.30	2.62	2.76	2.84	2.91	2.95	3.90
σ	0.61	0.60	0.60	0.42	0.40	0.40	0.40	0.40	0.41	—

recording errors and can not be discarded. They are indeed leverage points with the interpretation that the data are coming from two different groups.

The estimates of the regression coefficients for the regression of y on x and the error variance obtained the minimum density power divergence estimation for various α are presented in Table 2 and some of the fitted models are plotted in Figure 1. It is clear that the estimators corresponding to $\alpha = 0$ (which are the ordinary least square estimators also) are pulled away significantly by the four leverage points and hence it is not possible to separate out the two group of data by looking at the corresponding residuals. However, like the most robust (but inefficient) LMS estimators, the minimum DPD estimators with $\alpha \geq 0.25$ can successfully ignore the outliers to give excellent robust fits and are much closer to the fit generated by the LMS estimates. Based on the residuals of these minimum DPD estimators, we can also separate out the two group of observations – four large residuals correspond to the four giant stars. The selection of the "optimum" tuning parameters following Warwick and Jones (2005) [15] lead to $\alpha = 0.786$ and 0.932 respectively corresponding to the pilot estimators at $\alpha = 0.3$ and 0.5. In either case the respective line will lie in the extremely narrow region between the minimum DPD lines at $\alpha = 0.5$ and $\alpha = 1$ presented in Figure 1.

If we delete the four large leverage points, the OLS estimators of the intercept and slope parameters are −4.770 and 2.204 respectively; this slope parameter and the resulting fit is quite close to the robust minimum DPD fit for the full data at $\alpha = 0.25$. If we further delete the outlying point $(3.85, 4.6)$ along with the four large outliers, the OLS estimators of the intercept and slope parameters become −8.544 and 3.057, close to the MDPDE for the full data at $\alpha = 1$.

7.2. Belgium telephone call data

The real data set (Table 2, Chapter 2, Rousseeuw and Leroy, 1987 [12]) for our second example is from the Belgian Statistical Survey by the Ministry of Economy and contains the total number (in tens of millions) of international phone calls made in a year from 1950 to 1973. However, due to the use of another recording system (giving the total number of minutes of these calls) from the year 1964 to 1969, the data contains heavy contamination in the y-direction in that range. The years 1963 and 1970 are also partially affected for the same reason.

TABLE 3
The parameter estimates of the linear regression model for the Belgium Telephone Call data using several minimum density power divergence methods and the LMS method

	α									
	0	0.05	0.1	0.25	0.4	0.5	0.6	0.8	1	LMS
β_1	−26.01	−25.53	−24.94	−21.97	−5.24	−5.26	−5.28	−5.31	−5.36	−5.61
β_2	0.50	0.50	0.48	0.43	0.11	0.11	0.11	0.11	0.11	0.12
σ	5.38	5.40	5.41	5.29	0.11	0.11	0.12	0.12	0.12	—

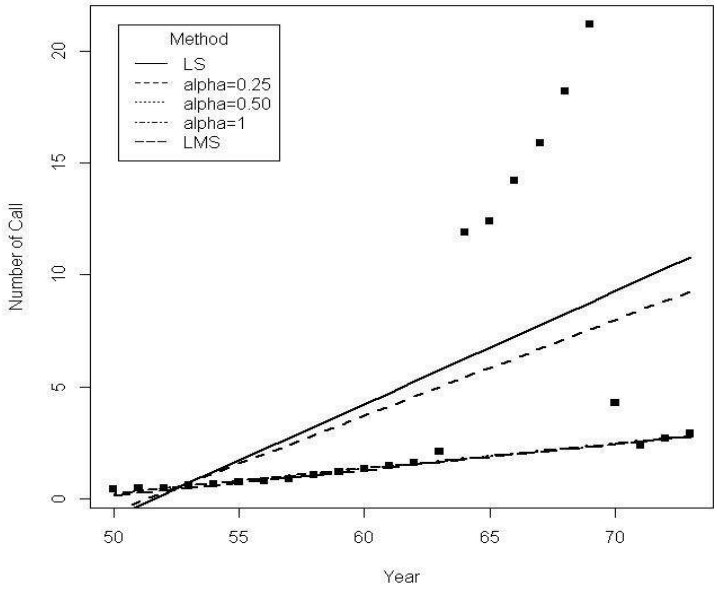

FIG 2. *Plots of the data-points and fitted regression lines for Belgium Telephone Call data using Least Square (LS), Least Median Square (LMS) and minimum DPD estimation for several* α.

The estimators of the regression coefficients for the different methods are presented in Table 3. Some of the fitted lines are also shown in Figure 2 along with the data points. It is clear that the estimators corresponding to $\alpha = 0$ (which are the ordinary least square estimators) are heavily affected by the outliers; however, like the LMS estimator, the minimum DPD estimators with $\alpha \geq 0.4$ are strongly robust with respect to the outliers giving excellent fits to the rest of the observations. In fact the slope parameter is practically constant for all $\alpha \geq 0.4$ (and the LMS). The Warwick and Jones optimal α parameters in this case are 0.486 and 0.631, and the corresponding estimators are well within the robust band described above. The least square estimators of the regression coefficients, after deleting the outlying observations corresponding to the years

TABLE 4
The parameter estimates for the Salinity data using minimum density power divergence approach for various values of the tuning parameter α and the LMS approach

	α									
	0	0.05	0.1	0.25	0.4	0.5	0.6	0.8	1	LMS
β_1	9.59	9.96	10.51	18.01	18.37	18.40	18.46	18.89	19.19	36.70
β_2	0.78	0.78	0.77	0.72	0.72	0.72	0.72	0.72	0.71	0.36
β_3	−0.03	−0.03	−0.04	−0.18	−0.20	−0.20	−0.20	−0.19	−0.18	−0.07
β_4	−0.30	−0.31	−0.33	−0.61	−0.63	−0.63	−0.63	−0.65	−0.66	−1.30
σ	1.23	1.23	1.22	0.97	0.91	0.87	0.83	0.76	0.71	—

1963 to 1970, are −5.1644 and 0.1085 respectively, quite close to all our robust estimators.

7.3. Salinity data

Finally, as an example of the multiple regression model with masking effects, we consider the "Salinity data" (Table 5, Chapter 3, Rousseeuw and Leroy, 1987 [12]) that were originally presented by Ruppert and Carroll (1980)[13]. The data set contains measurements of the salt concentration of the water and the river discharge taken in North Carolina's Pamlico Sound. Rousseeuw and Leroy (1987) [12] consider this data as a multiple linear model with salinity as the dependent variable and the independents variables being salinity lagged by two weeks (x_1), the number of biweekly periods elapsed since the beginning of the spring season (x_2), and the volume of river discharge into the sound (x_3). According to the physical description of the data given by Carroll and Ruppert (1985) [3], cases 5 and 16 in the data correspond to periods of very heavy discharge but Rousseeuw and Leroy contend that the cases 3 and 16 conspire to hide the discrepant number 5 producing the masking effect in the data.

Table 4 presents the estimators of the regression coefficients obtained by the minimum DPD estimators for several α. To get a clear understanding of the robustness properties of the estimators, we also present the residual plots for some of the estimators along with the LMS estimators for comparison in Figure 3.

It was already observed by Rousseeuw and Leroy (1987) [12] that the ordinary least square (OLS) estimators of the regression parameters (which are also the MLEs and correspond to $\alpha = 0$ within the DPD class) can not separate out the influential 5th and 16th observations due to the masking effect. However the robust LMS estimator remains unaffected by the masking effect. The 16th observation clearly stands out as a major outlier while the 5th observation is among the other high residual observations. In case of the minimum DPD estimators also the 16th observation clearly stands out unlike the OLS case. The 5th observation also has a large residual although it is less prominent than the LMS case. The minimum DPD estimator also produces residuals which are more evenly spread on either side of zero compared to the LMS residuals which are somewhat biased on the positive side.

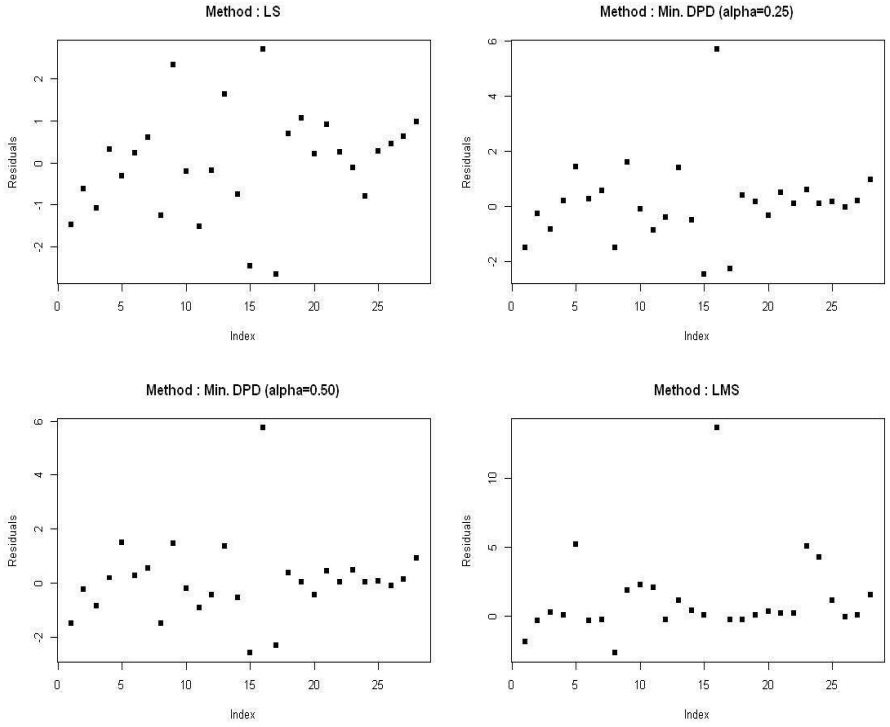

FIG 3. *Residual Plots of the fitted regression models for Salinity data using Least Squares (LS), Least Median of Squares (LMS) and minimum DPD estimation for several* α.

If we fit the linear regression ignoring the influential 5th and 16th observations, the OLS estimators of the regression coefficients are $(23.3862, 0.6998, -0.2500, -0.8356)$. These results are again very similar to those obtained by the minimum DPD estimators obtained for the full data when $\alpha \geq 0.25$. For both our pilot estimators the Warwick and Jones optimal value of α for this data set is equal to 1.

8. Concluding remarks

Problems which involve non-homogeneous independent data frequently arise in real life situations but are not always straightforward to deal with. Depending on the area of application, the number or robust options for inference may be limited. We trust that this work provides a general robust technique for the experimenter to deal with such situations.

In this paper we have chosen to illustrate the method developed on the most obvious domain of application, which is linear regression with normal errors. This, however, is by way of illustration only and the method applies generally

to any other domain where robust methods could be of use in case of non-homogeneous data. Among possible extensions which we propose to undertake in the near future is the Poisson regression problem; in terms of practical necessity it is another important area of application but, unlike the normal linear regression case where robust methods abound, has few options for robust inference.

Another obvious case of interest, which has not been addressed in the present paper, is the other fundamental paradigm of statistical inference – the hypothesis testing problem. In future we propose to extend the method to develop robust and efficient tests of hypothesis involving the common parameter of interest.

This paper has implicitly assumed that the case being dealt with involves individual observations from different distributions. Another situation where this method can be easily made useful is the multiple sample case where independent and identically distributed samples are obtained from different populations which involve the same parameter in their distribution. It is obvious, however, the overall divergence constructed in such situations must be a weighted average with weights proportional to individual the sample sizes.

Acknowledgements

The authors gratefully acknowledge the comments of the referees and the editorial board which helped to significantly improve the manuscript.

Appendix A: Proofs of the results

A.1. Proof of Theorem 3.1

Proof of consistency. To prove the consistency, we will consider the behavior of the density power divergence or equivalently $H_n(\theta)$ on a sphere Q_a with the center at the best fitting parameter θ^g and radius a. We will show that for any sufficiently small a the probability tends to 1 that $H_n(\theta) > H_n(\theta^g)$ for all points θ on the surface of Q_a, and hence that $H_n(\theta)$ has a local minimum in the interior of Q_a. Since at a local minimum the estimating equation (2.4) must be satisfied it will follow that for any $a > 0$ with probability tending to 1 as $n \to \infty$, the estimating equation (2.4) has a solution $\hat{\theta}_n(a)$ within Q_a.

To study the behavior of $H_n(\theta)$ on Q_a, we expand $H_n(\theta)$ by the Taylor series expansion around θ^g to get

$$\begin{aligned}&\frac{[H_n(\theta^g) - H_n(\theta)]}{1+\alpha}\\ &= \sum_j (-A_j)(\theta_j - \theta_j^g)\\ &\quad + \frac{1}{2}\sum_j \sum_k (-B_{jk})(\theta_j - \theta_j^g)(\theta_k - \theta_k^g)\end{aligned}$$

$$+\frac{1}{6}\sum_j\sum_k\sum_l(\theta_j-\theta_j^g)(\theta_k-\theta_k^g)(\theta_l-\theta_l^g)\frac{1}{n}\sum_{i=1}^n\gamma_{jkl}^{(i)}(y_i)M_{jkl}^{(i)}(y_i)$$
$$=S_1+S_2+S_3$$

where

$$A_j=\frac{1}{1+\alpha}\nabla_j H_n(\theta)|_{\theta=\theta^g}=\frac{1}{1+\alpha}\frac{1}{n}\sum_{i=1}^n\nabla_j V_i(Y_i;\theta)|_{\theta=\theta^g},$$

$$B_{jk}=\frac{1}{1+\alpha}\nabla_{jk} H_n(\theta)|_{\theta=\theta^g}=\frac{1}{1+\alpha}\frac{1}{n}\sum_{i=1}^n\nabla_{jk} V_i(Y_i;\theta)|_{\theta=\theta^g},$$

with ∇_j and ∇_{jk} representing the partial derivatives with the indicated components of θ and $0\le|\gamma_{jkl}(x)|\le 1$.

First note that for each i,j,

$$E_{g_i}\left[\nabla_j V_i(Y_i;\theta)\right]|_{\theta=\theta^g}=\nabla_j H_i(\theta^g)=0.$$

Thus by Assumption (A6), Equation (3.8) and a generalized version of the Khinchin's weak law of large numbers it follow that, for each j,

$$A_j\to 0\quad\text{ in }\quad L_1\quad\text{ and hence in probability.}$$

Therefore for any given a we have $|A_j|<a^2$, and hence $|S_1|<pa^3$ with probability tending to 1.

Similarly, by Assumption (A6), Equation(3.9) and by the definition of the matrix Ψ_n, it follows that for any j,k, $B_{jk}-(\Psi_n)_{jk}\to 0$ with probability tending to one. Consider the representation

$$\begin{aligned}2S_2 &= \sum\sum[-(\Psi_n)_{jk}(\theta_j-\theta_j^g)(\theta_k-\theta_k^g)]\\ &\quad+\sum\sum\{-B_{jk}+(\Psi_n)_{jk}\}(\theta_j-\theta_j^g)(\theta_k-\theta_k^g).\end{aligned}$$

For the second term in the above equation it follows from an argument similar to that for S_1 that its absolute value is less than p^2a^3 with probability tending to 1. The first term is a negative (nonrandom) quadratic form in the variables $(\theta_j-\theta_j^g)$ by Assumption (A4). By an orthogonal transformation this can be reduced to a diagonal form $\sum_i\lambda_i(\xi_i)^2$ with $\sum_i(\xi_i)^2=a^2$. The quantities λ_i and ξ_i are also function of n, which has been suppressed here for brevity. As each λ_i is negative, by ordering them and using Assumption (A4), one gets $\sum_i\lambda_i(\xi_i)^2\le-\lambda_0a^2$. Combining the first and the second terms, there exist $c>0$, $a_0>0$ such that for $a<a_0$, $S_2<-ca^2$, with probability tending to 1.

Finally, by Assumption (A5), with probability tending to 1, $|\frac{1}{n}\sum M_{jkl}^{(i)}(Y_i)|<2m_{jkl}<\infty$, and hence $|S_3|<ba^3$ on Q_a where $b=\frac{1}{3}\sum\sum\sum m_{jkl}$. Combining these inequalities, we see that

$$\max(S_1+S_2+S_3)<-ca^2+(b+s)a^3,$$

which is less than zero if $a<c/(b+s)$.

Thus, for sufficiently small a there exists a sequence of roots $\hat{\theta} = \hat{\theta}(a)$ such that $P(||\hat{\theta}-\theta||_2 < a) \to 1$ where $||\cdot||_2$ represents the L_2 norm. It remains to show that we can determine such a sequence independently of a. Let θ^* be the root closest to θ. This exists because the limit of a sequence of roots is again a root by the continuity of $H_n(\theta)$ as a function of θ. Then clearly $P(||\theta^*-\theta||_2 < a) \to 1$ for all $a > 0$. This concludes the proof of the existence of a sequence of consistent solutions to the estimating equation (2.4) with probability tending to 1. □

Proof of asymptotic normality. Now for the proof of the asymptotic normality of the minimum density power divergence estimator $\hat{\theta}_n$, we expand $H_n^j(\theta) = \nabla_j H_n(\theta)$ about θ^g to obtain:

$$H_n^j(\theta) = H_n^j(\theta^g) + \sum_k (\theta_k - \theta_k^g) H_n^{jk}(\theta^g) + \frac{1}{2}\sum_k \sum_l (\theta_k - \theta_k^g)(\theta_l - \theta_l^g) H_n^{jkl}(\theta^*)$$

where θ^* is a point on the line segment connecting θ and θ^g, and H_n^{jk} and H_n^{jkl} denote the indicated second and third partial derivatives of H_n. But since $H_n^j(\hat{\theta}) = 0$, evaluating the above at $\theta = \hat{\theta}_n$, we get

$$n^{1/2}\sum_k (\hat{\theta}_k - \theta_k^g)\left[H_n^{jk}(\theta^g) + \frac{1}{2}\sum_l (\hat{\theta}_l - \theta_l^g) H_n^{jkl}(\theta^*)\right] = -n^{1/2} H_n^j(\theta^g).$$

This has the form

$$\sum A_{jkn} Z_{kn} = T_{jn} \tag{A.1}$$

with

$$Z_{kn} = n^{1/2}(\hat{\theta}_k - \theta_k^g),$$
$$A_{jkn} = \left[H_n^{jk}(\theta^g) + \frac{1}{2}\sum_{l=1}^{p} (\hat{\theta}_l - \theta_l^g) H_n^{jkl}(\theta^*)\right],$$

and

$$T_{jn} = -n^{1/2} H_n^j(\theta^g).$$

In vector notation, we can rewrite equation (A.1) as

$$A_n Z_n = T_n, \tag{A.2}$$

$Z_n = (Z_{1n}, \ldots, Z_{pn})'$, $T_n = (T_{1n}, \ldots, T_{pn})'$ and $A_n = ((A_{jkn}))_{j=1,\ldots,p;k=1,\ldots,p}$. Note that $T_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n \nabla V_i(Y_i;\theta^g)$. A simple calculation shows that $E_{g_i}[V_i(Y_i;\theta^g)] = 0\ \forall i$ and each $V_i(Y_i;\theta^g)$ are independent with zero mean and finite variances. So by Assumption (A7) and a multivariate extension of the Lindebarg-Levy CLT, it follows that

$$\frac{1}{1+\alpha}\Omega_n^{-\frac{1}{2}} T_n \xrightarrow{\mathcal{D}} N_p(0, I_p).$$

Thus using equation (A.2), we get

$$\frac{1}{1+\alpha}\Omega_n^{-\frac{1}{2}} A_n Z_n \xrightarrow{\mathcal{D}} N_p(0, I_p). \tag{A.3}$$

Next from Assumption (A5) it follows that $H_n^{jkl}(\theta^*)$ is bounded with probability tending to 1, so that the consistency of $\hat{\theta}$ implies that the second term of A_{jkn} converges to zero in probability. Further as in the proof of the consistency part, $\frac{1}{1+\alpha}H_n^{jk}(\theta^g) - (\Psi_n)_{jk} \to^P 0$ for all j, k, and thus it follows that

$$\Omega_n^{-\frac{1}{2}}\left[\frac{1}{1+\alpha}A_n - (\Psi_n)\right] Z_n \xrightarrow{P} 0.$$

Combining this with equation(A.3), we finally get

$$\Omega_n^{-\frac{1}{2}}\Psi_n Z_n \xrightarrow{\mathcal{D}} N_p(0, I_p). \qquad \square$$

A.2. Proof of Lemma 5.1

Since both g_i and h_i belongs to the location-scale model family as specified above, we can assume that $h_i(x) = f(\frac{y-l_i(\mu_1)}{\sigma})$ and $g_i(x) = f(\frac{y-l_i(\mu_2)}{\sigma})$. Then we need to show that $\int D_\alpha(\epsilon g_i, h_i)$ is minimized when $\mu_1 = \mu_2$. Now note that

$$\int D_\alpha(\epsilon g_i, h_i) = \int h_i^{1+\alpha} - \frac{1+\alpha}{\alpha}\,\epsilon^\alpha \int h_i^\alpha g_i + \frac{\epsilon^{1+\alpha}}{\alpha} \int g_i^{1+\alpha}. \tag{A.4}$$

Because of the special form of the densities g_i and h_i the first and the last integral of the above will be independent of the parameter μ_2 and μ_1 respectively. Thus it is enough to prove that the middle integral $\int h_i^\alpha g_i$ is maximized at $\mu_1 = \mu_2$. But

$$\int h_i^\alpha g_i = \frac{1}{\sigma^{1+\alpha}} \int f\left(\frac{y-l_i(\mu_1)}{\sigma}\right)^\alpha f\left(\frac{y-l_i(\mu_2)}{\sigma}\right) dy = \frac{1}{\sigma^\alpha}\int f(y)^\alpha f(y+\eta)dy$$

where $\eta = \frac{l_i(\mu_1)-l_i(\mu_2)}{\sigma}$. Since $\alpha > 0$, an application of Holder's inequality gives

$$\int f(y)^\alpha f(y+\eta)dy \le \left(\int f(y)^{1+\alpha}dy\right)^{\frac{\alpha}{1+\alpha}} \left(\int f(y+\eta)^{1+\alpha}dy\right)^{\frac{1}{1+\alpha}} = \int f^{1+\alpha}$$

with equality iff $\eta = 0$ or $l_i(\mu_1) = l_i(\mu_2)$ or $\mu_1 = \mu_2$ by the one-to-one property of $l_i(.)$. This proves the lemma. $\square$

A.3. Proof of Theorem 5.2

As mentioned before let n be fixed and $\mu_n = T_\alpha^\mu(H_{1,\epsilon,m}, \ldots, H_{n,\epsilon,m})$ and $\theta_m = (\mu_m, \sigma)$ where ϵ denotes a fixed level of contamination. If breakdown occurs at

the model, there exists sequences $\{K_{i,m}\}$ of model densities such that $|\theta_m| \to \infty$ as $m \to \infty$. Now, fix an i and consider

$$\begin{aligned} d_\alpha(h_{i,\epsilon,m}(\cdot), f_i(\cdot;\theta_m)) &= \int_{A_{i,m}} D_\alpha(h_{i,\epsilon,m}(\cdot), f_i(\cdot;\theta_m)) \\ &\quad + \int_{A_{i,m}^c} D_\alpha(h_{i,\epsilon,m}(\cdot), f_i(\cdot;\theta_m)) \end{aligned} \tag{A.5}$$

where $A_{i,m} = \{x : g_i(x) > \max(k_{i,m}(x), f_i(x;\theta_m))\}$ and $D(g,f)$ is as defined before.

Now since g_i belongs to the model family $\mathcal{F}_{i,\theta}$, from (BP1), $\int_{A_{i,m}} k_{i,m}(x) \to 0$, and from (BP2), $\int_{A_{i,m}} f_i(x;\theta_m) \to 0$, so under $k_{i,m}(\cdot)$ and $f_i(\cdot;\theta_m)$, the set $A_{i,m}$ converges to a set of zero probability as $m \to \infty$. Thus, on $A_{i,m}$,

$$D_\alpha(h_{i,\epsilon,m}(\cdot), f_i(\cdot;\theta_m)) \to D_\alpha((1-\epsilon)g_i, 0) \quad \text{as} \quad m \to \infty$$

and so by the dominated convergence theorem (DCT)

$$\left| \int_{A_{i,m}} D_\alpha(h_{i,\epsilon,m}(x), f_i(x;\theta_m))dx - \int_{A_{i,m}} D_\alpha((1-\epsilon)g_i(x), 0)dx \right| \to 0. \tag{A.6}$$

And further by (BP1) and (BP2) we have

$$\left| \int_{A_{i,m}} D_\alpha((1-\epsilon)g_i, 0) - \int_{g_i>0} D_\alpha((1-\epsilon)g_i, 0) \right| \to 0. \tag{A.7}$$

Thus by (A.6) and (A.7) above

$$\left| \int_{A_{i,m}} D_\alpha(h_{i,\epsilon,m}(x), f_i(x;\theta_m))dx - \int_{g_i>0} D_\alpha((1-\epsilon)g_i(x), 0)dx \right| \to 0. \tag{A.8}$$

But we have $D_\alpha((1-\epsilon)g_i, 0) = \frac{(1-\epsilon)^{1+\alpha}}{\alpha\sigma^\alpha} M_f^\alpha$. Hence

$$\int_{A_{i,m}} D_\alpha(h_{i,\epsilon,m}(x), f_i(x;\theta_m))dx \to \frac{(1-\epsilon)^{1+\alpha}}{\alpha\sigma^\alpha} M_f^\alpha. \tag{A.9}$$

Next by (BP1) and (BP2), $\int_{A_{i,m}^c} g_i(x) \to 0$ as $n \to \infty$, so under $g_i(.)$, the set $A_{i,m}^c$ converges to a set of zero probability. Hence similarly, we get

$$\left| \int_{A_{i,m}^c} D_\alpha(h_{i,\epsilon,m}(x), f_i(x;\theta_m))dx - \int D_\alpha(\epsilon k_{i,m}(x), f_i(x;\theta_m))dx \right| \to 0. \tag{A.10}$$

Now by (BP3), we have

$$\int D_\alpha(\epsilon k_{i,m}(x), f_i(x;\theta_m))dx \quad \geq \quad \int D_\alpha(\epsilon f_i(x;\theta_m), f_{\theta_m})$$

$$= \int f_i(x;\theta_m)^{1+\alpha} C_\alpha(\epsilon-1)$$
$$= \frac{C_\alpha(\epsilon-1)}{\sigma^\alpha} M_f^\alpha.$$

Using (A.8), (A.9) and (A.10), we get

$$\liminf_{m\to} d_\alpha(h_{i,\epsilon,m}(.), f_i(.;\theta_m)) \geq \frac{C_\alpha(\epsilon-1)}{\sigma^\alpha} M_f^\alpha + \frac{(1-\epsilon)^{1+\alpha}}{\alpha\sigma^\alpha} M_f^\alpha.$$

Since above is true for all $i = 1, \ldots, n$, taking average over i, we get

$$\liminf_{m\to} \frac{1}{n}\sum_{i=1}^n d_\alpha(h_{i,\epsilon,m}(.), f_i(.;\theta_m)) \geq \frac{C_\alpha(\epsilon-1)}{\sigma^\alpha} M_f^\alpha + \frac{(1-\epsilon)^{1+\alpha}}{\alpha\sigma^\alpha} M_f^\alpha$$
$$= a_1(\epsilon) \text{ (say)}. \tag{A.11}$$

We will have a contradiction to our assumption that $\{k_{i,m}\}$ are sequences for which breakdown occurs if we can show that there exists a constant value θ^* in the parameter space such that for the same sequences $\{k_{i,m}\}$,

$$\limsup_{m\to\infty} \frac{1}{n}\sum_{i=1}^n d_\alpha(h_{i,\epsilon,m}(.), f_i(.;\theta_m)) < a_1(\epsilon) \tag{A.12}$$

as then the $\{\theta_m\}$ sequence above could not minimize $\frac{1}{n}\sum_{i=1}^n d_\alpha(h_{i,\epsilon,m}(.), f_i(.;\theta_m))$ for every m.

We will now show that equation (A.12) is true for all $\epsilon < 1/2$ under the model when we choose θ^* to be the true value $\theta^g = (\mu^g, \sigma)$ of the parameter. For any fixed i and θ, let $B_{i,m} = \{x : k_{i,m}(x) > max(g_i(x), f_i(x;\theta))\}$. Since g_i belongs to the model $\mathcal{F}_{i,\theta}$, from (BP1) we get $\int_{B_{i,m}} g_i(x) \to 0$ and $\int_{B_{i,m}} f_\theta(x) \to 0$. Similarly from (BP1), $\int_{B_{i,m}^c} k_{i,m} \to 0$ as $m \to \infty$. Thus, under $k_{i,m}$, the set $B_{i,m}^c$ converges to a set of zero probability, while under g_i and $f_i(\cdot;\theta)$, the set $B_{i,m}$ converges to a set of zero probability. Thus, on $B_{i,m}$, $D_\alpha(h_{i,\epsilon,m}(\cdot), f_i(\cdot;\theta_m)) \to D_\alpha(\epsilon K_{i,m}, 0)$ as $m \to \infty$. So, by DCT

$$\left| \int_{B_{i,m}} D_\alpha(h_{i,\epsilon,m}(x), f_i(x;\theta_m))dx - \int_{k_{i,m}>0} D_\alpha(\epsilon k_{i,m}(x), 0)dx \right| \to 0.$$

But as before for $k_{i,m} > 0$ we have $D_\alpha(\epsilon k_{i,m}, 0) = \frac{\epsilon^{1+\alpha}}{\alpha} \int k_{i,m}^{1+\alpha}$ and hence

$$\left| \int_{B_{i,m}} D_\alpha(h_{i,\epsilon,m}(x), f_i(x;\theta_m))dx - \frac{\epsilon^{1+\alpha}}{\alpha} \int k_{i,m}^{1+\alpha} \right| \to 0.$$

Similarly we have

$$\left| \int_{B_{i,m}^c} D_\alpha(h_{i,\epsilon,m}(x), f_i(x;\theta_m))dx - \int D_\alpha((1-\epsilon)g_i(x), f_i(x;\theta)) \right| \to 0.$$

Therefore,

$$\limsup_{m\to\infty} d_\alpha(h_{i,\epsilon,m}(\cdot), f_i(\cdot;\theta)) = \int D_\alpha((1-\epsilon)g_i(\cdot), f_i(\cdot;\theta)) + \frac{\epsilon^{1+\alpha}}{\alpha} \limsup_{m\to\infty} \int k_{i,m}^{1+\alpha}.$$

Averaging over $i = 1, \ldots, n$, we get

$$\limsup_{m\to\infty} \frac{1}{n}\sum_{i=1}^{n} d_\alpha(h_{i,\epsilon,m}(\cdot), f_i(\cdot;\theta)) = \frac{1}{n}\sum_{i=1}^{n} \int D_\alpha((1-\epsilon)g_i(\cdot), f_i(\cdot;\theta)) + \frac{\epsilon^{1+\alpha}}{\alpha} \limsup_{m\to\infty} \frac{1}{n}\sum_{i=1}^{n} \int k_{i,m}^{1+\alpha} \quad \text{(A.13)}$$

However note that since $g_i(y) = f_i(y;\theta^g)$, using Lemma 5.1, each $D_\alpha((1-\epsilon)g_i(\cdot), f_i(\cdot;\theta))$ is minimized over θ at $\theta = \theta^g$ and

$$D_\alpha((1-\epsilon)g_i(\cdot), f_i(\cdot;\theta^g)) = D_\alpha((1-\epsilon)f_i(\cdot;\theta^g), f_i(\cdot;\theta^g)) = \frac{C_\alpha(-\epsilon)}{\sigma^\alpha} M_f^\alpha.$$

So taking $\theta = \theta^g$ in above equation (A.13)and then using (BP3), we get

$$\begin{aligned} & \limsup_{m\to\infty} \frac{1}{n}\sum_{i=1}^{n} d_\alpha(h_{i,\epsilon,m}(\cdot), f_i(v;\theta^g)) \\ = \; & \frac{C_\alpha(-\epsilon)}{\sigma^\alpha} M_f^\alpha + \frac{\epsilon^{1+\alpha}}{\alpha} \limsup_{m\to\infty} \frac{1}{n}\sum_{i=1}^{n} \int k_{i,m}^{1+\alpha} \\ \leq \; & \frac{C_\alpha(-\epsilon)}{\sigma^\alpha} M_f^\alpha + \frac{\epsilon^{1+\alpha}}{\alpha\sigma^\alpha} M_f^\alpha = a_3(\epsilon) \quad \text{say.} \end{aligned} \quad \text{(A.14)}$$

Consequently, asymptotically there is no breakdown for ϵ level contamination when $a_3(\epsilon) < a_1(\epsilon)$. Notice that $a_1(\epsilon)$ and $a_3(\epsilon)$ are strictly decreasing and increasing respectively in ϵ and $a_1(1/2) = a_3(1/2)$, so that asymptotically there is no breakdown and

$$\limsup_{n\to\infty} |T_\alpha(H_{1,\epsilon,m}), \ldots, T_\alpha(H_{n,\epsilon,m})| < \infty$$

for $\epsilon < 1/2$. □

A.4. Proof of Lemma 6.1

Since the true density belongs to the model family and the model density is normal with mean $x_i^T\beta$ and variance σ^2, Assumptions (A1)–(A2) follow directly from the property of the normal density function. Assumption (A4) follows from Equation (6.12) of Assumption (R2). Considering the form of $V_i(y_i;\theta,x_i)$, Assumption (A5) follows from equation (6.11) of assumption (R1). To prove Equation (3.8), take any $j = 1, \ldots, p$ and consider

$$\nabla_j V_i(y_i;\theta,x_i) = \kappa e^{-\frac{\alpha(y_i - x_i^T\beta)^2}{2\sigma^2}}(y_i - x_i^T\beta)x_{ij}$$

where $\kappa = -\frac{1+\alpha}{(2\pi)^{\alpha/2}\sigma^{\alpha+2}}$. Then

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} E_i\left[|\nabla_j V_i(y_i;\theta,x_i)| I(|\nabla_j V_i(y_i;\theta,x_i)| > N)\right] \\
= \;& |\kappa|\frac{1}{n}\sum_{i=1}^{n} E_i\left[e^{-\frac{\alpha(y_i - x_i^T\beta)^2}{2\sigma^2}}|y_i - x_i^T\beta||x_{ij}|\right. \\
&\qquad\qquad \left. \times I(e^{-\frac{\alpha(y_i - x_i^T\beta)^2}{2\sigma^2}}|y_i - x_i^T\beta||x_{ij}| > \frac{N}{|\kappa|})\right] \\
\leq \;& |\kappa|\frac{1}{n}\sum_{i=1}^{n}|x_{ij}| E_i\left[e^{-\frac{\alpha z_i^2}{2\sigma^2}}|z_i| I(e^{-\frac{\alpha z_i^2}{2\sigma^2}}|z_i| > \frac{N}{|\kappa|(\sup\limits_{n>1}\max\limits_{1\leq i\leq n}|x_{ij}|)})\right] \\
&\qquad (z_i = y_i - x_i^T\beta) \\
= \;& |\kappa| E_1\left[e^{-\frac{\alpha z_1^2}{2\sigma^2}}|z_1| I(e^{-\frac{\alpha z_1^2}{2\sigma^2}}|z_1| > \frac{N}{|\kappa|(\sup\limits_{n>1}\max\limits_{1\leq i\leq n}|x_{ij}|)})\right]\left(\frac{1}{n}\sum_{i=1}^{n}|x_{ij}|\right) \\
&\qquad (z_i\text{'s are i.i.d.}).
\end{aligned}$$

Now by DCT, we have

$$\lim_{N\to\infty} E_1\left[e^{-\frac{\alpha z_1^2}{2\sigma^2}}|z_1| I(e^{-\frac{\alpha z_1^2}{2\sigma^2}}|z_1| > \frac{N}{|\kappa|(\sup\limits_{n>1}\max\limits_{1\leq i\leq n}|x_{ij}|)})\right] = 0$$

because $\sup_{n>1}\max_{1\leq i\leq n}|x_{ij}| = O(1)$ by Assumption (R1). Also, we have

$$\sup_{n>1}\left(\frac{1}{n}\sum_{i=1}^{n}|x_{ij}|\right) \leq \sup_{n>1}\max_{1\leq i\leq n}|x_{ij}| = O(1).$$

Thus Equation (3.8) of (A6) follows for all $j = 1,\ldots,p$. Similarly, it follows for $j = p+1$ also. Further the Assumption (A6), Equation (3.9) and Assumption (A7) also holds similarly using Equation (6.10) and equation (6.13) respectively. □

A.5. Proof of Theorem 6.2

The consistency part follows from Lemma 6.1 and Theorem 3.1. Also, in view of Lemma 6.1, it follows from Theorem 3.1 that the asymptotic distribution of $\Omega_n^{-\frac{1}{2}}\Psi_n[\sqrt{n}(\hat{\theta}-\theta)]$ is $(p+1)$-dimensional normal with mean 0 and variance $I_{(p+1)}$. Now using the form of the matrices Ψ_n and Ω_n from equation (6.7) and (6.9) respectively, we get that

$$\Omega_n^{-\frac{1}{2}}\Psi_n[\sqrt{n}(\hat{\theta}-\theta)] = \begin{pmatrix} \frac{\zeta_\alpha}{\sqrt{\zeta_{2\alpha}}}(X^TX)^{\frac{1}{2}}(\hat{\beta}-\beta) \\ \\ \frac{\varsigma_\alpha}{\sqrt{\varsigma_{2\alpha}-\frac{\alpha^2}{4}\zeta_\alpha^2}}\sqrt{n}(\hat{\sigma}^2-\sigma^2) \end{pmatrix}$$

which asymptotically follows $N_{(p+1)}(0, I_{(p+1)})$ distribution. Hence it follows that the asymptotic distributions of $\frac{\zeta_\alpha}{\sqrt{\zeta_{2\alpha}}}(X^TX)^{\frac{1}{2}}(\hat{\beta}-\beta)$ and $\frac{\varsigma_\alpha}{\sqrt{\varsigma_{2\alpha}-\frac{\alpha^2}{4}\zeta_\alpha^2}}\sqrt{n}(\hat{\sigma}^2-\sigma^2)$ are independent $N_p(0, I_p)$ and $N(0,1)$ respectively. The theorem now follows immediately. □

A.6. Proof of Theorem 6.3

For $\alpha = 0$, the minimum DPD estimator is the same as the maximum likelihood estimator which is known to satisfy all the three equivariance properties in the case of linear regression. So let $\alpha > 0$. First note that

$$\begin{aligned}
&\hat{\beta}(\{(x_i, y_i) : i = 1, \ldots, n\}) \\
= \ &\arg_\beta \min_{(\beta^T\sigma)} \left[\frac{1}{(2\pi)^{\alpha/2}\sigma^\alpha\sqrt{1+\alpha}} - \frac{1+\alpha}{\alpha}\frac{1}{(2\pi)^{\alpha/2}\sigma^\alpha}\frac{1}{n}\sum_{i=1}^n e^{-\frac{\alpha(y_i - x_i^T\beta)^2}{2\sigma^2}} \right]
\end{aligned}$$

Now, for any column vector v of the same dimension as β, we have

$$\begin{aligned}
&\hat{\beta}(\{(x_i, y_i + x_i^Tv) : i = 1, \ldots, n\}) \\
= \ &\arg_\beta \min_{(\beta^T\sigma)} \left[\frac{1}{(2\pi)^{\alpha/2}\sigma^\alpha\sqrt{1+\alpha}} - \frac{1+\alpha}{\alpha}\frac{1}{(2\pi)^{\alpha/2}\sigma^\alpha}\frac{1}{n}\sum_{i=1}^n e^{-\frac{\alpha(y_i - x_i^T(\beta-v))^2}{2\sigma^2}} \right] \\
= \ &\arg_{(\beta-v)} \min_{((\beta-v)^T\sigma)} \left[\frac{1}{(2\pi)^{\alpha/2}\sigma^\alpha\sqrt{1+\alpha}} \right. \\
&\qquad \left. - \frac{1+\alpha}{\alpha}\frac{1}{(2\pi)^{\alpha/2}\sigma^\alpha}\frac{1}{n}\sum_{i=1}^n e^{-\frac{\alpha(y_i - x_i^T(\beta-v))^2}{2\sigma^2}} \right] + v \\
= \ &\hat{\beta}(\{(x_i, y_i) : i = 1, \ldots, n\}) + v
\end{aligned}$$

which shows that the estimator $\hat{\beta} = \hat{\beta}(\{(x_i, y_i) : i = 1, \ldots, n\})$ is regression equivariant.

Next, for any constant c, cY_i follows normal distribution with mean $x_i^T\beta$ and variance $c^2\sigma^2$ so that we have

$$\begin{aligned}
&\hat{\beta}(\{(x_i, cy_i) : i = 1, \ldots, n\}) \\
= \ &\arg_\beta \min_{(\beta^T\sigma)} \left[\frac{1}{(2\pi)^{\alpha/2}(|c|\sigma)^\alpha\sqrt{1+\alpha}} \right. \\
&\qquad \left. - \frac{1+\alpha}{\alpha}\frac{1}{(2\pi)^{\alpha/2}(|c|\sigma)^\alpha}\frac{1}{n}\sum_{i=1}^n e^{-\frac{\alpha(cy_i - x_i^T\beta)^2}{2c^2\sigma^2}} \right] \\
= \ &c\arg_{(\beta/c)} \min_{((\beta/c)^T\sigma)} \frac{1}{|c|^\alpha} \left[\frac{1}{(2\pi)^{\alpha/2}\sigma^\alpha\sqrt{1+\alpha}} \right. \\
&\qquad \left. - \frac{1+\alpha}{\alpha}\frac{1}{(2\pi)^{\alpha/2}\sigma^\alpha}\frac{1}{n}\sum_{i=1}^n e^{-\frac{\alpha(y_i - x_i^T(\beta/c))^2}{2\sigma^2}} \right] \\
= \ &c\hat{\beta}(\{(x_i, y_i) : i = 1, \ldots, n\})
\end{aligned}$$

This implies that the estimator $\hat{\beta} = \hat{\beta}(\{(x_i, y_i) : i = 1, \ldots, n\})$ is scale equivariant.

Finally, for any non-singular square matrix A,we get

$$\begin{aligned}
& \hat{\beta}(\{(A^T x_i, y_i) : i = 1, \ldots, n\}) \\
= \; & \arg_\beta \min_{(\beta^T \sigma)} \left[\frac{1}{(2\pi)^{\alpha/2}\sigma^\alpha \sqrt{1+\alpha}} - \frac{1+\alpha}{\alpha} \frac{1}{(2\pi)^{\alpha/2}\sigma^\alpha} \frac{1}{n} \sum_{i=1}^{n} e^{-\frac{\alpha(y_i - x_i^T A\beta)^2}{2\sigma^2}} \right] \\
= \; & A^{-1} \arg_{(A\beta)} \min_{((A\beta)^T \sigma)} \left[\frac{1}{(2\pi)^{\alpha/2}\sigma^\alpha \sqrt{1+\alpha}} \right. \\
& \qquad \left. - \frac{1+\alpha}{\alpha} \frac{1}{(2\pi)^{\alpha/2}\sigma^\alpha} \frac{1}{n} \sum_{i=1}^{n} e^{-\frac{\alpha(y_i - x_i^T (A\beta))^2}{2\sigma^2}} \right] \\
= \; & A^{-1} \hat{\beta}(\{(x_i, y_i) : i = 1, \ldots, n\})
\end{aligned}$$

proving the affine equivariance of the estimator $\hat{\beta} = \hat{\beta}(\{(x_i, y_i) : i = 1, \ldots, n\})$. □

References

[1] Andersen, R. (2008). *Modern Methods for Robust Regression*. SAGE Publications, Inc., Los Angeles, USA.

[2] Basu, A., Harris, I. R., Hjort, N. L. and Jones, M. C. (1998). Robust and efficient estimation by minimizing a density power divergence. *Biometrika*, **85**, 549–559. MR1665873

[3] Carroll, R. J. and Ruppert, D. (1985). Transformations in regression: A robust analysis. *Technometrics*, **27**, 1–12. MR0772893

[4] Durio, A. and Isaia, E. D. (2011). The minimum density power divergence approach in building robust regression models. *Informatica*, **22/1**, 43–56. MR2885658

[5] Ghosh, A. and Basu, A. (2013). Selection of the optimal robustness tuning parameter in fitting linear regression models: The density power divergence approach. Technical Report, BIRU, Indian Statistical Institute, Kolkata, India.

[6] Hampel, F. R. (1968). *Contributions to the Theory of Robust Estimation*. Ph.D. Thesis, University of California, Berkeley, U.S.A. MR2617979

[7] Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist Assoc.*, **69**, 383–393. MR0362657

[8] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, New York. MR0829458

[9] Huber, P. J. (1983). Minimax aspects of bounded-influence regression (with discussion). *J. Amer. Statist. Assoc.*, **78**, 66–80. MR0696850

[10] Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, Berlin. MR0620321

[11] Park, C. and Basu, A. (2004). Minimum disparity estimation: Asymptotic normality and breakdown point results. *Bulletin of Informatics and Cybernetics*, Special Issue in Honor of Professor Takashi Yanagawa, **36**, 19–33. MR2139489

[12] Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. John Wiley & Sons, New York. MR0914792

[13] Ruppert, D. and Carroll, R. J. (1980). Trimmed least squares estimation in the linear model. *J. Amer. Statist. Assoc.*, **75**, 828–838. MR0600964

[14] Simpson, D. G. (1987). Minimum Hellinger distance estimation for the analysis of count data. *J. Amer. Statist. Assoc.*, **82**, 802–807. MR0909985

[15] Warwick, J. and Jones, M. C. (2005). Choosing a robustness tuning parameter. *J. Stat. Comput. Simul.*, **75**, 581–588. MR2162547