Robust estimation for non-homogeneous data and the selection of the optimal tuning parameter: the density power divergence approach

Abhik Ghosh and Ayanendranath Basu*

Interdisciplinary Statistical Research Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata 700 108, India

(Received 7 October 2013; accepted 4 February 2015)

The density power divergence (DPD) measure, defined in terms of a single parameter $\alpha$, has proved to be a popular tool in the area of robust estimation [1]. Recently, Ghosh and Basu [5] rigorously established the asymptotic properties of the MDPDEs in case of independent non-homogeneous observations. In this paper, we present an extensive numerical study to describe the performance of the method in the case of linear regression, the most common setup under the case of non-homogeneous data. In addition, we extend the existing methods for the selection of the optimal robustness tuning parameter from the case of independent and identically distributed (i.i.d.) data to the case of non-homogeneous observations. Proper selection of the tuning parameter is critical to the appropriateness of the resulting analysis. The selection of the optimal robustness tuning parameter is explored in the context of the linear regression problem with an extensive numerical study involving real and simulated data.

Keywords: density power divergence; non-homogeneous data; tuning parameter; linear regression-model

AMS 2000 Subject Classification: Primary: 62F35; Secondary: 62J05

1. Introduction

Density-based minimum distance inference is an important component of parametric statistical inference. In minimum distance inference, the parameter of interest is estimated by minimizing a suitable measure of discrepancy between the data and an assumed parametric model. The
discrepancy is quantified by a divergence measure between two density functions or two distribution functions. In this paper, we will concentrate on the density-based approach. Consider a random sample \(X_1, \ldots, X_n\) of size \(n\) from a population whose true density function is \(g\). Suppose, we model \(g\) by a parametric family of densities \(\{f_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}\) and let \(\rho(f_1, f_2)\) be a density-based divergence measure between two probability densities \(f_1\) and \(f_2\) with respect to a common dominating measure. Then, the minimum distance estimator of the unknown parameter \(\theta\) based on the divergence \(\rho(\cdot, \cdot)\) is defined by the minimizer of \(\rho(\hat{g}_n, f_\theta)\) with respect to \(\theta\) over the parameter space \(\Theta\), where \(\hat{g}_n\) is some nonparametric estimate of the true density \(g\) based on the observed sample. In the case of discrete models the vector of relative frequencies provides a natural choice for \(\hat{g}_n\), but in the case of continuous models one generally needs to use a kernel density estimator in place of \(\hat{g}_n\). See Vajda [13], Pardo [10] and Basu et al. [2] for an elaborate description of the theory of minimum distance inference and several illustrative examples.

It is important to emphasise that the measures of divergence that are used in minimum distance estimation are not necessarily mathematical metrics. They need not be symmetric in their arguments, or may not satisfy the triangle inequality. The only properties that are demanded are that the measure should be non-negative and should equal zero if and only if the two arguments are identically equal. We refer to all such divergences as ‘statistical distances’ or simply ‘distances’.

Many of the minimum distance estimators based on density functions have strong robustness properties, which make them particularly useful in situations where model misspecifications are suspected or the data set contains outliers inconsistent with the model assumptions. One of the notable estimators in this connection is the minimum density power divergence estimator (MDPDE), which provides a natural downweighting of the outlying observations with a weight proportional to the density of the observation in relation to the assumed parametric model [1]. A particularly attractive feature of this method of estimation is that the objective function to be minimized can be constructed without taking recourse to any nonparametric density estimation technique like kernel density estimation or methods based on splines. As a result the method becomes substantially simpler, both in theory and practice, than minimum Hellinger distance estimation and related methods. Basu et al. [2] provide an extended description of the methods based on the DPD and other related minimum distance methods.

We now consider the situation where the observations are independent and share a common parameter in their distributions, but are not identically distributed; we refer to such a scenario as one involving non-homogeneous observations. To be applicable to the case of independent but non-homogeneous data, the minimum DPD method requires an extension of the theory existing in the case of independent and identically distributed (i.i.d.) data. Recently, Ghosh and Basu [5] have provided this extension. Part of the present paper deals with an extensive simulation study which illustrates the performance of the method in the case of linear regression. The results clearly bear out the desirable properties of the minimum DPD method, and complements the theoretical findings of Ghosh and Basu [5].

The DPD family is indexed by a single tuning parameter \(\alpha\), which controls the trade-off between robustness and efficiency with larger \(\alpha\) indicating enhanced robustness. To be useful in practice, the minimum DPD proposal must be accompanied by a recommendation for the choice of an optimal tuning parameter. In many real situations, the choice of \(\alpha\) has a significant effect on the value of the estimate. The present paper also deals with the selection of the optimal tuning parameter \(\alpha\) in problems where the data are independent but non-homogeneous. In the process, we appropriately extend the techniques that have been proposed for the selection of the optimal tuning parameter in the case of i.i.d. data. We propose this as a general method, which is not specific only to the regression case; however for the purpose of illustration we explore the optimal selection technique of the tuning parameter through extensive simulation studies and real data examples.
The rest of the paper is organized as follows. In Section 2, we briefly review the density power divergence (DPD) measure and discuss the application of this divergence in robust estimation, both in the case of i.i.d. and non-homogeneous data. In this section, we also develop the selection of the optimal tuning parameter in the case of non-homogeneous data and briefly discuss the mechanics of the method in the case of linear regression. Section 3 contains all the numerical illustrations of the paper which includes simulation studies, real data examples and selection of the optimal tuning parameter in specific examples. Section 4 has some concluding remarks.

2. The DPD and related inference

2.1 The i.i.d. data problem

The DPD measure \( d_\alpha(g,f) \) between the densities \( g \) and \( f \) is defined, in terms of the tuning parameter \( \alpha (\geq 0) \), as

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

\[
d_0(g,f) = \int g \ln \left( \frac{g}{f} \right), \quad (2)
\]

where \( \ln \) represents the natural logarithm. Given an i.i.d. random sample \( Y_1, \ldots, Y_n \) and a parametric family \( \mathcal{F} = \{ f_\theta : \theta \in \Theta \in \mathbb{R}^p \} \) of densities which models the true data-generating density \( g \), the minimum DPD estimator \( \hat{\theta}_\alpha \) of \( \theta \) is defined by the relation \( d_\alpha(\hat{g}, f_{\hat{\theta}_\alpha}) = \min_{\theta \in \Theta} d_\alpha(\hat{g}, f_\theta) \), where \( \hat{g} \) represents a suitable nonparametric estimate of \( g \). As already mentioned, a positive feature of the DPD measure is that it does not require any nonparametric smoothing technique for any value of \( \alpha \); substituting the empirical distribution function in place of the true unknown distribution function allows us to build suitable approximations to the divergence measures in Equation (1) or Equation (2). Simplifications of this nature are not possible in the case of most of the other density-based minimum distance methods (such as the minimum Hellinger distance estimator).

When estimating \( \theta \) by minimizing \( d_\alpha(g,f_\theta) \), the last part of the expression in the right-hand side of Equation (1) is independent of the unknown parameter; hence, the essential objective function to be minimized, after approximating the true distribution with the empirical, turns out to be

\[
\int f_\theta(y)^{1+\alpha} dy - \left( 1 + \frac{1}{\alpha} \right) \frac{1}{n} \sum_{i=1}^{n} f_\theta(Y_i)^\alpha. \quad (3)
\]

Under differentiability of the model, the MDPDE is obtained by solving the estimating equation

\[
\frac{1}{n} \sum_{i=1}^{n} u_\theta(Y_i) f_\theta(Y_i)^\alpha - \int u_\theta(y) f_\theta(y)^{1+\alpha} dy = 0, \quad (4)
\]

where \( u_\theta(y) = \nabla \ln f_\theta(y) \) is the usual likelihood score function and \( \nabla \) is the gradient with respect to \( \theta \). From the first term of Equation (4), it is clear that each likelihood score value is weighted by a power of the model density, so that observations unlikely under the model are automatically downweighted; thus, the robustness of the procedure is intuitively obvious. The strength of downweighting increases with \( \alpha \); \( \alpha = 0 \) indicates no downweighting, and in this case the MDPDE is the maximum likelihood estimator (MLE) of \( \theta \). The second term in Equation (4) is simply the adjustment needed to make the estimating equation unbiased under the model.
The popularity of the MDPDE is mainly due to its high robustness with minimal loss in efficiency and no use of nonparametric kernel smoothing even under continuous models. The MDPDE can also be seen to belong to the class of M-estimators \[9\]. Basu et al. \[1\] presented the asymptotic properties of the MDPDE along with its influence function (IF), which is a classical measure of the theoretical robustness of any estimator \[7\]. Whenever the true density \( g \) belongs to the model family, i.e. \( g = f_\theta \), the IF of the MDPDE (\( T_\alpha \), say) has an interesting form given by

\[
IF(y, T_\alpha, F_\theta) = \left( \int u_\theta u_\theta^T f_\theta^{1+\alpha} \right)^{-1} \left[ u_\theta (y) f_\theta^\alpha (y) - \int u_\theta f_\theta^{1+\alpha} \right], \tag{5}
\]

where \( y \) is the contamination point and \( F_\theta \) is the distribution function of \( f_\theta \). Clearly, the IF of the MDPDE is bounded at any \( \alpha > 0 \) for most parametric models, implying their robustness. Furthermore, Basu et al. \[1\] also showed that for the particular case of normal models with both parameters unknown, the breakdown point for the MDPDE of the mean parameter is \( \alpha(1 + \alpha)^{-3/2} \); recently Ghosh et al. \[6\] generalized this breakdown point result by showing the asymptotic breakdown point of 0.5 for the general location-scale model and a general class of minimum distance estimators including the MDPDEs with \( \alpha > 0 \).

### 2.2 The case of non-homogeneous observations

Adapting the DPD method to the case of non-homogeneous observations in a fully rigorous manner requires a substantial and nontrivial extension of the approach followed in the case of i.i.d. data. Ghosh and Basu \[5\] have recently provided this generalization. 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. Some examples of \( g_i \) are provided by sequences of normal distributions with different means \( \mu_i \) and a common variance \( \sigma^2 \), or a sequence of Gamma distributions with shape parameters \( \alpha_i \) and a common scale \( \beta \). A sequence of Bernoulli distributions with different success probabilities \( p_i \) may also be a useful example in the case of logistic regression, where each \( p_i \) may be linked to a common parameter \( \beta \) through the explanatory variables. We want to model \( g_i \) by the family \( F_i, \theta = \{f_i(\cdot; \theta) | \theta \in \Theta \} \) for all \( i = 1, 2, \ldots \); thus while the distributions are possibly different, they all share a common parameter \( \theta \).

Ghosh and Basu \[5\] proposed the minimization of the average divergence between the data points and the models, given by

\[
H_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} d_\alpha(\hat{g}_i, f_i(\cdot; \theta)) ,
\]

where \( \hat{g}_i \) is a density estimate of \( g_i \); ignoring the constant, the above may be approximated by

\[
H_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left[ \int f_i(y; \theta)^{1+\alpha} dy - \left( 1 + \frac{1}{\alpha} \right) f_i(Y_i; \theta)^\alpha \right] = \frac{1}{n} \sum_{i=1}^{n} V_i(Y_i; \theta), \tag{6}
\]

where \( V_i(\cdot; \theta) \) is the indicated term within the square brackets in the above equation. These authors also derived the asymptotic distribution of the MDPDE \( \hat{\theta}_n \). Under regularity conditions listed in \[5\], it was proved that \( \Omega_n^{-1/2} \psi_n n^{1/2}(\hat{\theta}_n - \theta_\alpha) \) has an asymptotic \( p \)-dimensional normal distribution with (vector) mean 0 and covariance matrix \( I_p \), the \( p \)-dimensional identity matrix,
where $\theta_\alpha$ is the minimizer of $(1/n) \sum_{i=1}^{n} d_a(g_i, f_i(y; \theta))$, $u_i(y; \theta) = \nabla \ln \hat{f}_i(y; \theta)$,

$$\Psi_n = \frac{1}{n} \sum_{i=1}^{n} J^{(i)} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + \alpha} E_{g_i}[\nabla \hat{f}_i(Y_i; \theta_\alpha)]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left[ \int u_i(y; \theta_\alpha) u_i^T(y; \theta_\alpha) f_i^{1+\alpha}(y; \theta_\alpha) \, dy \
- \int \{\nabla u_i(y; \theta_\alpha) + \alpha u_i(y; \theta_\alpha) u_i^T(y; \theta_\alpha)\} [g_i(y) - f_i(y; \theta_\alpha)] f_i(y; \theta_\alpha)^\alpha \, dy \right].$$

and

$$\Omega_n = \frac{1}{n} \sum_{i=1}^{n} \text{Var}_{g_i}[\nabla \hat{V}_i(Y_i; \theta_\alpha)]$$

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

where

$$\xi_i = \int u_i(y; \theta_\alpha) f_i(y; \theta_\alpha)^\alpha g_i(y) \, dy.$$  

Since the focal point of this proposal is the robustness of the estimator, one should check its IF to investigate this issue. Ghosh and Basu [5] extended the definition of IF from the i.i.d. case [7] to the present case of non-homogeneous observations and derived the same for the proposed MDPDE. In fact, in the case of the non-homogeneous setup, the contamination can be in any one direction or in all the directions with respect to the index $i$. When the contamination is only in the $i_0$th direction (at the contamination point $t_{i_0}$), the IF of the MDPDE $T_\alpha$ turns out to be [5]

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

where $G_i$ is the distribution function of $g_i$. On the other hand, when contamination exists in all the data points, the IF becomes [5]

$$IF(t_1, \ldots, t_n, T_\alpha, G_1, \ldots, G_n) = \Psi_n^{-1} \frac{1}{n} \sum_{i=1}^{n} [f_i(t_i; \theta)^\alpha u_i(t_i; \theta) - \xi_i],$$

where $t_i$ is the contamination point in the $i$th direction. In particular, letting $t_i = y$, $G_i = F_\theta$ and $f_i = f$ in above, we get back the IF of the MDPDE for the i.i.d. case given in Equation (5). See Ghosh and Basu [5] for a detailed discussion of the IF and summary measures based on it. For most parametric models, the IF of the MDPDE is again seen to be bounded for all $\alpha > 0$, implying robustness.

Ghosh and Basu [5] also proved the breakdown result for the MDPDE under a special location-scale type model $f_i(y; \theta) = (1/\sigma) f((y - l(\mu))/\sigma)$ with $\theta = (\mu, \sigma)$ under the non-homogeneous setup. Under Assumptions (BP1)–(BP3) of Ghosh and Basu [5], the asymptotic breakdown point of the MDPDE of the location parameter $\mu$ with fixed $\sigma$ is shown at least $1/2$ at the model distribution.
2.3 Selecting the optimal tuning parameter

We get the most efficient but highly non-robust MLE as the solution for the case $\alpha = 0$; as $\alpha$ increases so does the robustness of the estimator, albeit at the cost of efficiency. Thus, a natural question that arises in applying the method to practical real data scenarios is how to choose $\alpha$ appropriately. There have been some attempts to select the ‘optimal’ $\alpha$ for determining the MDPDE of $\theta$ in the i.i.d. case (see, e.g. [8,14]). We will modify the approach of Warwick and Jones in determining the optimal value of the tuning parameter for determining the MDPDEs of the estimators in the non-homogeneous data context.

For the sake of completeness, we briefly describe the method in case of i.i.d. data. Let $f_\theta$, be the target density, so that $\theta^*$ is the target parameter. Let $g(x) = (1 - \epsilon)f_\theta(x) + \epsilon\delta(x)$ be a contaminated version of the target density, where $\delta(x)$ is the Dirac delta function at the value $y$. In general $g$ need not involve the model density $f_\theta$ at all, but the above setup is useful in motivating the choice of the selection parameter. For a given $\alpha$, let $\hat{\theta}_\alpha$ denote the value of $\theta$ which minimizes $d_\alpha(g,f_\theta)$. In a given situation involving an i.i.d. random sample from $g$, let $\hat{\theta}_\alpha$ denote the MDPDE of $\theta$, Warwick and Jones [14] proposed to minimize an asymptotic approximation of the summed mean-squared error (MSE) $E[(\hat{\theta}_\alpha - \theta^*)^T(\hat{\theta}_\alpha - \theta^*)]$ to choose the optimal $\alpha$. Under standard regularity conditions, the asymptotic bias and variance of $\hat{\theta}_\alpha$ are given by $(\theta_\alpha - \theta^*)$ and $(1/n)J_\alpha(\theta_\alpha)^{-1}K_\alpha(\theta_\alpha)J_\alpha(\theta_\alpha)^{-1}$, respectively, where $J_\alpha(\theta)$ and $K_\alpha(\theta)$ are as given in Equations (3) and (4) of Warwick and Jones [14]. Thus, we have

$$E[(\hat{\theta}_\alpha - \theta^*)^T(\hat{\theta}_\alpha - \theta^*)] = (\theta_\alpha - \theta^*)^T(\theta_\alpha - \theta^*) + \frac{1}{n} \text{Trace}[J_\alpha(\theta_\alpha)^{-1}K_\alpha(\theta_\alpha)J_\alpha(\theta_\alpha)^{-1}].$$

(12)

The asymptotic variance component may be estimated by substituting $\hat{\theta}_\alpha$ for $\theta_\alpha$ and the sample average for the population average under $g$ in the definition of $J_\alpha(\theta)$ and $K_\alpha(\theta)$. To estimate the bias component, $\theta_\alpha$ may be substituted by its estimate $\hat{\theta}_\alpha$. But there is no obvious choice of $\theta^*$. Warwick and Jones [14] considered several ‘pilot’ estimators of $\theta$ as the input for $\theta^*$ in the minimization process and compared its effect through simulation studies. Based on their explorations, they recommended the use of the MDPDE corresponding to $\alpha = 1$ as the pilot estimate.

The approach of Hong and Kim [8] for choosing the optimal $\alpha$ may be considered as a special case of the above approach with the pilot estimator being the MDPDE $\hat{\theta}_\alpha$ itself. Thus, their minimization depends only on the second term of the right-hand side of Equation (12).

We will now generalize the approach of Warwick and Jones [14] for determining the optimal value of the robustness tuning parameter $\alpha$ in our non-homogeneous setup. Note that there could be some other optimality criteria apart from the minimization of the MSE. However, the MSE is a popular and widely used criterion with easy interpretation and this criterion appears to work well in case of i.i.d. data [14]. Thus, we have chosen to continue with this particular optimality criterion in the present paper.

Now, from the asymptotic distribution of our proposed MDPDE as discussed in Section 2.2, it follows that asymptotically

$$E[(\hat{\theta}_\alpha - \theta^*)^T(\hat{\theta}_\alpha - \theta^*)] = (\theta_\alpha - \theta^*)^T(\theta_\alpha - \theta^*) + \frac{1}{n} \text{Trace}[\Psi^{-1}_n\Omega_n\Psi^{-1}_n].$$

(13)

Thus, we need to minimize an estimate of the above approximation in order to obtain an optimum value of $\alpha$. Based on the given non-homogeneous observations, we can estimate the second part of the above as $(1/n)\text{Trace}[\Psi^{-1}_n\Omega_n\Psi^{-1}_n]$, where $\Psi_n$ and $\Omega_n$ are some consistent estimators of $\Psi_n$ and $\Omega_n$, respectively, corresponding to the tuning parameter $\alpha$. The estimation of $\Psi_n$ can be done in a natural manner by replacing $\theta_\alpha$ by $\hat{\theta}_\alpha$ and the population average with respect to $g_i$'s...
by the corresponding sample values so that we have

\[ \hat{\psi}_{n,\alpha} = \frac{1}{n} \sum_{i=1}^{n} \hat{J}^{(i)} \]

\[ = \frac{1}{n} \sum_{i=1}^{n} \left[ \int \{ \nabla u_i(y; \hat{\theta}_\alpha) + (\alpha + 1) u_i(y; \hat{\theta}_\alpha) u_i^T(y; \hat{\theta}_\alpha) \} f_i^{1+\alpha}(y; \hat{\theta}_\alpha) \, dy \right. \]

\[ - \left. \{ \nabla u_i(y_i; \hat{\theta}_\alpha) + \alpha u_i(y_i; \hat{\theta}_\alpha) u_i^T(y_i; \hat{\theta}_\alpha) \} f_i(y_i; \hat{\theta}_\alpha)^\alpha \right]. \] (14)

However, we will normally have only one observation from each \( g_i \), and for such situations it is not possible to estimate the variance based on only one observation. Therefore, we must go beyond the Warwick–Jones proposal to estimate \( \Omega_n \), and for this purpose we replace \( \theta_\alpha \) by \( \hat{\theta}_\alpha \) and \( g_i \)s by the model densities \( f_i(\cdot; \hat{\theta}_\alpha) \). Thus, we use the approximation

\[ \hat{\Omega}_{n,\alpha} = \frac{1}{n} \sum_{i=1}^{n} \left[ \int u_i(y; \hat{\theta}_\alpha) u_i^T(y; \hat{\theta}_\alpha) f_i(y; \hat{\theta}_\alpha)^{2\alpha+1} \, dy - \hat{\xi}_i \hat{\xi}_i^T \right]. \] (15)

with

\[ \hat{\xi}_i = \int u_i(y; \hat{\theta}_\alpha) f_i(y; \hat{\theta}_\alpha)^{1+\alpha} \, dy. \] (16)

To estimate the squared bias term \( (\theta_\alpha - \theta^*)^T(\theta_\alpha - \theta^*) \), we replace \( \theta_\alpha \) by its estimate \( \hat{\theta}_\alpha \). However, as in the i.i.d. case, there is no obvious choice of \( \theta^* \). We will use several robust estimates of \( \theta \) as our ‘pilot’ estimator in the place of \( \theta^* \) in the minimization process. We will explore the role of the pilot estimators on the choice of optimum \( \alpha \) through several simulation studies and real data examples in the context of linear regression and make a recommendation at the end.

As in the case of independent and identically distributed observations, here also the approach of Hong and Kim [8] turns out to be a special case of the Warwick and Jones approach, where one minimizes only the variance component. This process may be viewed as one where the MDPDE itself is chosen to be the pilot estimate. This particular case of the pilot estimator is referred to as the ‘no bias’ case in the rest of the manuscript.

Note that several (intuitively) reasonable approximations have been used in this section. The closeness of these approximations may be of some concern to the reader; however, as we will see later in our numerical illustrations, this is indeed not a practical problem for the implementation of the proposed method.

### 2.4 An application: selecting the optimum \( \alpha \) in normal linear regression

The linear regression problem with normal errors constitutes the most prominent example of estimation under independent and non-homogeneous data. While the asymptotic properties of the MDPDEs in the regression setup have been established recently in a formal way by Ghosh and Basu [5], Basu et al. [1] had briefly touched upon the use of the DPD measure in fitting robust regression models; a slightly expanded discussion was provided in the University of Oslo Technical Report of the same set of authors (having the same title and same year of citation). Scott [12] gave an example involving the use of the MDPDE corresponding to \( \alpha = 1 \) case (the \( L_2 \) distance) in the regression scenario. Some simulation results illustrating the performance of the MDPDEs of the regression parameters were presented by Durio and Isaia [4].
Consider the linear regression model

\[ y_i = x_i^T \beta + \epsilon_i, \quad i = 1, \ldots, n, \]  

(17)

where the error \( \epsilon_i \)'s are i.i.d. normal with mean zero and variance \( \sigma^2 \), \( x_i^T = (1, x_{i1}, \ldots, x_{ip}) \) is the vector of predictor variables corresponding to the \( i \)th observation and \( \beta = (\beta_0, \beta_1, \ldots, \beta_p)^T \) represents the regression coefficients. We will assume that the \( x_i \)'s are fixed; under this setup the MDPDE of the parameter \( \theta = (\beta^T, \sigma^2)^T \) can be obtained by minimizing the expression in Equation (6) with \( f_i \) being the \( N(x_i^T \beta, \sigma^2) \) density.

Ghosh and Basu [5] presented the theoretical asymptotic properties and the robustness properties of the proposed MDPDE under the linear regression model. They have also demonstrated that, whenever the error variance \( \sigma^2 \) is fixed, the asymptotic breakdown point of the proposed MDPDE of regression coefficient \( \beta \) is exactly 0.5 at the model. The advantages of the proposed MDPDE of the regression parameter has also been discussed in comparison with the existing robust regression methods. In the present paper, for the sake of completeness, we only present the IF of the proposed MDPDE \((T^\beta_a, T^\sigma_a)\) of \((\beta, \sigma)\) under contamination only in the \( i_0 \)th direction as given by

\[
IF_{i_0}(t_{i_0}, T^\beta_a, G_1, \ldots, G_n) = (1 + \alpha)^{3/2} (X^T X)^{-1} x_{i_0} (t_{i_0} - x_{i_0}^T \beta) e^{-\alpha (t_{i_0} - x_{i_0}^T \beta)^2/2\sigma^2}, \tag{18}
\]

\[
IF_{i_0}(t_{i_0}, T^\sigma_a, G_1, \ldots, G_n) = \frac{2(1 + \alpha)^{5/2}}{n(2 + \alpha^2)} (t_{i_0} - x_{i_0}^T \beta)^2 - \sigma^2) e^{-\alpha (t_{i_0} - x_{i_0}^T \beta)^2/2\sigma^2} + \frac{2\alpha(1 + \alpha^2)}{n(2 + \alpha^2)}.	ag{19}
\]

Clearly, these IFs depend on the residual of regression and the predictor values and they are both bounded in the contamination point \( t_{i_0} \) for all \( \alpha > 0 \) and for any direction \( i_0 \). Hence, the proposed MDPDEs with \( \alpha > 0 \) are always robust with respect to contamination in any data point. However, the IFs at \( \alpha = 0 \) are clearly unbounded implying the non-robust nature of the corresponding MLEs. Similar robustness properties can also be observed for contamination in all the direction and the extent of robustness always increases as \( \alpha \) increases; see Ghosh and Basu [5] for details.

Here, we consider the problem of finding the optimum \( \alpha \) in the context of the above linear regression model. As explained earlier, we can find the expressions for the estimated bias and variance components, as a function of \( \alpha \), based on the observed data. Then, we will minimize it with respect to \( \alpha \) in \([0, 1]\) to obtain the optimum value of the robustness tuning parameter. The choice of the optimum \( \alpha \) depends on the choice of the ‘pilot’ estimator used in estimating the bias component. We will explore this dependency and the performance of the proposed methodology through simulation studies in the following sections.

The robustness of the MDPDEs increases with \( \alpha \), but at the cost of their asymptotic efficiency. We consider the estimators corresponding to \( \alpha > 1 \) to be of little practical use on account of their low efficiency, and feel that nothing will be lost in restricting the search of the optimal parameter to the interval \([0, 1]\) as we have done.

Durio and Isaia [4] have also proposed a strategy to find the optimum \( \alpha \) in the context of the normal linear regression method. Unlike the current proposal which is quick, simple and intuitive, their method involves an elaborate Monte-Carlo testing and has a high computational complexity.

3. Numerical studies

In this section, we numerically explore the properties of the proposed estimators along the aims of the present paper. We do this with both real and simulated data and in either case touch upon the
selection of the optimal tuning parameter together with a description of the robustness properties of the proposed method.

3.1 Real data examples

Example 1 (Hertzsprung–Russell data of the star cluster) The Hertzsprung–Russell diagram of the star cluster CYG OB1 contains 47 stars in the direction of Cygnus [11, Table 3, Chapter 2]. For these 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 analyzed by Rousseeuw and Leroy [11] 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. These four stars (with indices 11, 20, 30 and 34) are known as giants in astronomy and represent huge leverage points for the linear regression model.

The MDPDEs of the model parameters for the regression of \( y \) on \( x \) for different values of \( \alpha \) together with the least median of squares (LMS) estimates are presented in Table 1. It is clear that the four leverage points exert a significant pull on the MLEs of the regression parameters to the extent that the corresponding residuals fail to separate the two groups of data. Note that the MLEs of the regression parameters are also the MDPDEs corresponding to \( \alpha = 0 \), as well as the ordinary least-squares (OLS) estimators. However, like the highly robust (but inefficient) LMS estimator, the minimum DPD estimators with \( \alpha \geq 0.3 \) can successfully ignore the outliers and produce robust fits which can easily separate the two groups.

![Figure 1. Different regression fits for the Hertzsprung–Russell data.](image)

Table 1. The parameter estimates for the Hertzsprung–Russell data using the minimum DPD approach for various values of the tuning parameter \( \alpha \) and the LMS approach.

| \( \alpha \) | 0.0 | 0.05 | 0.1 | 0.3 | 0.4 | 0.5 | 0.6 | 0.8 | 1.0 | LMS |
|---|---|---|---|---|---|---|---|---|---|---|
| \( \beta_0 \) | 6.69 | 6.74 | 6.78 | -5.80 | -6.57 | -7.22 | -7.57 | -7.89 | -8.03 | -12.30 |
| \( \beta_1 \) | -0.39 | -0.40 | -0.41 | 2.44 | 2.62 | 2.76 | 2.84 | 2.91 | 2.95 | 3.90 |
| \( \sigma \) | 0.61 | 0.60 | 0.60 | 0.40 | 0.40 | 0.40 | 0.40 | 0.40 | 0.41 | — |
In Table 2, the optimal $\alpha$ corresponding to different pilot estimators are presented. The last column represents the Hong and Kim proposal. It is clearly seen that except for pilot estimators which correspond to MDPDEs with very small values of $\alpha$, the selected optimal value of $\alpha$ is fairly stable. In Figure 1, we plot the estimated OLS line, the estimated LMS line and the MDPDE estimator line with the optimal tuning parameter corresponding to the pilot estimator given by the MDPDE at $\alpha = 0.5$; from Table 2 it may be seen that this optimal parameter is $\alpha = 0.93$.

**Example 2 (Belgium telephone call data)** This example has a heavy contamination in the $y$-direction. The data set [11, Table 2, Chapter 2] is the result of a 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. Due to the use of another recording system (giving the total number of minutes of these calls) from 1964 to 1969, the data contain large outliers in the $y$-direction in that range. The years 1963 and 1970 are also partially affected.

The estimates of the regression coefficients for the regression of the number of calls on the independent variable (year) obtained by the minimum DPD estimation with several $\alpha$ are presented in Table 3. Clearly, the MLE corresponding to $\alpha = 0$ are highly affected by the outliers but like the LMS method the MDPDEs with $\alpha \geq 0.3$ are highly resistant to the outliers giving excellent fit to the rest of the observations. Notice that in this case the estimated optimal tuning parameter is much more variable (see Table 4); this appears to be a consequence of the fact that in this case all the estimators in the range $\alpha \in [0.3, 1]$ are practically identical, and any choice of $\alpha$ in this range works equally well. In Figure 2, we plot the estimated OLS line, the estimated LMS line and the MDPDE estimator line with the optimal tuning parameter corresponding to the pilot estimator given by the MDPDE at $\alpha = 0.5$; from Table 4 it may be seen that this optimal parameter is $\alpha = 0.63$.

**Example 3 (Salinity data)** Finally as an example of the multiple regression model with masking effect, we consider the ‘Salinity data’ [3,11]. The set contains measurements of water’s salt concentration and river discharge taken in North Carolina’s Pamlico Sound. Rousseeuw and Leroy [11] fitted a multiple linear regression to these data with salinity as the dependent variable;
Figure 2. Different regression fits for the telephone data.

Table 5. The parameter estimates for the salinity data using the minimum DPD approach for various values of $\alpha$ and the LMS approach.

| $\alpha$ | 0.0 | 0.05 | 0.1 | 0.3 | 0.4 | 0.5 | 0.6 | 0.8 | 1.0 | LMS |
|----------|-----|------|-----|-----|-----|-----|-----|-----|-----|-----|
| $\beta_0$ | 9.59 | 9.96 | 10.51 | 18.28 | 18.37 | 18.40 | 18.46 | 18.89 | 19.19 | 36.70 |
| $\beta_1$ | 0.78 | 0.78 | 0.77 | 0.72 | 0.72 | 0.72 | 0.72 | 0.71 | 0.71 | 0.36 |
| $\beta_2$ | -0.03 | -0.03 | -0.04 | -0.19 | -0.19 | -0.20 | -0.20 | -0.19 | -0.18 | -0.07 |
| $\beta_3$ | -0.30 | -0.31 | -0.33 | -0.62 | -0.63 | -0.63 | -0.63 | -0.65 | -0.66 | -1.30 |
| $\sigma$ | 1.23 | 1.23 | 1.22 | 0.93 | 0.91 | 0.87 | 0.83 | 0.76 | 0.71 | — |

the independent variables are 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, cases 5 and 16 in the data correspond to periods of very heavy discharge but the cases 3 and 16 conspire to hide the discrepant number 5 producing a masking effect in the data.

Table 5 presents the estimates of the regression coefficients obtained by the minimum DPD method for several $\alpha$.

It was observed by Rousseeuw and Leroy [11] that the OLS/ML estimators (corresponding to $\alpha = 0$) cannot separate the influential 5th and 16th observations due to the masking effect. But the robust LMS estimator is not affected by this masking and easily distinguishes these two observations by their large residuals. Table 5 shows that the MDPDEs with $\alpha \geq 0.3$ remain pretty stable. The OLS fit after discarding the influential 5th and 16th observations generates $(23.3862, 0.6998, -0.25, -0.8356)$ as the vector of regression coefficients, quite close to the robust MDPDEs obtained here.

The tuning parameter selection criterion, when applied to these data, lead to an optimal value of $\alpha = 1$ except for pilot estimators corresponding to very small (close to zero) values of $\alpha$ (Table 6). Clearly, the heavy contamination present here requires that a very strong downweighting of the bias component is necessary.
Table 6. The optimum $\alpha$ for the salinity data with several ‘pilot’ $\alpha$ values.

| Pilot $\alpha$ | 0.00 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 1.00 | No bias |
|----------------|------|------|------|------|------|------|------|------|------|------|------|---------|
| Optimum $\alpha$ | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 |

3.2 Simulation study

In this section, we perform a detailed numerical exploration of the performance of the MDPDEs in the linear regression context. This exploration has two components. First, we study the general stability of the estimators under contamination. The second part involves a (primarily graphical) study of the effect of the pilot estimator on the selection of the optimal $\alpha$.

3.2.1 Simulation scheme

We consider the simple linear regression case under the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where $\epsilon_i$s are i.i.d. normal with mean 0 and variance $\sigma^2$. Here, we construct the empirical bias and MSE of the MDPDEs of $\beta_0$, $\beta_1$ and $\sigma$ for various values of the tuning parameter $\alpha$ ($\alpha = 0$ gives the MLE/OLS). We simulate, on purpose, $\epsilon_i$s and $x_i$s from distributions that create specific scenarios of large residuals and leverage points as follows:

1. We choose $x$ randomly from the $N(10, 5)$ distribution, where the latter argument is the variance component; the error $\epsilon$ is chosen randomly from a $N(0, 3)$ distribution. This gives the case of a fully pure model which has no artificially introduced large residuals or leverage points.
2. Next we again choose $x$ randomly from the $N(10, 5)$ distribution, and $\epsilon$ is now chosen randomly from the contaminated $(1 - e_{err})N(0, 3) + e_{err}N(10, 3)$ distribution. In this case, approximately $100e_{err}^\%$ of the residuals are large, but there are no artificial leverage points.
3. Next we choose $x$ randomly from the $(1 - e_s)N(10, 5) + e_sN(16, 5)$ distribution, while $\epsilon$ is selected randomly from the $N(0, 3)$ distribution. About $100e_s^\%$ of the observations now represent leverage points, but there are no large artificial residuals.
4. Finally, we choose $x$ randomly from the $(1 - e_s)N(10, 5) + e_sN(16, 5)$ distribution and $\epsilon$ randomly from the $(1 - e_{err})N(0, 3) + e_{err}N(10, 3)$ distribution. In this case, approximately $100e_{err}^\%$ of the residuals are large and about $100e_s^\%$ of the observations represent leverage points.

For each simulation experiment, we chose the true values of the parameters to be $(\beta_0, \beta_1, \sigma) = (3, 2, \sqrt{3})$, and the bias and MSES are calculated around these target values.

3.2.2 Bias and MSE of MDPDE

We have used 1000 replications to estimate the empirical bias and MSE of the regression parameters. The results, for several combinations of $\alpha$, $e_{err}$ and $e_s$, are given in Tables 7–9. In Table 7, it may be clearly seen that while the estimators corresponding to small values of $\alpha$ can have large biases depending on the combination of $e_{err}$ and $e_s$ values, the large $\alpha$ MDPDEs clearly control this bias neatly and give quite satisfactory results. The same is seen in the case of the MSE in Table 8.

In Table 9, the overall MSE of $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma})^T$ is presented for several values of the tuning parameter. Here, the MSE is the sum of the MSES of the three individual estimators, which is nothing but
the squared norm of the vector bias plus the trace of the empirical variance–covariance matrix. Once again the stable behavior of the estimators corresponding to large values of $\alpha$ is clear.

3.2.3 Choice of optimum $\alpha$

Table 10 represents the estimated values of the optimum MSE and its standard error for different values of the pilot estimators. The results are based on 1000 replications. The entries

Table 7. Simulated bias of $\hat{\beta}_1$ in simple linear regression (sample size = 50).

| $\varepsilon_{err}$ | $\varepsilon_x$ | $\alpha$            |
|---------------------|-----------------|---------------------|
|                     |                 | 0 (OLS) 0.1 0.25 0.4 0.5 0.6 0.75 1 |
| 0                   |                 | 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 |
| 0.05                |                 | 0.01 0.00 0.00 0.00 0.00 0.00 0.00 0.00 |
| 0.1                 | 0               | 0.00 0.01 0.01 0.01 0.01 0.01 0.01 0.01 |
| 0.15                | 0               | 0.01 0.00 0.00 0.00 0.00 0.00 0.00 0.00 |
| 0.2                 | 0               | 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 |
| 0                   | 0.05            | -0.13 -0.09 -0.06 -0.04 -0.03 -0.03 -0.03 -0.03 |
| 0                   | 0.1             | -0.66 -0.50 -0.16 -0.08 -0.07 -0.06 -0.05 -0.04 |
| 0                   | 0.15            | -0.96 -0.85 -0.17 -0.06 -0.05 -0.04 -0.03 -0.02 |
| 0                   | 0.2             | -1.21 -1.20 -1.04 -0.41 -0.20 -0.11 -0.05 -0.02 |
| 0.05                | 0.05            | 0.01 0.00 0.00 0.00 0.00 0.00 0.00 0.00 |
| 0.05                | 0.1             | -0.38 -0.31 -0.13 -0.04 -0.03 -0.02 -0.02 -0.01 |
| 0.05                | 0.15            | -0.70 -0.55 -0.31 -0.17 -0.13 -0.10 -0.08 -0.06 |
| 0.05                | 0.2             | -1.16 -1.10 -0.91 -0.36 -0.19 -0.11 -0.07 -0.04 |
| 0.1                 | 0.05            | -0.38 -0.31 -0.13 -0.04 -0.03 -0.02 -0.02 -0.01 |
| 0.1                 | 0.1             | 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 |
| 0.1                 | 0.15            | -0.70 -0.55 -0.31 -0.17 -0.13 -0.10 -0.08 -0.06 |
| 0.1                 | 0.2             | -1.16 -1.10 -0.91 -0.36 -0.19 -0.11 -0.07 -0.04 |
| 0.15                | 0.05            | -0.44 -0.26 -0.07 -0.02 -0.01 -0.00 -0.00 -0.00 |
| 0.15                | 0.1             | -1.38 -1.32 -0.84 -0.18 -0.07 -0.04 -0.02 -0.02 |

Table 8. Simulated MSE of $\hat{\beta}_1$ in simple linear regression (sample size = 50).

| $\varepsilon_{err}$ | $\varepsilon_x$ | $\alpha$            |
|---------------------|-----------------|---------------------|
|                     |                 | 0 (OLS) 0.1 0.25 0.4 0.5 0.6 0.75 1 |
| 0                   |                 | 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 |
| 0.05                |                 | 0.02 0.02 0.01 0.01 0.01 0.01 0.01 0.01 |
| 0.1                 | 0               | 0.07 0.06 0.03 0.02 0.02 0.02 0.02 0.02 |
| 0.15                | 0               | 0.10 0.09 0.06 0.04 0.03 0.03 0.03 0.03 |
| 0.2                 | 0               | 0.07 0.06 0.05 0.04 0.04 0.03 0.02 0.02 |
| 0                   | 0.05            | 0.03 0.02 0.01 0.01 0.02 0.02 0.02 0.02 |
| 0                   | 0.1             | 0.44 0.26 0.05 0.03 0.02 0.02 0.02 0.02 |
| 0                   | 0.15            | 0.92 0.74 0.08 0.02 0.02 0.02 0.02 0.02 |
| 0                   | 0.2             | 1.46 1.44 1.25 0.49 0.24 0.13 0.07 0.04 |
| 0.05                | 0.05            | 0.55 0.12 0.02 0.02 0.02 0.02 0.02 0.02 |
| 0.05                | 0.1             | 0.64 0.47 0.05 0.02 0.02 0.01 0.01 0.02 |
| 0.05                | 0.15            | 0.51 0.33 0.12 0.06 0.04 0.04 0.03 0.03 |
| 0.05                | 0.2             | 1.36 1.23 0.94 0.34 0.17 0.10 0.06 0.04 |
| 0.1                 | 0.05            | 0.19 0.14 0.05 0.02 0.02 0.02 0.02 0.02 |
| 0.1                 | 0.1             | 1.56 1.44 0.60 0.10 0.04 0.04 0.04 0.04 |
| 0.1                 | 0.15            | 2.09 1.99 0.90 0.14 0.06 0.03 0.03 0.03 |
| 0.1                 | 0.2             | 2.18 2.14 1.88 0.88 0.50 0.29 0.14 0.08 |
| 0.15                | 0.05            | 0.24 0.12 0.04 0.02 0.02 0.02 0.02 0.02 |
| 0.15                | 0.1             | 1.94 1.81 1.08 0.23 0.09 0.05 0.04 0.04 |
Table 9. Simulated MSE of the parameter vector \((\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma})^T\) in simple linear regression (sample size = 50).

| \(e_{\text{err}}\) | \(e_{x}\) | 0 (OLS) | 0.1 | 0.25 | 0.4 | 0.5 | 0.6 | 0.75 | 1 |
|-----------------|----------|---------|------|------|-----|-----|-----|-----|----|
| 0               | 0        | 1.59    | 1.62 | 1.71 | 1.86| 1.97| 2.09| 2.29| 1.85|
| 0.05            | 0        | 3.94    | 2.31 | 1.28 | 1.26| 1.32| 1.39| 1.51| 1.63|
| 0.1             | 0        | 11.05   | 7.84 | 3.55 | 2.50| 2.49| 2.47| 2.56| 2.64|
| 0.15            | 0        | 17.15   | 14.40| 8.79 | 4.34| 3.29| 2.82| 2.86| 3.00|
| 0.2             | 0        | 16.55   | 15.19| 11.75| 6.81| 4.51| 3.20| 2.42| 2.11|
| 0               | 0.05     | 2.54    | 1.88 | 1.60 | 1.62| 1.68| 1.76| 1.94| 2.25|
| 0               | 0.1      | 37.26   | 22.17| 4.37 | 2.55| 2.36| 2.33| 2.42| 2.65|
| 0               | 0.15     | 83.21   | 67.16| 7.08 | 2.26| 2.11| 2.11| 2.19| 2.40|
| 0               | 0.2      | 134.38  | 133.41| 116.13| 46.24| 22.97| 12.62| 7.02| 4.21|
| 0.05            | 0.05     | 61.77   | 15.97| 2.38 | 2.17| 2.25| 2.34| 2.50| 2.79|
| 0.05            | 0.1      | 66.58   | 48.45| 5.80 | 1.69| 1.57| 1.59| 1.68| 1.84|
| 0.05            | 0.15     | 51.26   | 33.02| 11.80| 5.01| 3.75| 3.24| 2.89| 2.72|
| 0.05            | 0.2      | 137.49  | 122.29| 90.79| 32.60| 16.08| 9.18 | 5.47| 4.03|
| 0.1             | 0.05     | 29.40   | 20.67| 7.63 | 2.69| 2.15| 2.11| 2.18| 2.41|
| 0.1             | 0.1      | 183.04  | 166.61| 70.25| 11.34| 5.02| 4.21| 3.89| 4.10|
| 0.1             | 0.15     | 241.31  | 227.94| 105.08| 16.66| 6.84| 3.79| 3.27| 3.20|
| 0.1             | 0.2      | 226.48  | 217.87| 187.77| 87.64| 49.56| 28.71| 13.94| 7.80|
| 0.15            | 0.05     | 41.91   | 24.83| 9.36 | 3.29| 2.32| 1.93| 1.77| 1.94|
| 0.15            | 0.1      | 237.78  | 219.51| 132.81| 29.23| 11.40| 6.59 | 4.65| 4.20|

corresponding to \(\theta(\alpha = 0)\) are the average, optimal mean square error obtained through the appropriate approximation of Equation (13) as discussed in Section 2.3, when the MDPDE with \(\alpha = 0\) is the pilot estimator. The entries for the other columns of \(\alpha\) are similarly interpreted.

It is clear from the table that the optimum choice of \(\alpha\) obtained by minimizing only the variance component (no bias) is the best in terms of the MSE of corresponding MDPDE both in the presence and absence of outliers and (or) leverage points. However, in general the bias term must be incorporated, and our suggestion, based on the tables presented here and other simulations, is to use the MDPDE corresponding to \(\alpha = 0.5\) as the pilot estimator while computing the bias component. We acknowledge, however, that several values of \(\alpha\) in a two-sided neighborhood of \(\alpha = 0.5\) worked almost equally well in our numerical studies.

Next, we provide an extensive graphical investigation of the role of the pilot estimator in determining the optimal choice of \(\alpha\). For this purpose, we have chosen different contaminations and leverage point scenarios. Based on the finding of Tables 7–9, we have some idea of where the optimal tuning parameter should be for each perturbation scenario. We want to explore to what extent the selection of our optimal tuning parameter is consistent with the findings of the above tables and to what extent the results are influenced by the choice of the pilot estimator.

For each scenario, we have selected a number of possible pilot estimators which include a few MDPDEs and also the iteratively reweighted least-squares (IRLS) estimators of the regression parameters based on Huber’s \(\psi\) function; for the last mentioned estimator, the input residuals \(r_i\) at any stage are given as

\[
 r_i = \frac{r_{0i}}{(1.345 \times \hat{\sigma} \sqrt{1 - h_i})},
\]

where \(r_{0i}\) are the residuals at the previous iteration, \(\hat{\sigma} = \text{MAD}/0.6745\), \text{MAD} is the median absolute deviation about the median, and \(h_i\) is the leverage value from the least-squares fit. We will refer to this as the IRLS-Huber estimator. We have not used the LMS estimator as the pilot because of its low efficiency. For each scenario, and each pilot estimator, 100 replications are run.
generating 100 ‘optimal’ tuning parameters. We have then made a histogram of these 100 values to get an idea of whether the findings are in the acceptable range. Our findings can be briefly summarized as follows:

(1) When the data are generated from the pure model, an overwhelming majority of the estimated tuning parameters are equal to or around $\alpha = 0$, irrespective of the pilot estimator (Figure 3). This is how it should be, and indicates that in such cases the algorithm is choosing the correct tuning parameter most of the time. Interestingly, for each pilot estimator, there is always a small chunk of values at or around $\alpha = 1$. The latter phenomenon appears to be always true, except for the pilot estimator corresponding to MDPDE($\alpha = 0$) and the IRLS-Huber estimator.

(2) This situation changes progressively as the level of perturbation increases. However, the MDPDE at $\alpha = 0$ (the MLE) is clearly inadequate, since it appears to choose $\alpha = 0$ itself as the optimal value of the tuning parameter in a large number of cases even when the error distribution is fairly heavily contaminated, or when there are a large number of leverage points. This is not entirely unexpected, as the elimination of a large bias component allows it to stay at the top in such cases.
Figure 3. Histogram of optimal $\alpha$ values obtained using different pilot estimators based on 100 replications (sample size = 50) for $\varepsilon_{\text{err}} = 0$ and $e_x = 0$.

(3) As the perturbation becomes heavier, the large bar around $\alpha = 0$ melts away (see Figures 1–9 in the supplementary material). The estimated tuning parameter shifts toward larger values along the $\alpha$ axis. For small perturbations, there is a significant concentration of optimal values slightly above, but not equal to, the value $\alpha = 0$ (see Figure 1 of the supplementary material), as would seem appropriate.

(4) The degree of shift of the optimal values to the right along the $\alpha$ axis appears to be consistent with the degree of perturbation.

(5) The IRLS (Huber) pilot estimators do not appear to generate useful optimal values all the time. It would appear that the MDPDEs corresponding to $\alpha = 0.3, 0.5$ and 1 are the most useful ones in the present context. However, apart from these empirical suggestions, stronger theoretical recommendations may be of some practical interest. In the future, we hope to develop estimates of the bias component without the use of a 'pilot' estimator.

(6) On the whole, there is very little difference in the overall shape of the histograms between the pilot estimators with $\alpha = 0.5$ and $\alpha = 1$.

4. Concluding remarks

In this paper, we have tried to explore, in detail, the applicability of the minimum DPD estimation method in the case of non-homogeneous data. The robustness of these methods is generated in a natural way as a function of a single tuning parameter, and since this parameter controls the trade-off between robustness and efficiency, a proper selection of the tuning parameter in any given problem becomes important. Together with the study of the robustness of the estimators, we have tried to select the optimal tuning parameter appropriately, following the approach of Warwick and Jones [14], and have carefully studied the role of the pilot estimator. We trust that on the whole this establishes the practical utility of the minimum DPD estimators, and, together with the works of Scott [12] and Durio and Isaia [4], nicely complements the theoretical findings of Ghosh and Basu [5]. We hope that the practitioners will find the method to be of value in solving real-life situations.
Acknowledgements

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

Disclosure statement

No potential conflict of interest was reported by the authors.

Supplementary material

Supplemental data for this article can be accessed at 10.1080/02664763.2015.1016901 description of location.

References

[1] A. Basu, I.R. Harris, N.L. Hjort and M.C. Jones, Robust and efficient estimation by minimizing a density power divergence, Biometrika 85 (1998), pp. 549–559.
[2] A. Basu, H. Shioya, and C. Park, Statistical Inference: The Minimum Distance Approach, Chapman & Hall/CRC, Boca Raton, FL, 2011.
[3] D. Ruppert and R.J. Carroll, Trimmed least squares estimation in the linear model, J. Amer. Statist. Assoc. 75 (1980), pp. 828–838.
[4] A. Durio and E.D. Isaia, The minimum density power divergence approach in building robust regression models, Informatica 22/1 (2011), pp. 43–56.
[5] A. Ghosh and A. Basu, Robust estimation for independent non-homogeneous observations using density power divergence with applications to linear regression, Electron. J. Stat. 7 (2013), pp. 2420–2456.
[6] A. Ghosh, I.R. Harris, A. Maji, A. Basu, and L. Pardo, A generalized divergence for statistical inference, Technical Report, BIRU/2013/3, Bayesian and Interdisciplinary Research Unit, Indian Statistical Institute, Kolkata, India, 2013.
[7] F.R. Hampel, E. Ronchetti, P.J. Rousseeuw, and W. Stahel, Robust Statistics: The Approach Based on Influence Functions, John Wiley & Sons, New York, 1986.
[8] C. Hong and Y. Kim, Automatic selection of the tuning parameter in the minimum density power divergence estimation, J Korean Statist. Soc. 30 (2001), pp. 453–465.
[9] P.J. Huber, Robust estimation of a location parameter, Ann. Math. Stat. 35 (1964), pp. 73–101.
[10] L. Pardo, Statistical Inference Based on Divergences, CRC/Chapman-Hall, Boca Raton, FL, 2006.
[11] P.J. Rousseeuw and A.M. Leroy, Robust Regression and Outlier Detection, John Wiley & Sons, New York, 1987.
[12] D.W. Scott, Parametric statistical modeling by minimum integrated square error, Technometrics 43 (2001), pp. 274–285.
[13] I. Vajda, Theory of Statistical Inference and Information, Kluwer Academic, Dordrecht, 1989.
[14] J. Warwick and M.C. Jones, Choosing a robustness tuning parameter, J. Statist. Comput. Simul. 75 (2005), pp. 581–588.