The Minimum Density Power Divergence Approach in Building Robust Regression Models

The Minimum Density Power Divergence Approach in Building Robust Regression Models

A. Durio[*]and E. D. Isaia [†]

February 22, 2011

Abstract

It is well known that in situations involving the study of large datasets where influential observations or outliers maybe present, regression models based on the Maximum Likelihood criterion are likely to be unstable. In this paper we investigate the use of the Minimum Density Power Divergence criterion as a practical tool for parametric regression models building. More precisely, we suggest a procedure relying on an index of similarity between estimated regression models and on a Monte Carlo Significance test of hypothesis that allows to check the existence of outliers in the data and therefore to choose the best tuning constant for the Minimum Density Power Divergence estimators. Theory is outlined, numerical examples featuring several experimental scenarios are provided and main results of a simulation study aiming to verify the goodness of the procedure are supplied.

Keywords: Minimum density power divergence estimators, Monte Carlo significance test, Outliers detection, Robust regression, Similarity between functions.

1 Introduction

In applied statistics regression is certainly one of the widespread tool in establishing the relationship between a set of predictors and a response variable. However, in many circumstances a careful data preparation is not

[*] Department of Statistics and Applied Mathematics, University of Turin (Italy), durio@econ.unito.it

[†] Department of Statistics and Applied Mathematics, University of Turin (Italy), isaia@econ.unito.it

feasible and data may hence be heavily contaminated by a substantial number of outliers. In these situations, the estimates of the parameters of the regression model according to the Maximum Likelihood criterion are fairly unstable. Since outliers can play havoc with standard statistical methods (Daniel et al. (1968), Rousseeuw and Leroy (1987), Davies (1993)), many robust estimators have been proposed since 1960 to be less sensitive to outliers. The development of robust methods is underlined by the appearance of a wide number of papers and books on the topic including the more recent Huber (1981), Hampel et al. (1987), Staudte and Sheather (1990), Dodge and Jurečkova (2000), Seber and Lee (2003), Rousseeuw et al. (2004) and Maronna et al. (2006). In parametric estimation, the estimators with good robustness proprieties relative to maximum likelihood are those based on a minimum divergence methods. The minimum divergence estimators are M-estimators and their proprieties are strictly linked on the distance used as measure the divergence.

In the following we investigate the use of the Minimum Density Power Divergence criterion as a valuable tool for useful parametric regression models building. Our work can be seen as an attempt to explore the pratical utility of robust estimators based on this minimun distance method that are in literature proposed from a theoretical point of view by Basu et al. (1998). One of the main critical issue in the use of a robust family estimators is the tuning parameters selection. The $MDPD$ estimators family is indexed by a single parameter which controls the trade-off between robustness and asymptotic efficiency of the estimator. In the work of Warwick and Jones (2005) the best value for the parameter of the Basu family estimators is selected minimizing an asymptotic estimation of the mean squared error. In the paper of Fujisawa and Eguchi (2006) is proposed an adaptive methods for selecting the tuning parameter based on an emphirical approximations of the Cramer-von Mises divergence. We propose a data-driven way to choose the tuning parameter based on a Monte Carlo Significance test on the similarity between a robust and a classical estimators. More precisely, we introduce and discuss an intuitive procedure which relies on an index of similarity between estimated regression models and on a Monte Carlo Significance test of statistical hypothesis. The procedure we suggest allows (a) to verify the presence of outliers in the data and, if they are present, (b) to select the best tuning constant for the Minimum Density Power Divergence estimators. We propose a data-driven way to choose the tuning parameter relevant issue are solved in a data-driven way

Theory is outlined and numerical examples featuring several scenarios are provided and for each of them main results of a simulation study, aiming to verify the goodness of the whole procedure, are supplied and commented.

2 The methods and the proposed procedure

In this section we first introduce, for a parametric regression problem, the Minimum Density Power Divergence Estimators ($MDPDE$), originally proposed by Basu, Harris, Hjort, and Jones (1998). The procedure to choose the tuning parameter is illustrated in the secon subsection in which we also describe the similarity index between two estimators and the simplified Monte Carlo Significance Test.

2.1 The regression model and the estimators

Let $\{(x_{i1}, \dots, x_{ip}, y_i)\}_{i=1,\dots,n}$ be the observed dataset, where each observation stems from a random sample drawn from the $p+1$ random variable $(X_1, \dots, X_p, Y)$. The regression model for the observed data set we study is $y_i = m_{\boldsymbol{\beta}}(\mathbf{x}_i) + \varepsilon_i$, with $i = 1, \dots, n$, and the object of our interest is the regression mean

$$m_{\boldsymbol{\beta}}(\mathbf{x}_i) = \mathbb{E}[Y|\mathbf{x}_i] = \beta_0 + \sum_{j=1}^{p} \boldsymbol{\beta}_j \mathbf{x}_{ij},$$

where the errors $\{\varepsilon_i\}_{i=1,\dots,n}$ are assumed to be independent random variables with zero mean and unknown finite variances. If we furthermore assume that the errors are i.i.d. $\mathcal{N}(0, \sigma_0)$, then the estimate of the vector of the parameters according to the Maximum Likelihood (ML) criterion is

$$\hat{\boldsymbol{\beta}}_{ML} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left[\frac{1}{(2\,\pi\,\sigma_0^2)^{n/2}} \exp\left(\frac{\sum_{i=1}^{n} (y_i - m_{\boldsymbol{\beta}}(\mathbf{x}_i))^2}{2\,\sigma_0^2} \right) \right] \tag{2.1}$$

and in this case the solutions of equation (2.1) are equivalent to the ones given by the ordinary least-squares method; as an alternative we consider the family ($MDPDE$).

Given the r.v. $\mathbf{X}$ of dimension $d \geq 1$ with density $\varphi(\mathbf{x}|\boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0 \in \mathcal{S} \subseteq \mathbb{R}^p$ and $p \geq 1$, for which we introduce the model $f(\mathbf{x}|\boldsymbol{\theta})$, with $\boldsymbol{\theta} \in \mathcal{S}$, the density power divergence between f and φ is defined, for $\alpha > 0$, as

$$d_\alpha(f, \varphi) = \int_{\mathbb{R}^d} \left\{ f^{1+\alpha}(\mathbf{x}|\boldsymbol{\theta}) - \left(1 + \frac{1}{\alpha}\right) \varphi(\mathbf{x}|\boldsymbol{\theta}_0) f^\alpha(\mathbf{x}|\boldsymbol{\theta}) + \frac{1}{\alpha} \varphi^{1+\alpha}(\mathbf{x}|\boldsymbol{\theta}_0) \right\} d\mathbf{x},$$

while for $\alpha = 0$ it is defined as the Kullback-Leibler divergence.

Let $\mathbf{X}_1, \dots, \mathbf{X}_n$ be a random sample of size $n \geq 2$ from $\mathbf{X}$, the Minimum Density Power Divergence Estimator for $\boldsymbol{\theta}_0$ corresponds to the vector $\hat{\boldsymbol{\theta}}_\alpha$ minimizing the divergence $d_\alpha(f, \varphi)$ between the probability mass function

$\hat{\varphi}_n$ associated with the empirical distribution of the sample and f, that is for $\alpha > 0$

$$\hat{\boldsymbol{\theta}}_\alpha = \underset{\boldsymbol{\theta}\in\mathcal{S}}{\operatorname{argmin}} \left[\int_{\mathbb{R}^d} f^{1+\alpha}(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} - \left(1 + \frac{1}{\alpha}\right) \frac{1}{n} \sum_{i=1}^{n} f^\alpha(\mathbf{X}_i|\boldsymbol{\theta}) \right]. \tag{2.2}$$

In general, it can be shown that as the tuning parameter α increases the robustness of the Minimum Density Power Divergence estimator increases while its efficiency decreases (Basu et al. (1998)). For $\alpha = 0$ the $MDPDE$ becomes the Maximum Likelihood estimator, while for $\alpha = 1$ the divergence $d_1(f, \varphi)$ yields the L_2 metric and the estimator minimizes the L_2 distance between the densities (e.g Scott, 2001, Durio and Isaia, 2003).

The Minimum Density Power Divergence criterion can easily be applied to parametric regression problems. In fact if we assume that the random variables $Y|\mathbf{x}$ are distributed as a $\mathcal{N}(m_{\boldsymbol{\beta}}(\mathbf{x}), \sigma_0)$ random variable with density function ϕ, then, according to equation (2.2), the estimate of the vector $\boldsymbol{\theta}_\alpha = [\beta_0, \ldots, \beta_p, \sigma_0]$, is given by

$$\hat{\boldsymbol{\theta}}_\alpha = \underset{\boldsymbol{\beta},\sigma}{\operatorname{argmin}} \left[\frac{1}{\sigma^\alpha \sqrt{(2\pi)^\alpha (1+\alpha)}} - \frac{\alpha+1}{\alpha} \frac{1}{n} \sum_{i=1}^{n} \phi^\alpha(y_i|m_{\boldsymbol{\beta}}(\mathbf{x}_i), \sigma) \right], \tag{2.3}$$

as the integral of equation (2.2) becomes

$$\int_{\mathbb{R}} \phi^{1+\alpha}(y|m_{\boldsymbol{\beta}}(\mathbf{x}), \sigma)\, dy = \frac{1}{\sqrt{1+\alpha}}\, \phi^\alpha(0|m_{\boldsymbol{\beta}}(\mathbf{x}), \sigma).$$

The vector $\hat{\boldsymbol{\theta}}_\alpha$ obtained by equation (2.3) contains the estimates of the $p + 1$ parameters of the model and the estimate of the standard deviation of the errors, i.e. $\hat{\boldsymbol{\theta}}_\alpha = \left[\hat{\boldsymbol{\beta}}_{MD,\alpha}, \hat{\sigma}_{MD,\alpha}\right]$. In the following we unambiguously indicate with $\hat{\boldsymbol{\beta}}_{MD,\alpha}$ the estimate of the vector $\boldsymbol{\beta}$ in accord with the Minimum Density Power Divergence criterion and therefore we denote by $\hat{m}_{MD,\alpha}(\mathbf{x})$ the corresponding estimated regression model.

We remark that equation (2.3) is a feasible computationally closed-form expression so that $MDPD$ criteria can be performed by any standard non linear optimization code, for instance the `nlm` routine of the `R` library, although, whatever the algorithm, its convergence to optimal solutions strongly depends on its initial configurations.

2.2 The choice of the best $\boldsymbol{\alpha}$ tuning parameter

As stated above, our purpose is to check the presence of outliers in the data set and, if they are present, to choose in the family of the Minimum Density

Power Divergence estimators the best one, that is to select the tuning α parameter such that we obtain concurrently the most robust and the most efficient estimator.

Since we already pointed out that the robustness of the Minimum Density Power Divergence estimator increases as α increases, when outliers are present the vectors of the estimates $\hat{\boldsymbol{\beta}}_{ML}$, and $\hat{\boldsymbol{\beta}}_{MD,\alpha}$, for some $0 < \alpha \leq 1$, will be not equal and hence the estimated regression models tend to be dissimilar.

In order to compare the performance of $MDPDE$ with respect to MLE or, more generally, the performance between any two estimators in the family of the Minimum Density Power Divergence Estimators, we resort to the normalized index of similarity between regression models originally proposed by Durio and Isaia (2010).

Letting T_0 and T_1 be two regression estimators and $\hat{\boldsymbol{\beta}}_{T_0}$, $\hat{\boldsymbol{\beta}}_{T_1}$ the corresponding vectors of the estimated parameters, the similarity index takes into account the space region between $\hat{m}_{T_0}(\mathbf{x})$ and $\hat{m}_{T_1}(\mathbf{x})$ with respect to the space region where data points locate. If we introduce the sets

$$\begin{aligned}
\boldsymbol{I}^p &= [\min(x_{i1}); \max(x_{i1})] \times \ldots \times [\min(x_{ip}); \max(x_{ip})] \\
\boldsymbol{I} &= [\min(y_i); \max(y_i)]
\end{aligned}$$

the similarity index is defined as

$$\begin{aligned}
& sim(T_0, T_1) \overset{def}{=} \frac{\int_{\boldsymbol{D}^{p+1}} d\mathbf{t}}{\int_{\boldsymbol{C}^{p+1}} d\mathbf{t}} \\
& \boldsymbol{C}^{p+1} = \boldsymbol{I}^p \times \boldsymbol{I} \\
& \boldsymbol{D}^{p+1} = \left\{(\mathbf{x}, y) \in \mathbb{R}^{p+1} : \zeta(\mathbf{x}) \leq y \leq \xi(\mathbf{x}), \mathbf{x} \in I^p\right\} \cap \boldsymbol{C}^{p+1},
\end{aligned} \tag{2.4}$$

where $\zeta(\mathbf{x}) = min\left(\hat{m}_{T_0}(\mathbf{x}), \hat{m}_{T_1}(\mathbf{x})\right)$, $\xi(\mathbf{x}) = max\left(\hat{m}_{T_0}(\mathbf{x}), \hat{m}_{T_1}(\mathbf{x})\right)$ and clearly $0 \leq sim(T_0, T_1) \leq 1$.

If the vectors $\hat{\boldsymbol{\beta}}_{T_0}$ and $\hat{\boldsymbol{\beta}}_{T_1}$ are close to each other, then $sim(T_0, T_1)$ will be close to zero On the other hand, if the estimated models $\hat{m}_{T_0}(\mathbf{x})$ and $\hat{m}_{T_1}(\mathbf{x})$ are dissimilar we are likely to observe a value of $sim(T_0, T_1)$ tending to unit.

We therefore suggest to use the $sim(T_0, T_1)$ statistic given by equation (2.4) to verify the following system of hypothesis

$$\begin{cases} H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \\ H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0 \end{cases} \tag{2.5}$$

where $\boldsymbol{\beta}_0$ is the value of $\hat{\boldsymbol{\beta}}_{T_0}$ computed on the observed data.

Since it is not realistic to look for a closed-form of the $sim(T_0, T_1)$ distribution, in order to check the above system of hypothesis we resort to the simplified Monte Carlo Significance Test (Barnard, 1963, Hope, 1968).

Denoting with $sim_{T_0T_1}$ the value of the $sim(T_0, T_1)$ statistic computed on the observed data, the simplified Monte Carlo Significance test (MCS test) consists in rejecting the null hypothesis of system (2.5) if $sim_{T_0T_1}$ is the $m\,\alpha$-th most extreme statistic relative to the corresponding quantities $sim^*_{T_0T_1}$ computed on each of the random samples of the reference set. The reference set consists in $m-1$ random samples of size n each generated under the null hypothesis, that is drawn at random from the regression model $\hat{m}_{T_0}(\mathbf{x})$ with $\sigma = \hat{\sigma}_{T_0}$. We remark that if we set the type-I error probability equal to $0.001(0.002, 0.01, 0.05)$ then the size of the reference set will be $999(499, 99, 19)$.

In order to meet our target, we introduce a procedure consisting in three steps. For a given dataset, we start verifying the presence of outliers checking, with the aid of the MCS test, the similarity between MLE and the less efficient but more robust $MDPDE$ with $\alpha = 1$.

If outliers are present, i.e. if the MCS test leads us to reject the null hypothesis of system (2.5), we look for the best $MDPD$ estimator checking the similarity between $MDPDE$ with $\alpha = 1$ and $MDPDE$ with $\alpha < 1$, increasing α until the MCS test allows us to accept for the first time the null hypothesis of system (2.5).

The whole 3 Steps procedure can be summarized as follows

Step 1: considering the Maximum Likelihood estimator and the Minimum Density Power Divergence estimator with $\alpha = 1$, i.e. we set $T_0 = ML$ and $T_1 = MD_{\alpha=1}$, we check for outliers testing $sim(ML, MD_{\alpha=1})$.

Step 2: if the MCS test of Step 1 leads us to accept H_0, then we can state that outliers are absent and the best model is the one corresponding, for its inherent properties, to the Maximum Likelihood criterion.

Step 3: if from Step 1 we reject H_0, in order to choose the best Minimum Density Power Divergence estimator we check the similarity between the regression models estimated by $MD_{\alpha=1}$ and $MD_{\alpha<1}$. We perform the MCS test increasing α until for the first time it allows us to accept the null hypothesis. The corresponding value of the tuning parameter $\alpha = \alpha^\star$ gives the best Minimum Density Power Divergence estimator for the given dataset.

3 Numerical examples and simulation

In this section we provide and comment some numerical examples featuring several experimental situations in order to show how the 3 Steps procedure we propose works in practice.

Furthermore, with the aim to verify the goodness of the whole procedure, we introduce and comment a simulation study that we apply to each experimental scenario.

3.1 Numerical examples

The first example considers a situation where no outliers are present, while the next two scenarios involve a substantial number of outliers (20%) affecting the data. A last numerical example investigates the behaviour of the procedure and consequentially the performance the Minimum Density Power Divergence estimators when the number of outliers increases (from 2.4% up to 20%).

Example I: as a first example, we consider a simulated dataset of $n = 600$ points generated according to the model

$$Y = 0.5\, X_1 + 0.5\, X_2 + \varepsilon \tag{3.1}$$

where $X_1, X_2 \sim \mathcal{U}(0, 1)$ and $\varepsilon \sim \mathcal{N}(0, 0.1)$.
Since in this situation no outliers are present, we expect that the regression models estimated according to the Maximum Likelihood and to the Minimum Density Power Divergence criteria can be considered similar and this for any $0 < \alpha \leq 1$. This example is provided to show that the procedure do not falls in select a less efficient $MPDP$ estimators when its robust proprieties are not necessary.
We start considering the ML estimator and the $MDPD$ estimator with $\alpha = 1$ (i.e. we set $T_0 = ML$ and $T_1 = MD_{\alpha=1}$) and we obtain the following estimates $\hat{\boldsymbol{\beta}}_{ML} = [-0.0053, 0.5095, 0.5073]$ with $\hat{\sigma}_{ML} = 0.0927$ and $\hat{\boldsymbol{\beta}}_{MD,\alpha=1} = [-0.0086, 0.5106, 0.5152]$ with $\hat{\sigma}_{MD,\alpha=1} = 0.0974$, while according to equation (2.4) the similarity index is $sim_{ML,MD_{\alpha=1}} = 0.00176$. If we apply the MCS test (level of significance 99.8%), it leads us to accept H_0, as we obtain $\max(sim^*_{ML,MD_{\alpha=1}}) = 0.00651 > sim_{ML,MD_{\alpha=1}} = 0.00176$.
Clearly for this scenario, where outliers are absent and the two estimated models $\hat{m}_{ML}(\mathbf{x})$ and $\hat{m}_{MD,\alpha=1}(\mathbf{x})$ can be considered similar, we state that the best model is the one given, for its inherent properties, by the Maximum Likelihood criterion, i.e. $\hat{m}_{ML}(\mathbf{x}) = -0.0053 + 0.5095\, x_1 + 0.5073\, x_2$.

Table 1: main results of the MCS test of Step 3 applied to $sim(MD_{\alpha=1}, MD_{\alpha<1})$ for the simulated dataset of Example I where outliers are absent.

α	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\sigma}$	$sim_{MD_{\alpha=1},MD_{\alpha<1}}$	H_0
0.10	-0.0061	0.5094	0.5093	0.0930	0.00139	Acc.
0.30	-0.0071	0.5094	0.5118	0.0941	0.00076	Acc.
0.60	-0.0079	0.5097	0.5138	0.0960	0.00027	Acc.
0.90	-0.0084	0.5103	0.5149	0.0973	0.00041	Acc.

Even if it is not necessary, but just for spirit of inquiry, we perform Step 3 of our procedure. This means setting $T_0 = MD_{\alpha=1}$ and $T_1 = MD_{\alpha<1}$ and repeatedly applying the MCS test. The results of Table 1 show that the pairs of $MDPD$ estimators can be considered similar for any value of $\alpha < 1$ and this confirms the goodness of the strategy we suggest even in simple case where outliers are absent.

Example II: we consider now a variant of the situation of Example I. This new scenario involves a simulated dataset of $n_1 = 480$ points generated according to the model

$$Y = 0.5\, X_1 + 0.5\, X_2 + \varepsilon \tag{3.2}$$

and $n_2 = 120$ points, that we consider as outliers, drawn from the model

$$Y = 0.7\, X_1 + 0.7\, X_2 + \varepsilon \tag{3.3}$$

where $X_1, X_2 \sim \mathcal{U}(0, 1)$ and $\varepsilon \sim \mathcal{N}(0, 0.1)$.
In this case too we start estimating the parameters of the regression model resorting to MLE and $MDPDE$ with $\alpha = 1$ and we obtain the following vectors of the estimates $\hat{\boldsymbol{\beta}}_{ML} = [0.0074, 0.5817, 0.5714]$ with $\hat{\sigma}_{ML} = 0.1876$ and $\hat{\boldsymbol{\beta}}_{MD,\alpha=1} = [0.0126, 0.4922, 0.5035]$ with $\hat{\sigma}_{MD,\alpha=1} = 0.1130$. If we compute the similarity index between the two estimated regression models we have $sim_{ML,MD_{\alpha=1}} = 0.04226$ and the MCS test (level of significance 99.8%) leads us to reject the null hypothesis as $\max(sim^*_{ML,MD_{\alpha=1}}) = 0.00868$.
The two estimated regression models can thus be considered dissimilar. This result is quite obvious since we are in presence of a substantial cluster of outliers (20%) and the estimator based upon the L_2 norm tends to estimate the heaviest cluster of data (Durio and Isaia, 2004).
In order to look for the the best estimator in the family of the Minimum Density Power Divergence estimators, we move to Step 3 of our procedure,

Table 2: main results of the MCS test of Step 3 applied to $sim(MD_{\alpha=1}, MD_{\alpha<1})$ for the simulated dataset of example II where outliers are present.

α	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\sigma}$	$sim_{MD_{\alpha=1},MD_{\alpha<1}}$	H_0
0.10	0.0185	0.5521	0.5491	0.1810	0.03413	Rej.
0.15	0.0185	0.5521	0.5491	0.1810	0.03215	Rej.
0.20	0.0227	0.5407	0.5409	0.1770	0.03093	Rej.
0.25	0.0269	0.5287	0.5323	0.1718	0.02711	Rej.
...	...	...	...	...	...	...
0.65	0.0205	0.4880	0.5028	0.1180	0.00321	Rej.
0.70	0.0186	0.4889	0.5030	0.1162	0.00233	Rej.
0.75	*0.0170*	*0.4897*	*0.5032*	*0.1150*	*0.00146*	*Acc.*
0.80	0.0157	0.4904	0.5033	0.1142	0.00087	Acc.
0.85	0.0146	0.4910	0.5034	0.1136	0.00058	Acc.
0.90	0.0138	0.4915	0.5035	0.1133	0.00022	Acc.

i.e. we set $T_0 = MD_{\alpha=1}$ and $T_1 = MD_{\alpha<1}$ and we repeatedly apply the MCS test. The results displayed in Table 2 show that the best value of the tuning parameter corresponds to $\alpha^\star = 0.75$, while for any $\alpha > 0.75$ we always accept H_0. It follows that the optimal estimate of the regression model is $\hat{m}_{MD,\alpha=0.75}(\mathbf{x}) = 0.0170 + 0.4897\, x_1 + 0.5032\, x_2$ with $\hat{\sigma}_{MD,\alpha=0.75} = 0.1150$.

Example III: as a third example we examine a situation where the response is explained by four predictors. To this end we consider a simulated dataset of $n_1 = 480$ points generated according to the model

$$Y = \sum_{i=1}^{4} 0.25\, X_i + \varepsilon \tag{3.4}$$

and $n_2 = 120$ points, that we consider as outliers, drawn from the model

$$Y = \sum_{i=1}^{4} 0.35\, X_i + \varepsilon \tag{3.5}$$

where $X_i \sim \mathcal{U}(0, 1)$ and $\varepsilon \sim \mathcal{N}(0, 0.1)$.
Considering Step 1 of our procedure we set $T_0 = ML$ and $T_1 = MD_{\alpha=1}$ and from equations (2.1) and (2.3) we obtain the following vectors of the estimate

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{ML} &= [-0.0299, 0.2977, 0.2837, 0.2871, 0.2725] \\ \hat{\boldsymbol{\beta}}_{MD,\alpha=1} &= [0.0097, 0.3060, 0.2616, 0.2070, 0.2483] \end{aligned}$$

with $\hat{\sigma}_{ML} = 0.1351$ and $\hat{\sigma}_{MD,\alpha=1} = 0.1179$.
If we calculate the similarity index between the two estimated models we have $sim_{ML,MD_{\alpha=1}} = 0.01997$ and the MCS test (significance level 99.8%) leads us to reject the null hypothesis, as $\max(sim^*_{ML,MD_{\alpha=1}}) = 0.01059$.
Since the two estimated regression models $\hat{m}_{ML}$ and $\hat{m}_{MD,\alpha=1}$ can be judged dissimilar, we move to Step 3 of the procedure and, setting $T_0 = MD_{\alpha=1}$ and $T_1 = MD_{\alpha<1}$, we repeatedly apply the MCS test. Doing so, we find that the best value of the tuning parameter is $\alpha^\star = 0.65$, while for any $\alpha > 0.65$ we always accept H_0. It follows that the best estimate of the regression model is $\hat{m}_{MD,\alpha=0.65}(\mathbf{x}) = 0.0057 + 0.2982\ x_1 + 0.2662\ x_2 + 0.2233\ x_3 + 0.2499\ x_4$ with $\hat{\sigma}_{MD,\alpha=0.65} = 0.1209$.

Example IV: in this last example we examine the behaviour of the Minimum Density Power Divergence estimators as the number of outliers increases. To this end, we consider a sample of $n_1 + n_2 = 200$ points generated according to the one predictor model

$$Y = X_{n_i} + \varepsilon \tag{3.6}$$

where $X_{n_1=180} \sim \mathcal{U}(0, 0.5)$, $X_{n_2=20} \sim \mathcal{U}(0.5, 1)$ and $\varepsilon \sim \mathcal{N}(0, 0.1)$,
We furthermore we generate $m = 5(10, 20, 30, 40, 50)$ points, that we consider as outliers, from the model

$$Y = 0.5\,X + \varepsilon \tag{3.7}$$

where $X \sim \mathcal{U}(0.7, 1)$ and $\varepsilon \sim \mathcal{N}(0, 0.05)$.
According to this specific layout, Step 1 leads us invariably to reject the null hypothesis of similarity between he estimated models $\hat{m}_{ML}$ and $\hat{m}_{MD,\alpha=1}$, that is the MCS test always detects the presence of outliers infecting the data and this for any value of m.
Moving to Step 3, which implies setting $T_0 = MD_{\alpha=1}$ and $T_1 = MD_{\alpha<1}$ and repeatedly applying the MCS test, we obtain for $m = 5, 10, 20, 30, 40, 50$ the following best values $\alpha^\star = 0.065, 0.200, 0.370, 0.425, 0.515, 0.655$ (see first row of Table 3 and left panel of Figure 2).
These results indicate that the $\alpha^\star$ values increase as the number of outliers grows up. This behaviour is not surprising and it is fully justified if we think that increasing the number of the outliers the optimal tuning parameter tends to unit, which is to say that it tends to the estimator based on the L_2 norm and this for the intrinsic properties of the estimators based on the Minimum Density Power Divergence criterion.

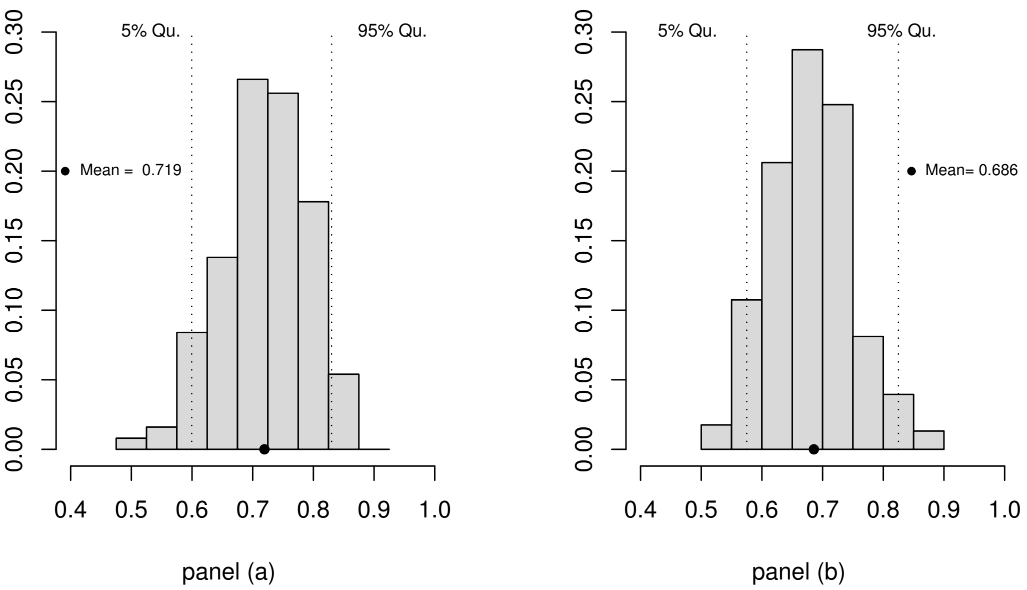

Figure 1: panel(a) histograms of the best values of the tuning constant obtained from simulation for Example II and panel(b) histograms of the best values of the tuning constant for Example III.

3.2 The simulation

We turn now our attention to investigate the goodness of the 3 Steps procedure, that is to verify if the $\alpha^{\star}$ values we obtained on the datasets of Examples I through IV remain somewhat constant if we resample form each generating model. To this end we decide to set up a simulation study where

1. we generate $h = 1000$ samples from the given generating models
2. on each sample we perform Step 1, 2 and 3 of our procedure. With regard to Step 3, we let the tuning constant varying from $\alpha = 0.01$ up to $\alpha = 0.90$ with increments of 0.005.
3. for each sample we record the best value $\alpha^{\star s}$ of the tuning constant.

Example I: for this simple scenario, where the 600 points are generated from model (3.1) and hence no outliers infect the samples, the results of the simulation corroborate the solution we obtained on our original sample. This in the sense that we always accept the hypothesis of similarity between the

Table 3: $\alpha^\star$ values for the sample data of Example IV and summary statistics of the distributions of the $\alpha^{\star s}$ obtained from simulation.

	number of outliers					
	$m = 5$	$m = 10$	$m = 20$	$m = 30$	$m = 40$	$m = 50$
$\alpha^\star$	**0.075**	**0.200**	**0.370**	**0.425**	**0.515**	**0.655**
Min	0.055	0.125	0.300	0.385	0.475	0.600
Mean	0.078	0.194	0.366	0.430	0.512	0.653
St.Dev.	0.011	0.035	0.029	0.019	0.023	0.024
5% Qu.	0.060	0.150	0.325	0.412	0.475	0.612
25% Qu.	0.070	0.175	0.350	0.412	0.500	0.644
50% Qu.	0.075	0.200	0.375	0.425	0.512	0.653
75% Qu.	0.085	0.225	0.400	0.450	0.525	0.675
95% Qu.	0.095	0.250	0.400	0.465	0.550	0.688
Max	0.105	0.275	0.425	0.465	0.563	0.700

regression models estimated by ML and by $MDPD_{\alpha=1}$ and also between $MDPD_{\alpha=1}$ and any $MDPD_{\alpha<1}$.

Example II: in this situation we generate data points from models (3.2) and (3.3) and we always reject the null hypothesis on Step 1, i.e without fail we recognize the presence of outliers.
With regard to Step 3, the results of the simulation are quite encouraging and this in the sense that all the 1000 $\alpha^{\star s}$ values range from 0.520 up to 0.875 with a mean of 0.719 (recall that for our original sample we found $\alpha^\star = 0.75$), a median equal to 0.70 and a standard deviation of 0.072. Furthermore, see Figure 1 panel(a), we can state that 90% of the $\alpha^{\star s}$ fall in the range $[0.60; 0.85]$ while the 50% of them belong to the interval $[0.70; 0.75]$.

Example III: in this context we generate data points according to models (3.4) and (3.5) and we systematically reject the null hypothesis on Step 1, that is we always detect the presence of outliers.
Moving to Step 3, the outcomes of the simulation are promising and this in the sense that all the $\alpha^{\star s}$ obtained from simulation range from 0.510 up to 0.890 with a mean of 0.686 (remember that for our original sample we had $\alpha^\star = 0.65$), a median equal to 0.70 and a standard deviation of 0.063. Furthermore, see Figure 1 panel (a), we can assert that 90% of the $\alpha^{\star s}$ values fall in the range $[0.575; 0.825]$ while the 50% of them belong to the interval $[0.650; 0.725]$.

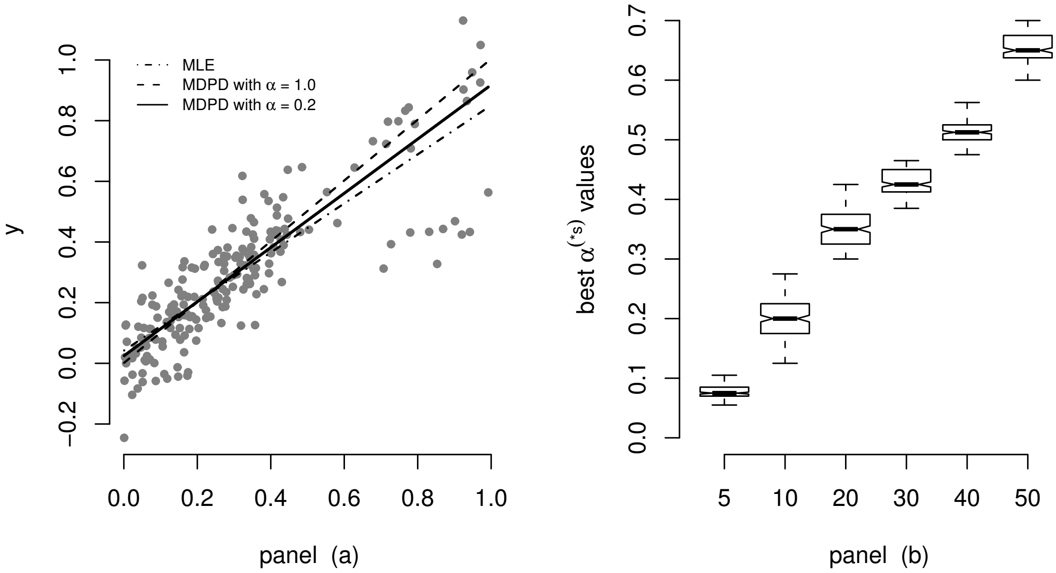

Figure 2: panel (a) displays data points of Example IV ($m = 10$) and the estimated regression models $\hat{m}_{ML}$, $\hat{m}_{MD,\alpha=1}$ and $\hat{m}_{MD,\alpha^{\star}=0.20}$. The panel (b) shows the boxplots of the distributions of the $\alpha^{\star s}$ obtained from simulation for the six different values of m.

Example IV: in this situation, fixed the number m of the outliers, and we generate data points from models (3.6) and (3.7) and thus we consider six sub-scenarios.

Applying Step 1 of our procedure to each sub-scenario, we always reject the hypothesis of similarity between $\hat{m}_{ML}$ and $\hat{m}_{MD,\alpha=1}$. This means that the estimator based on the L_2 norm detects the presence of outiliers even when $m = 5$, although this behaviour is essentially due to the specific expression of the generating model (3.6).

Table 3 shows, for each sub-scenario, some summaries of the distribution of the $\alpha^{\star s}$ values obtained from simulation along with the $\alpha^{\star}$ values computed on the original datasets (first row).

In this case too the results of the simulation are quite promising and this in the sense that all the $\alpha^{\star}$ values computed on the original samples are very close to the mean values of the distributions of the $\alpha^{\star s}$ obtained from simulation. Furthermore all the observed $\alpha^{\star}$ values lay in the intervals $[25\%\ \texttt{Qu.}; 75\%\ \texttt{Qu.}]$ of the distribution of the $\alpha^{\star s}$ and in some cases they

coincide with the median.
We remark (see Figure 2, panel(b)) that the six 90% intervals for $\alpha^{\star s}$, that is [5% $\texttt{Qu.}$; 95% $\texttt{Qu.}$], are not overlapping as m increases and, since all the notches of the plots do not overlap, we may affirm that there is a strong evidence that the m medians differ among them (Chambers et al., 1983).

4 Conclusions and future works

Given that "..all models are wrong, but some are useful" (Box, 1979), exploiting the inherent properties of the estimates based on the Maximum Likelihood and the Minimum Density Power Divergence criteria, we introduce and outline a procedure which can be helpful in parametric regression models building especially in those situations involving the study of large datasets where a substantial number of outliers or clustered data maybe present and data cleaning is impractical and statistical efficiency is a secondary concern.

The procedure we suggest allows simultaneously to detect the presence of outliers in the data and, if they are present, to select the best tuning constant for the Minimum Density Power Divergence estimators, for which computationally closed-form expressions are available so that solutions can be obtained applying any standard non linear optimization code. We also pointed out that, despite feasible computationally closed-forms expressions are available for the estimators, particular care must be taken in choosing the initial guesses of the minimizing routine, we suggest a random generations of initial guesses.

The core of the procedure relies on the concept of similarity between estimated regression models, for which a normalized index is introduced and a Monte Carlo significance test of statistical hypothesis is provided. From a computational point of view, the similarity index given by equation can easily be evaluated resorting to the algorithm suggested by Durio and Isaia (2010), advantageous in terms of parsimony of computing time, notably in high dimensional problems.

The procedure we advise seems to behave very well in all the experimental situations we explored and the results of a simulation, applied to several scenarios, validate this impression. We compute our simulations on various scenarious, in other to match some tipical situations that frequently arise in scientific areas (e.g. engineering, chemical, pharmaceutical) and a future work would be its applications on real datasets.

A more deeply study on the procedure performance would be a comparison about the results in terms of best tuning parameter given by our algorthm with respect to those obtained by other metods.

Futher features on the topic would be test our procedure for classical robust estimators such as regression quantile, regression rank scores or the densiti-based minimum divergence estimators proposed by Basu et al. (2001) that are based on those introduced by Windam (1995) .

Acknowledgements

Authors are indebted to the coordinating editor and to the anonymous referees for carefully reading the manuscript and for their many important remarks and suggestions.

References

Barnard, G. A. (1963). Contribution to the discussion of paper by M.S. Bartlett. *Journal of the Royal Statistical Society, B 25*, 294.

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

Basu, A., Harris, I. R., Hjort, N., and Jones, M. (2001). A comparison of related density-based minimum divergece estimators. *Biometrika 88*, 865–873.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. Launer and G. Wilkinson (Eds.), *Robustness in Statistics*, pp. 201–235. Academic Press.

Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983). *Graphical Methods for Data Analysis.* Wadsworth & Brooks/Cole.

Daniel, C. and Woods, F. S. (1977). Detection of influential observation in linear regression. *Technometrics 19*, 15–18.

Davies, P. L. (1993). Aspects of robust linear regression. *Annals of Statistics 21*, 43–99.

Dodge, Y. and Jurečkova, J. (2000). *Adaptive Regression.* Springer-Verlag.

Durio, A. and Isaia, E. D. (2003). A parametric regression model by minimum L_2 criterion: a study on hydrocarbon pollution of electrical transformers. *Developments in Applied Statistics, Metodološki Zvezki 19*, 69–83.

Durio, A. and Isaia, E. D. (2004). On robustness to outliers of parametric L_2 estimate criterion in the case of bivariate normal mixtures: a simulation study. In Hubert, A. S. M., Pison, G. and Aelst, S. V. (Eds.), *Theory and Applications of Recent Robust Methods*, pp. 93–104. Birkhäuser.

Durio, A. and Isaia, E. D. (2010). Clusters detection in regression problems: a similarity test between estimated. *Communications in StatisticsTheory and Methods*,39, 508516.

Fujisawa, H. and Eguchi, F. (2006). Robust estimation in the normal mixture model. *Journal of Statistical Planning Inference, B 136(11)*, 3989–4011.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1987). *Robust Regression and Outlier Detection.* John Wiley.

Hope, A. C. (1968). A simplified monte carlo significance test procedure. *Journal of the Royal Statistical Society, B 30*, 582–598.

Huber, P. J. (1981). *Robust Statistics.* John Wiley, New York.

Maronna, R., Martin, D., and Yohai, V. (2006). *Robust Statistics: Theory and Methods.* John Wiley, New York.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection.* John Wiley.

Rousseeuw, P. J., Van Alest, S., Van Driessen, K. and Agulló, J. (2004). Robust multivariate regression. *Technometrics 46*, 293–305.

Scott, D. W. (2001). Parametric statistical modeling by minimum integrated square error. *Technometrics 43*, 274–285.

Seber, G. A. F. and Lee, A. J. (2003). *Linear Regression Analysis* (2 ed.). John Wiley.

Staudte, R. G. and Sheather, S. J. (1990). *Robust Estimation and Testing.* John Wiley.

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

Windam, M. P. (1995). Robustifying model fitting. *Journal of the Royal Statistical Society, B 57*, 599–609.