Clusters Detection in Regression Problems: A Similarity Test Between Estimated Models

A. Durio, E. D. Isaia
Clusters Detection in Regression Problems: A Similarity Test Between Estimated Models
COMMUNICATIONS IN STATISTICS. THEORY AND METHODS (2010) 39
DOI: 10.1080/03610920903140254

Clusters Detection in Regression Problems: a Similarity Test Between Estimated Models

Alessandra Durio[1] and Ennio Davide Isaia[1]

Department of Statistics & Applied Mathematics
University of Torino
Piazza Arbarello, 8
10122 - Torino, Italy
(e-mail: durio@econ.unito.it,isaia@econ.unito.it)

Abstract Purpose of this paper is to investigate on the use of the Minimum Integrated Square Error criterion as a practical tool in building useful regression models, notably in all those situations involving the study of large data sets where a substantial number of outliers can be present or data are clustered. We suggest a technique of regression analysis which consists in comparing the results arising from L_2 estimates with the ones obtained applying some common M-estimators. A new index of similarity between functions is proposed and a Monte Carlo test of hypothesis based on it is introduced. Rejecting the hypothesis of similarity between the estimated regression models implies a careful investigation of data structure. Results of a simulation study, referring to several experimental scenarios, are provided to illustrate the approach we propose.
Keywords: M-estimators, Minimum integrated square error, Montecarlo significance test, Robust regression.

1 Introduction

In applied statistics regression is one of the most used tool in establishing the relationship between a set of predictor variables and a response variable. Given that "... all models are wrong, but some are useful" [Box, 1979], in the following we investigate on the use of Integrated Square Error as practical estimation tool for parametric regression models. The approach based on minimizing the Integrated Square Error (L_2 criterion) is particularly helpful in all those situations where, due to large sample size, a careful data preparation is not feasible and hence data may be heavily contaminated by a substantial number of outliers [Scott, 2001].

In the following we suggest a technique of regression analysis based on the comparison of the L_2 estimates with the ones obtained applying some common M-estimators; this comparison is based on a new index of similarity between functions. Afterwards, a Monte Carlo Significance test of hypothesis based on this statistics is introduced in

order to verify the hypothesis of similarity between the L_2 estimated regression model and the one obtained resorting to an M-estimator. Whenever the hypothesis of similarity between the two estimated regression models is rejected, we suggest to investigate more carefully the data structure in order to check about the presence of clusters or outliers. In this sense L_2 criterion can be viewed as a practical diagnostic tool in building useful models.

Theory is outlined and main results of a simulation study, referring to several experimental scenarios, are provided to illustrate and corroborate the approach we propose.

2 Robust estimators in regression models

A problem with regression techniques is the effect of outliers; these may occur for three main reasons, namely *recording errors*, *inclusion of cases with special characteristics* and *modeling errors* caused by choosing the wrong model. An extreme situation arises when the outliers are numerous and they underlie a set of clustered data. Data set with a large proportion of outliers can be found, for instance, if there is an omitted categorical variable (e.g. gender, species, geographical location, ...) where the data behave differently for each category.

Since outliers can play havoc with standard statistical methods, many robust techniques have been developed since the early work of [Huber, 1964] to be less sensitive to outliers, e.g. [Hampel, 2001] and [Maronna *et al.*, 2006].
Let $\{(x_{i1}, \ldots, x_{ip}, y_i)\}_{i=1,\ldots,n}$ be the observed data set, where each observation stems from a random sample drawn from the $p+1$ random variable $(X_1, \ldots, X_p, Y)$. The regression model for the observed data set we study is $y_i = m_{\boldsymbol{\beta}}(\mathbf{x}_i) + \varepsilon_i$ for $i = 1, \ldots, n$, where the object of our interest is the regression mean

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

and the errors $\{\varepsilon_i\}_{i=1,\ldots,n}$ are assumed to be independent random variables with zero mean and unknown finite variances.

The class of M-estimators of the vector $\boldsymbol{\beta}$ is defined as [Huber, 1981]

$$\hat{\boldsymbol{\beta}}_M = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho\left(y_i - m_{\boldsymbol{\beta}}(\mathbf{x}_i)\right) \tag{2}$$

where $\rho : \mathbb{R} \to \mathbb{R}$ is absolutely continuous, usually a convex function with derivative ψ.

The classical least-squares estimator is obtained from equation (2) simply setting $\rho(y_i - m_{\boldsymbol{\beta}}(\mathbf{x}_i)) = (y_i - m_{\boldsymbol{\beta}}(\mathbf{x}_i))^2$. Assuming that the r.vs ε_i are independent and identically distributed as the r.v. $\varepsilon \sim \mathcal{N}(0, \sigma)$, the least-squares estimator gives the Maximum Likelihood Estimate (MLE) of the vector $\boldsymbol{\beta}$ in equation (1), i.e.

$$\hat{\boldsymbol{\beta}}_{MLE} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left[y_i - m_{\boldsymbol{\beta}}(\mathbf{x}_i) \right]^2 \tag{3}$$

For our purpose, since in presence of outliers Maximum Likelihood estimates are very unstable, in the class of M-estimators we resort to the robust Huber M-estimator (HME) for which

$$\rho(y_i - m_{\boldsymbol{\beta}}(\mathbf{x}_i)) = \begin{cases} \dfrac{1}{2}(y_i - m_{\boldsymbol{\beta}}(\mathbf{x}_i))^2 & \text{if } \left| y_i - m_{\boldsymbol{\beta}}(\mathbf{x}_i) \right| \leq k \\ k \left| y_i - m_{\boldsymbol{\beta}}(\mathbf{x}_i) \right| \left(1 - \dfrac{k}{2}\right) & \text{if } \left| y_i - m_{\boldsymbol{\beta}}(\mathbf{x}_i) \right| > k \end{cases}$$

where the *tuning constant* k is set to $1.345\,\sigma$ [Hampel *et al.*, 1986].

As an alternative to the class of M-estimators, we investigate on parametric linear regression models based on Minimum Integrated Square Error [Durio and Isaia, 2003]. Our choice can be motivated by the fact that in the α-family of estimators, L_2 estimator, briefly L_2E, is the more robust to outliers, even if it is less efficient than MLE [Basu *et al.*, 1998].

Given the r.v. X, with unknown density $f(x|\boldsymbol{\theta}_0)$, for which we introduce the model $f(x|\boldsymbol{\theta})$, the estimate of the vector $\boldsymbol{\theta}_0$ according to the minimum Integrated Square Error criterion is given by

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}} \int_{\mathbb{R}} \left[f(x|\boldsymbol{\theta}) - f(x|\boldsymbol{\theta}_0) \right]^2 dx = \\ &= \arg\min_{\boldsymbol{\theta}} \left[\int_{\mathbb{R}} f^2(x|\boldsymbol{\theta})\, dx - 2\, \mathbb{E}\left[f(x|\boldsymbol{\theta}_0) \right] \right] \end{aligned} \tag{4}$$

The first term of equation (4) may be replaced by [Basu *et al.*, 1998]

$$\int_{\mathbb{R}} f^2(x|\boldsymbol{\theta})\, d\, x = \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}} f^2(x_i|\boldsymbol{\theta})\, d\, x_i \tag{5}$$

and if we replace in (4) the so called expected height of the density, $\mathbb{E}\left[f(x|\boldsymbol{\theta}_0)\right]$, with its estimate $\widehat{\mathbb{E}}\left[f(x|\boldsymbol{\theta}_0)\right] = n^{-1}\sum_{i=1}^{n} f(x_i|\boldsymbol{\theta})$, the estimator for $\boldsymbol{\theta}_0$ minimizing the L_2 distance will be

$$\hat{\boldsymbol{\theta}}_{L_2E} = \arg\min_{\boldsymbol{\theta}} \left[\frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}} f^2(x_i|\boldsymbol{\theta})\, d\, x_i - \frac{2}{n}\sum_{i=1}^{n} f(x_i|\boldsymbol{\theta})\right] \tag{6}$$

In the case of parametric regression, assuming that the random variables $Y|\mathbf{x}$ are distributed as a $\mathcal{N}(m_{\boldsymbol{\beta}_0}(\mathbf{x}), \sigma_0)$, the model becomes

$$f_{Y|\mathbf{x}}(y|\boldsymbol{\beta}_0, \sigma_0) = \phi(y|m_{\boldsymbol{\beta}_0}(\mathbf{x}), \sigma_0)$$

and the L_2 estimates of the parameters in $\boldsymbol{\beta}_0$ and σ_0 are simultaneously given by equation (6), which in this case becomes

$$(\hat{\boldsymbol{\beta}}, \hat{\sigma})_{L_2E} = \arg\min_{\boldsymbol{\beta},\sigma} \left[\frac{1}{2\sigma\sqrt{\pi}} - \frac{2}{n}\sum_{i=1}^{n}\phi(y_i|m_{\boldsymbol{\beta}}(\mathbf{x}_i), \sigma)\right] \tag{7}$$

since from equation (5) we have $\int_{\mathbb{R}} \phi^2(y|m_{\boldsymbol{\beta}}(\mathbf{x}), \sigma)\, dy = (2\,\sigma\,\sqrt{\pi})^{-1}$.

Clearly equation (7) is a feasible computationally closed-form expression so that L_2 criterion can be performed by any standard non linear optimization code, for instance resorting to the `nlm` routine of the `R` library.

3 The similarity index

To compare the L_2E performance with respect to some other common estimators we introduce an index of similarity between regression models.

To this end, let 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. In order to measure the discrepancy between the two estimated regression models, we resort to an index of similarity which 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 the whole data points lay. To this goal, introduced the sets

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

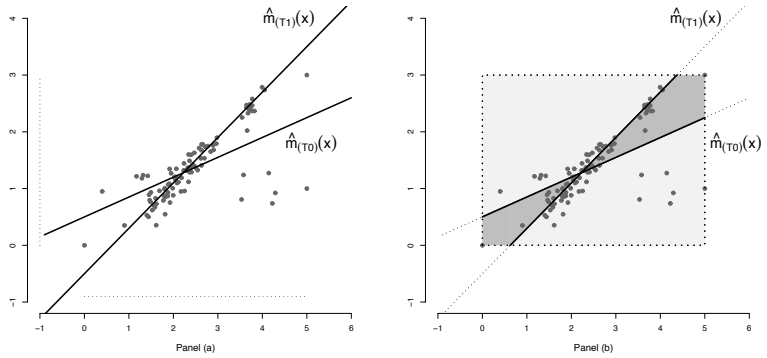

Figure1. Data points and two estimated regression models, namely $\hat{m}_{T0}(x)$ and $\hat{m}_{T1}(x)$. Panel (b) shows the sets $\boldsymbol{C}^2$ (in light gray) and $\boldsymbol{D}^2$ (in dark gray).

we define the *similarity index* as

$$
\begin{aligned}
&sim(T_0, T_1) \stackrel{def}{=} \frac{\int_{\boldsymbol{D}^{p+1}} d\mathbf{t}}{\int_{\boldsymbol{C}^{p+1}} d\mathbf{t}} \\
&\boldsymbol{C}^{p+1} = \boldsymbol{I}^p \times \boldsymbol{I} \\
&\boldsymbol{D}^{p+1} = \left\{(\mathbf{x}, y) \in \mathbb{R}^{p+1} : \zeta(\mathbf{x}) \leq y \leq \xi(\mathbf{x}), \mathbf{x} \in I^p\right\} \cap \boldsymbol{C}^{p+1}
\end{aligned}
\tag{8}
$$

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

Figure 1 shows how the similarity index given by equation (8) can be computed in the simple case when $p = 1$. In panel (a) data points and the two estimated models $\hat{m}_{T_0}(\mathbf{x})$ and $\hat{m}_{T_1}(\mathbf{x})$ are plotted. The dark-gray shaded area of panel (b) corresponds to $\int_{\boldsymbol{D}^{p+1}} d\mathbf{t}$, while the integral $\int_{\boldsymbol{C}^{p+1}} d\mathbf{t}$ is given by the area of the light-gray rectangle in which data points lay.

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 $sim(T_0, T_1)$ close to unit.

From a computational point of view it is rather onerous to evaluate the integral present in the numerator of equation (8), since it implies splitting the domain $\boldsymbol{D}^{p+1}$ into "simple regions" with respect to the variable y. Resorting to this direct method, we are obliged to find the regions of intersections of both models $\hat{m}_{T_0}(\mathbf{x})$ and $\hat{m}_{T_1}(\mathbf{x})$ with the

boundary of $\boldsymbol{D}^{p+1}$ as well as the region of intersection between the two regression models.

We therefore decide, in order to compute the similarity index $sim(T_0, T_1)$, to resort to a numerical routine based on he following algorithm

Algorithm 1

1 : move the origin of the system to $(\min(x_{i1}), \ldots, \min(x_{ip}), \min(y_i))$ and compute the integrals

$$A = \int_{\boldsymbol{I}^p} \left| \hat{m}_{T_0}(\mathbf{x}) - \hat{m}_{T_1}(\mathbf{x}) \right| \, d\mathbf{x}$$

$$A_j^- = 0.5 \left(\int_{\boldsymbol{I}^p} \left| \hat{m}_{T_j}(\mathbf{x}) \right| \, d\mathbf{x} - \int_{\boldsymbol{I}^p} \hat{m}_{T_j}(\mathbf{x}) \, d\mathbf{x} \right) \qquad \text{for } j = 0, 1$$

2 : compute the value of $\texttt{butterfly} = A - (A_0^- + A_1^-)$;
3 : check if both models are negative on the vertexes of the domain $\boldsymbol{I}^p$ and if so $\texttt{butterfly} = \texttt{butterfly} + 2 \min(A_0^-, A_1^-)$;
4 : move the origin of the system to $(\min(x_{i1}), \ldots, \min(x_{ip}), \max(y_i))$ and compute the integrals

$$A_j^+ = 0.5 \left(\int_{\boldsymbol{I}^p} \left| \hat{m}_{T_j}(\mathbf{x}) \right| \, d\mathbf{x} + \int_{\boldsymbol{I}^p} \hat{m}_{T_j}(\mathbf{x}) \, d\mathbf{x} \right) \qquad \text{for } j = 0, 1$$

5 : compute the value of $\texttt{butterfly} = \texttt{butterfly} - (A_0^+ + A_1^+)$;
6 : check if both models are positive on the vertexes of the domain $\boldsymbol{I}^p$ and finally $sim(T_0, T_1) = \texttt{butterfly} + 2 \min(A_0^+, A_1^+)$.

Algorithm (1) implies a numerical evaluation of nine integrals of a p-dimensional function over a simple domain and two checks on the values of the functions $\hat{m}_{T_0}(\mathbf{x})$ and $\hat{m}_{T_1}(\mathbf{x})$ on the vertexes of the domain $\boldsymbol{I}^p$.

It is worthwhile to remark that, given two models $\hat{m}_{T_0}(\mathbf{x})$ and $\hat{m}_{T_1}(\mathbf{x})$, the results we obtain applying Algorithm (1) are of the same order as those we can get computing the the integral present in the numerator of equation (8) with the direct method, i.e. splitting the domain $\boldsymbol{D}^{p+1}$ into "simple regions" with respect to the variable y.

Though the errors rising from numerical integration according to Algorithm (1) are comparable in magnitude with those arising from numerical roots finding routines applying the direct method, the approach outlined by Algorithm (1) is preferable in terms of parsimony of computing time, notably in high dimension problems.

4 Monte Carlo significance test

We propose to use the $sim(T_0, T_1)$ statistics given by equation (8) to verify the following system of hypothesis

$$\begin{cases} H_0 : \boldsymbol{\beta}_0 = \hat{\boldsymbol{\beta}}_{T_0} \\ H_1 : \boldsymbol{\beta}_0 \neq \hat{\boldsymbol{\beta}}_{T_0} \end{cases} \tag{9}$$

Since it seems not reasonable 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 (M.C.S. test), originally suggested by [Barnard, 1963] and later proposed by [Hope, 1968].

Let $sim_{T_0T_1}$ denote the value of the $sim(T_0, T_1)$ statistic computed on the observed data. The simplified Monte Carlo Significance Test consists in rejecting H_0 if $sim_{T_0T_1}$ is the $m\,\alpha$-th most extreme statistic relative to the corresponding quantities based on the random samples of the reference set, where the reference set consists in $m-1$ random samples, of size n each, generated under the null hypothesis, i.e. drawn at random from the model $\hat{m}_{T_0}(\mathbf{x})$ with $\sigma = \hat{\sigma}_{T_0}$. In other words we generate $m-1$ random samples under H_0 and for each of them we compute $sim^*_{T_0T_1}$ and we shall reject the null hypothesis, at the α significance level, if and only if the value of the test statistic $sim_{T_0T_1}$ is greater than all the $m-1$ values of $sim^*_{T_0T_1}$. We remark that if we set $m\,\alpha = 1$ and fix $\alpha = 0.01$, we have $m - 1 = 99$, while fixing $\alpha = 0.05$ would yield $m - 1 = 19$.

The rejection of the hypothesis of similarity between the two estimated models must be interpreted as a warning signal suggesting a more careful investigation of the data structure in order to check about the presence of clusters or outliers.

5 Main results of a simulation study

At this point it is worthwhile to have a look at some results from a simulation study we carried out to check the goodness of our procedure and to illustrate the behaviour of L_2 criterion in presence of clustered data.

To this end, we set up some different experimental configurations according to a specified data generating model and for each of them

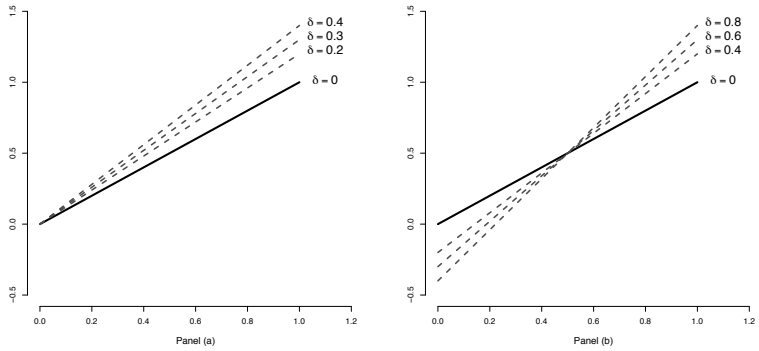

Figure2. Main model $Y = X$ for Scenario I and perturbing models (a) $Y = (\delta+1)X$ with $\delta = 0.20(0.30, 0.40)$ and (b) $Y = -0.5\,\delta + (\delta + 1)X$ with $\delta = 0.40(0.60, 0.80)$.

we consider, beside L_2 estimator, the Maximum Likelihood estimator and the robust Huber M-estimator. Fixing in equation (8) $T_1 = L_2E$, we shall perform the M.C.S. test two times: the first one fixing $T_0 = MLE$ for $sim(MLE, L_2E)$, the second one fixing $T_0 = HME$, for $sim(HME, L_2E)$.

For each scenario, we draw $H = 100$ random samples of size $n = 600$ and on each of them we perform both the two M.C.S. tests at $\alpha = 0.01$ ($m = 99$) recording the number of times that they lead us to reject the null hypothesis of system (9) when it is actually false.

In the following we provide and comment three experimental configurations featuring 80% of data point belonging to one (main) cluster and the remaining 20% to a second cluster. In these situations L_2 criterion will fit the heaviest cluster [Durio and Isaia, 2004] and it behaves differently from the other two estimators as the clusters tend to be well separated.

More precisely, the three scenarios we provide are

Scenario I (one predictor):
we consider a simulated data set of $n_1 = 480$ points generated according to $Y = X + \varepsilon$, where $X \sim \mathcal{U}(0, 1)$ and $\varepsilon \sim \mathcal{N}(0, 0.1)$. We then introduce $n_2 = 120$ points according to the models

(a) $Y = (\delta + 1)X + \varepsilon$ with $\delta = 0.20(0.30, 0.40)$.

(b) $Y = -0.5\,\delta + (\delta + 1)X + \varepsilon$ with $\delta = 0.40(0.60, 0.80)$.

Scenario II (two predictors):
we consider a simulated data set of $n_1 = 480$ points generated according to $Y = \sum_{i=1}^{2} 0.5\, X_i + \varepsilon$, where $X_i \sim \mathcal{U}(0,1)$ and $\varepsilon \sim \mathcal{N}(0, 0.1)$. We then introduce $n_2 = 120$ points according to the models

(a) $Y = \sum_{i=1}^{2} (\delta + 0.5) X_i + \varepsilon$ with $\delta = 0.10(0.15, 0.20)$.

(b) $Y = -\delta + \sum_{i=1}^{2} (\delta + 0.5) X_i + \varepsilon$ with $\delta = 0.20(0.30, 0.40)$.

Scenario III (four predictors):
we consider a simulated data set of $n_1 = 480$ points generated according to $Y = \sum_{i=1}^{4} 0.25\, X_i + \varepsilon$, where $X_i \sim \mathcal{U}(0,1)$ and $\varepsilon \sim \mathcal{N}(0, 0.1)$. We then introduce $n_2 = 120$ points according to the models

(a) $Y = \sum_{i=1}^{4} (\delta + 0.25) X_i + \varepsilon$ with $\delta = 0.05(0.075, 0.10)$.

(b) $Y = -2\,\delta + \sum_{i=1}^{4} (\delta + 0.25) X_i + \varepsilon$ with $\delta = 0.10(0.15, 0.20)$.

Figure 2 shows, for Scenario I, the main model and the perturbing models in both (a) and (b) sub-configurations. Main and perturbing models of Scenario II and Scenario III have the same shape but obviously in high-dimensions.

If we consider the results displayed in Table 1 as empirical powers of the M.C.S. tests, we obviously remark that the power of the test increases as the parameter δ increases.

For each configuration the percentage of times we correctly reject the similarity between L_2E and MLE is greater then the corresponding percentage between L_2E and HME. This is due to the higher robustness of HME with respect to MLE and this confirms the properties of robustness of L_2E.

Comparing the (a) cases of the scenarios provided, it follows that when the number p of predictors increases the behaviour of the three estimators tends to be similar. However, even in the worst situations in which clusters are very confounded (low levels of the parameter δ), we had samples for which the M.C.S. test led to reject the hypothesis of similarity.

For all the scenarios the empirical power of the tests for (b) cases is lower than the one of the corresponding (a) cases. This is due to the fact (see Figure 2) that for these configurations the two models generating the data sets for (b) cases intersect in the centre of mass of the domain $\boldsymbol{I}^p$ and hence the two clusters are more confused than those of (a) cases.

Scenario I						
	$T_1 = MLE$	$T_1 = HME$	$T_1 = MLE$	$T_1 = HME$	$T_1 = MLE$	$T_1 = HME$
a	$\delta = 0.20$		$\delta = 0.30$		$\delta = 0.40$	
	14	10	92	82	100	98
b	$\delta = 0.40$		$\delta = 0.60$		$\delta = 0.80$	
	18	11	89	78	100	99
Scenario II						
	$T_1 = MLE$	$T_1 = HME$	$T_1 = MLE$	$T_1 = HME$	$T_1 = MLE$	$T_1 = HME$
a	$\delta = 0.10$		$\delta = 0.15$		$\delta = 0.20$	
	6	5	56	50	99	97
b	$\delta = 0.20$		$\delta = 0.30$		$\delta = 0.40$	
	5	5	36	30	86	80
Scenario III						
	$T_1 = MLE$	$T_1 = HME$	$T_1 = MLE$	$T_1 = HME$	$T_1 = MLE$	$T_1 = HME$
a	$\delta = 0.05$		$\delta = 0.075$		$\delta = 0.10$	
	6	4	26	24	92	84
b	$\delta = 0.10$		$\delta = 0.15$		$\delta = 0.20$	
	4	3	22	20	48	42

Table1. Main results of the simulation study (values in percentage).

Besides these considerations, it must be pointed out that the estimates of the parameters of the models are, in terms of their standard deviation, very stable in all the situations, even if L_2 estimates show a somewhat "natural" bigger dispersion.

We finally remark that the empirical powers of the test we obtained repeating the simulation for the three scenarios with different percentage of data points drawn from the second cluster are higher than those we provide here when this contamination percentage is 10% and they are lower when it is 40%.

Acknowledgements

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

References

[Barnard, 1963]G. A. Barnard. Contribution to the discussion of paper by m.s. bartlett. *Journal of the Royal Statistical Society, B*, 25:294, 1963.

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

[Box, 1979]G. E. P. Box. Robustness in the strategy of scientific model building. In R. Launer and G. Wilkinson, editors, *Robustness in Statistics*, pages 201–235. Academic Press, 1979.

[Durio and Isaia, 2003]A. Durio and E. D. Isaia. 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, 2003.

[Durio and Isaia, 2004]A. Durio and E. D. Isaia. On robustness to outliers of parametric l_2 estimate criterion in the case of bivariate normal mixtures: a simulation study. In A. Struyf M. Hubert, G. Pison and S. Van Aelst, editors, *Theory and Applications of Recent Robust Methods*, pages 93–104. Birkhäuser, 2004.

[Hampel *et al.*, 1986]F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics : The Approach Based on Influence Functions.* John Wiley, 1986.

[Hampel, 2001]F. Hampel. Robust statistics: A brief introduction and overview. Technical Report 94, Seminar für Statistik, Eidgenssische Technische Hochschule, 2001.

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

[Huber, 1964]P. J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:73–101, 1964.

[Huber, 1981]P. J. Huber. *Robust Statistics.* John Wiley, New York, 1981.

[Maronna *et al.*, 2006]R. Maronna, D. Martin, and V. Yohai. *Robust Statistics: Theory and Methods.* John Wiley, New York, 2006.

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