HIGH ROBUSTNESS AND AN ILLUSION OF "TRUTH"

Jan Ámos Víšek, Prague

Abstract

The popular misunderstanding of the concept of robustness is discussed and illustrated by examples. It is also shown that even the highly robust procedures do not guarantee automatically anything, and that some a posteriori test is (at least sometimes) inevitable.

Key words: Regression, least trimmed squares, least median of squares, diversity of estimates, selection of model
AMS classification: Primary 62F35, secondary 62F10

THE CONCEPTS OF ROBUSTNESS AND THEIR DISCUSSION

The experiences of the generation of statisticians have yielded a numerous families of the classical (parametric) probabilistic models. Much more recent experience says that even very carefully measured data frequently contain a fraction (1% - 10%) of (gross) errors, which are not necessarily always easy to distinguish from the "proper data" (a lot of nice examples may be found in [3, 4, 7]). In the theoretical counterpart it means that we should not assume that the random variables are governed by a classical probabilistic model, but by a model which may slightly differ from the classical one. Unfortunately, it appeared that many (classical) statistical functionals (estimators or test statistics) are not *distributionally* continuous on the space of (all) probabilistic models, i.e. a small change of the underlying model may cause a large change of the distribution of the statistic in question. The consequence for the applications is that the contamination of data may have a fatal influence on the behaviour of the functional.

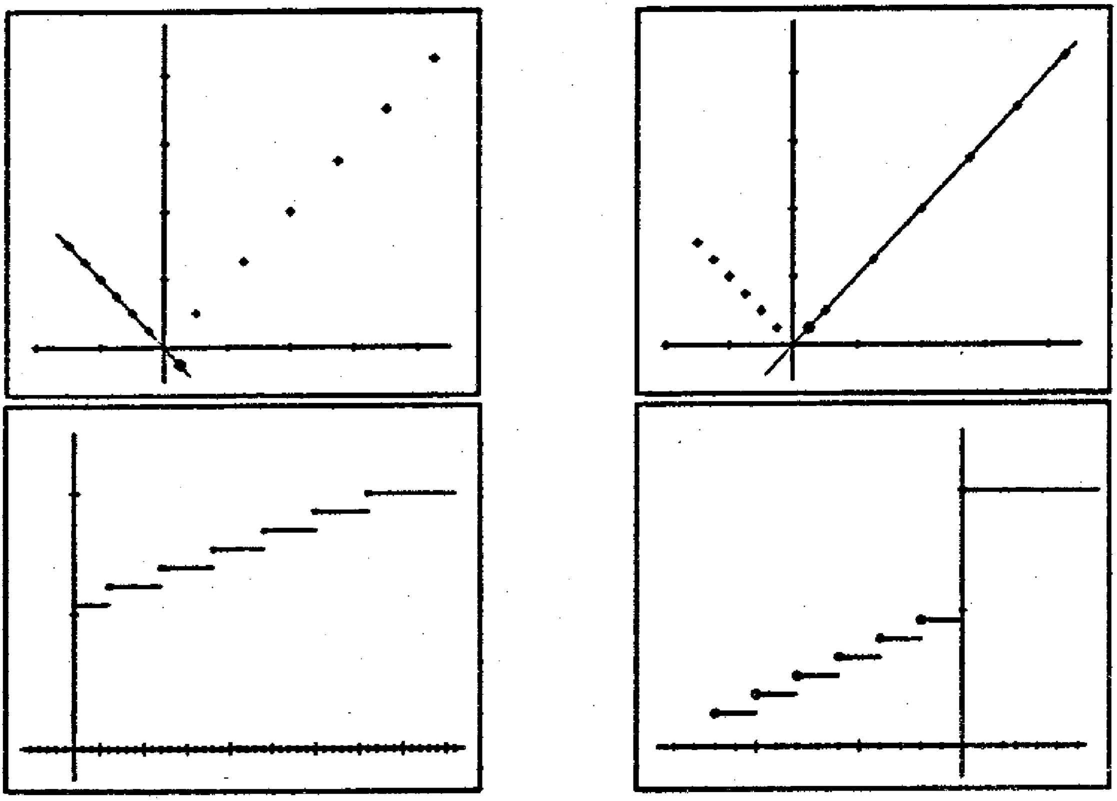
The description of the problem given in the previous lines immediately hints a possible remedy for the troubles, namely we should restrict ourselves on the distributionally continuous functionals (qualitative robustness). In other words, we should look for such functionals for which small changes of underlying probabilistic model cause only small changes of their distributions.

Of course, in a nonprobabilistic processing data various attempts how to cope with the contamination of data have also appeared. They have suggested to find such estimators for which any changes of a (small) fraction of data does not influence the value of the estimator too much (robustness with respect to the (outlying or inlying) data; [11], see the abstract and paragraph 4.6). In what follows we shall show that the latter approach may lead to considerable difficulties, namely that it is equivalent to the requirement that an estimator gives model highly stable but sometimes unfortunately for a minority of data. A critical discussion of the former one requires rather large space and will be therefore presented in the forthcoming paper.

What is common to the both approaches is the quite acceptable assumption that the data are mixture of some proper data, which may be described by a (usually) simple underlying model, and of a contamination which of course represents a minority of data ([9]). It implies that as a "true" underlying model is assumed such one which explains the majority of data (compare also [7], p. 56). Another reason supporting the idea of the construction of the model for the majority of the data is following. Usually we do not want to specify even the type of the distribution of the contamination, and hence it is not possible to say that the data are generated by such and such distribution which is a convex mixture of the given two distributions.

Earlier than we shall continue let us recall what we expect from a "good" estimator. Naturally, it is assumed that the estimator gives a result which is relevant for the identification of the underlying model, i. e. we expect that the estimator estimates the "true" underlying model. Taking into account what has been said a few lines above, in the case when data are not homogeneous we expect that the estimator estimates the model which fits to the majority of data. But then it is clear that if we could give an example in which a (small) change of a small fraction of data implies a large change of the underlying model it would demonstrate that the robustness with respect to the (outlying or inlying) data is the misleading concept. Indeed, if an estimator is robust with respect to the (outlying or inlying) data then the (small) change of the small fraction of data does not cause too large change of its value. On the other hand, as it was already said, in our example this change of data will indicate the large change of the underlying model. The conclusion is that the estimator is not able to recognize that the underlying models for the original data and for the modified ones are considerably different. The following exhibits demonstrates such situation (compare also [8] – exact-fit example).

EXHIBIT 1. "Decreasing" and "increasing" model
and the corresponding empirical distributions of residuals



Of course, the example which was presented by Exhibit 1 is a bit artificial, and in practice we would split data into two groups and process each of them separately. An example of the same final effect with the real data will be given below. A graphical analysis of that example also shows that it is sometimes impossible to divide the data into two parts and to study them separately. Unfortunately, the results of such graphical analysis would require a lot of space.

Since the concept of the qualitative robustness could not grasp the (all) desired features characterizing the robustness of the statistics (as a small sensitivity with respect to the gross-errors, a small sensitivity with respect to (large number of) the local shifts etc.) another concept of the robustness has been introduced, the concept of the quantitative robustness ([7]). One of attractive characteristics of the quantitative robustness is the breakdown point. Although the original definition ([6]) is a bit academic, the sample versions (see e. g. [5]) are more practical and transparent. Under the breakdown point (for given data containing n observations) we understand the ratio m/n where m is the smallest number of the observations which are necessary to change (in an arbitrary way) to turn the absolute value (or the norm) of the estimate into infinity (or for the estimate of variance: to turn the absolute value or the determinant of the estimate to zero or infinity). Sometimes instead of changing the observations we consider an extension of data, i. e. we assume that we add m observations to the n original ones; of course asymptotically it is equivalent. It is evident that generally the sample version of the breakdown point is sample-dependent, and hence it does not seem to be very useful for characterizing the estimator. Fortunately, it has proved that for many estimators the breakdown point is sample-independent. The breakdown point may be interpreted also as a limit value of the contamination level which turns the (asymptotic) bias of the estimate into infinity ([13]), but we shall not go into details.

It follows from the definition of the breakdown point that the higher is the breakdown point the more resistant against the contamination is the estimator. In the other words, if the estimator has the high breakdown point, in spite of the presence of the (large) contamination it gives an estimate which is not too far from the "true" model. As we shall see later the main trouble will be that it will not be clear, even theoretically, what is the "true" model.

It is clear that the upper bound of the (asymptotic) breakdown point is $\frac{1}{2}$, because if the contamination would represent more than 50 % of the data, it would be questionable what are the proper data and what is the contamination (although sometimes even such model may be reasonable; consider e. g. signal-noise model).

One of the most frequently discussed problem is the problem of estimating location parameter, and it is not difficult to find the estimator of location with the high breakdown point. The median, which was used much earlier than the robust statistics have been studied, is the example of the estimator with the (asymptotic) breakdown point equal to $\frac{1}{2}$.

In the regressional problem (which will be discussed in the rest of the paper it was much more complicated to find the estimators which attain (asymptotically) the upper bound $\frac{1}{2}$. Let us recall two of them (which will be used later in this discussion). First of all, however, we shall introduce some notation.

Let N denote the set of all positive integers and R^p the p-dimensional Euclidean space ($p \in N$). Finally, let $(\Omega, \mathcal{B}, P)$ be a prbability space. For all $i \in N$ we shall consider the regression model

$$Y_i(\omega) = X_i^{\mathrm{T}}(\omega)\,\beta^0 + e_i(\omega)$$

where $\{X_i(\omega)\}_{i=1}^{\infty}$, $X_i(\omega) : \Omega \to R^p$ is a sequence of independent random variables (carriers), $\beta^0 \in R^p$ and $\{e_i(\omega)\}_{i=1}^{\infty}$, $e_i(\omega) : \Omega \to R$ is another sequence of independent and identically distributed random variables (fluctuations), independent fom the sequence $\{X_i(\omega)\}_{i=1}^{\infty}$. For any $\beta \in R^p$ let $r_i(\beta,\omega) = Y_i(\omega) - X_i^{\mathrm{T}}(\omega)\,\beta$ be the i-th residual and $r^2_{(i:n)}(\beta,\omega)$ the i-th order statistic among $r_1^2(\beta,\omega), r_2^2(\beta,\omega), \ldots, r_n^2(\beta,\omega)$. Then the Least Median of Squares estimator is defined as

$$\hat{\beta}^{(n,h)}_{\mathrm{LMS}}(\omega) = \underset{\beta \in R^p}{\arg\min}\; r^2_{(h:n)}(\beta, \omega) \tag{1}$$

where $\frac{n}{2} \le h \le n$ (in the original Rousseeuw's proposal h was selected to be equal to $[\frac{n}{2}]$, hence the name of the estimator). Similarly the Least Trimmed Squares estimators is given as

$$\hat{\beta}^{(n,h)}_{\mathrm{LTS}}(\omega) = \underset{\beta \in R^p}{\arg\min} \sum_{i=1}^{h} r^2_{(i:n)}(\beta, \omega). \tag{2}$$

If we put $h = [\frac{n}{2}] + [\frac{p+1}{2}]$ the breakdown point of the both estimators is equal to $n^{-1} \cdot ([\frac{n-p}{2}] + 1)$ (see [18]). It means that asymptotically, these estimators have 50 % breakdown point. Now we are prepared to give the example which was promised above. We shall use the "Engine Knock Data" which are given in Table 1. They were published in [14] and later analyzed once again by Hettmansperger and Sheater [8].

TABLE 1. Engine Knock Data

case	Spark	Air	Intake	Exhaust	Knock
1	13.3	13.9	31	697	84.4
2	13.3	14.1	30	697	84.1
3	13.4	15.2	32	700	88.4
4	12.7	13.8	31	669	84.2
5	14.4	13.6	31	631	89.8
6	14.4	13.8	30	638	84.0
7	14.5	13.9	32	643	83.7
8	14.2	13.7	31	629	84.1
9	12.2	14.8	36	724	90.5
10	12.2	15.3	35	739	90.1
11	12.2	14.9	36	722	89.4
12	12.0	15.2	37	743	90.2
13	12.9	15.4	36	723	93.8
14	12.7	16.1	35	649	93.0
15	12.9	15.1	36	721	93.3
16	12.7	15.9	37	696	93.1

The "Knock" is assumed as the response variable, the others as the regressors. The LTS-estimates of the regression coefficients are given in the following Table 2.

TABLE 2. Estimates of the regression coefficients (with Air = 14.1)

Regressor	intercept	Spark	Air	Intake	Exhaust
$\hat{\beta}^{(n,h)}_{\mathrm{LTS}}$	35.1134	-0.0275	2.9490	0.4774	-0.0091

The values given in Table 2 were obtained so that the all 4 368 subsets containing 11 observations was taken into account and on each such subset the LS-estimator was applied. It means that the values given in Table 2 are precise solution of the minimization given in (2) (and not only an approximation as in the case of larger samples of the data). When we change the value of the "Air"-coordinate for the second observation to 15.1 (instead of the correct value 14.1) we obtain the following LTS-estimator of the regression coefficients.

TABLE 3. Estimates of the regression coefficients (data with Air = 15.1)

Regressor	intercept	Spark	Air	Intake	Exhaust
$\hat{\beta}^{(n,h)}_{\mathrm{LTS}}$	-88.7289	4.7194	1.0576	1.5693	0.0676

Comparing Tables 2 and 3 we find that the estimates for the correct data and for the (slightly) modified data are rather different. It means that we have obtained for real data a very similar effect as it was demonstrated in Exhibit 1. (Similar effect has been discovered for the correct and modified Engine Knock Data and for LMS-estimator by Hettmansperger and Sheater [8]. However it appeared later that it was due to an unsatisfactory approximation of the precise solution of (1) (see [25])).

It again confirms that the concept of robustness with respect to the (outlying or inlying) data is dubious. Of course, thinking about the robustness with respect to outlying or inlying data once again, one may perhaps come nearly immediately and without any other hints to the same conclusion which the examples given above have implied. Then it is even more surprising how the concept of robustness with respect to the outlying or inlying data is still overliving (see [12]). Maybe, it is an example of the situation which may be commented by the famous: *Any problem has a simple false solution.* Of course, one may immediately ask why then some of methods about which the authors claim that they are robust with respect to the outlying or inlying data are in fact robust in some other, easier and more corretly justifiable sense. It is a consequence of the fact that the procedures in question *accidently* coincide with such a procedure which might be derived in the statistical framework, see [15]. Moreover, using the statistical framework we may avoid some steps which cannot be heuristically fully justified. In the paper [11] which was refered at the beginning of this paper, such a step had the consequence of implementation of the metric into the observational space. Of course, it is usually a part of R^k, and so it may be assumed to be Euclidian space. On the other hand, no "natural" feelings about uncertainty implies any metric on the level of the observations. Notice also that the theory of probability corresponds with it, introducing the metrics just in the space of (all) distributions.

The discussion of the concept of qualitative robustness, i. e. robustness with respect to the small changes of the probabilistic model from the application point of view requires to introduce some other data and notions and hence will be discussed in some further paper. Instead of it we shall utilize the modified Engine Knock Data for the discussion of the (high) quantitative robustness. Earlier than doing that let us return for a while to the LMS and LTS-estimators.

DIVERSITY OF THE ESTIMATES

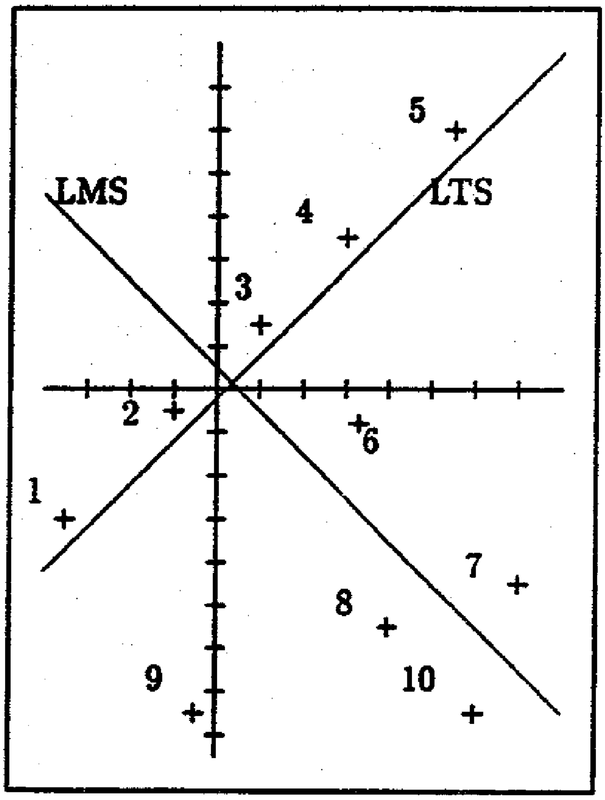
As it was already said the both estimators have the asymptotic breakdown point equal to $\frac{1}{2}$. It means that the contamination of data may be nearly 50 %, nevertheless the both estimates should be near to the "true" model. Due to this fact it seems that the both estimators may be used as diagnostic tools to reveal the influential points (outliers and wrong leverage points; see [7], p. 331 or [18], chapter 6.6). But for the modified Engine Knock Data we obtain following LMS-estimate of the regression coefficients (in following table we give also LTS-estimate, to facilitate comparison). The LMS-estimate of the coefficients have been evaluated by the software which is due to Pavel Boček and we are grateful for the possibility to use it. The method of the evaluation is based on the algorithm for the dual problem of linear programming (see [1]). This method was (much) quicker and gave smaller "median" of the squared residuals for any up to now processed data than any other available method (see [10]).

TABLE 4. Estimates of the regression coefficient (data with Air = 15.1)

Regressor	intercept	Spark	Air	Intake	Exhaust	11^{th} res.	sum of 11
$\hat{\beta}^{(n,h)}_{\text{LTS}}$	-88.72	4.7194	1.0576	1.5693	0.0676	0.539	0.728
$\hat{\beta}^{(n,h)}_{\text{LMS}}$	48.38	-0.7318	3.3925	0.1947	-0.011	0.450	1.184

("11^{th} res." means that it is the 11^{th} smallest value among the squared residuals; similarly "sum of 11" means that it is the sum of 11 smallest values among the squared residuals.) Analyzing Table 4 we conclude that the estimates are considerably different. One may have a suspicion that it is due to the small number of observations. Rousseeuw ([17], paragraph 7) proposed a rule of thumb that there should be at least 5 observations per dimension. However, it is possible to enlarge the number of observations in the Engine Knock Data (we have at hand already such data) and to show that the effect will be preserved. We shall not present these data here, instead of it we prefer to give an artificial data which enlighten the problem better.

EXHIBIT 2. LTS and LMS estimates for the artificial data



The Exhibit 2 shows that the key problem is that the data may be explained (or fitted, if you want) by two, rather different "true" models, and it is impossible to decide, without some additional a posterior criterium of quality (or of fit) of the estimate (not of the estimator), which of the estimates is better. Let us analyze the situation which is depicted in Exhibit 2. The data may be evidently explained as a mixture of "proper" data and the contamination. But it may be done in two ways. One model may consider the points $1, 2, 3, 4, 5$ and 6 as the proper data and the rest as the contamination, another model may propose to consider the points $2, 3, 6, 7, 8, 10$ as the proper data and the rest as the contamination. One may however object that with the increasing number of observations (i. e. asymptotically) the points from the proper model has to reach majority. But on the other hand it may appear that the points 1, 4, 5, 7, 8, 9 and 10 are atypical (although some of them may be good leverage points; keep in mind that we have assumed random carriers, although even for "deterministically" selected points in the factor space a similar situation may appear). So the observations which we shall obtain when we shall increase size of sample may fall into the region of the points 2, 3 and 6, and then even for a very large number of observations the effect of considerably different values of the (highly) robust estimators may take place.

FORMALIZING THE DIVERSITY OF ESTIMATES

First of all, let us recall that we consider the linear regression model

$$Y_i(\omega) = X_i^T(\omega)\,\beta + e_i(\omega) \quad i = 1, 2, \ldots \quad . \tag{3}$$

In what follows we shall understand by the linear regression model usually the equality (3) for the first n indeces and we shall abbreviate it by $(\mathcal{D}_n(\omega), \beta)$ where $\mathcal{D}_n(\omega)$ stays for $(Y_i(\omega), X_i(\omega))$ $i = 1, 2, \ldots, n$, possibly with an upper index which will distinguish among different models.

The task is to "explain" the *unique* sample of data given as a $(n \times (p+1))$-matrix of numbers

$$D_n = \begin{matrix} y_1, x_{11}, & \cdots, & x_{1p} \\ y_2, x_{21}, & \cdots, & x_{2p} \\ \vdots & & \vdots \\ y_n, x_{n1}, & \cdots, & x_{np} \end{matrix} \quad .$$

We usually *implicitly* assume that the data have been generated by a linear regression model $(\mathcal{D}_n^0(\omega),\ \beta^0)$, i. e. that for some $\omega_0 \in \Omega$ we have

$$y_i = Y_i^0(\omega_0) \quad \text{and} \quad x_{ij} = X_{ij}^0(\omega_0) \tag{4}$$

for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$. Notice that ω_0 generally depends on n, p, $\mathcal{D}_n^0(\omega)$, etc. We have used the word *implicitly* to indicate that we do not usually try to specify fully Ω, $\mathcal{B}$, P, $\{Y_i^0(\omega)\}_{i=1}^\infty$, $\{X_i^0(\omega)\}_{i=1}^\infty$, $\{e_i^0(\omega)\}_{i=1}^\infty$, $\beta^0 \in R^p$ and ω_0 so that (3) and (4) are fulfilled but we want only to estimate the unknown vector β^0 (for which some linear regression model fulfilling (3) and (4) may exist; of course there may be a whole family of models – maybe with different β's and generally also different distribution functions F – all of them fulfilling ((3) and 4) as well). Sometimes we try to estimate also the d. f. F of residuals, nevertheless again we do not specify fully $\mathcal{D}_n(\omega)$. In what follows in the case when we shall consider a collection of linear regression models, say $(\mathcal{D}_n^1(\omega), \beta^1)$, $(\mathcal{D}_n^2(\omega),\ \beta^2), \ldots, (\mathcal{D}_n^k(\omega), \beta^k)$, we shall abbreviate the equalities (4) for ℓ-th linear regression model as $\mathcal{D}_n^\ell(\omega_\ell) = D_n$.

In order to find an estimate $\hat{\beta}$ (a vector of numbers) of β^0 we construct in the framework of linear regression models $\{\mathcal{D}_n^0(\omega),\ \beta^0\}_{n=1}^\infty$ a sequence of estimators $\hat{\beta}(\mathcal{D}_n^0(\omega))$: $\Omega \to R^p$, $n = 1, 2, \ldots$ such that $\hat{\beta}(\mathcal{D}_n^0(\omega)) \to \beta^0$ a. s. or in probability as $n \to \infty$. Possibly the estimator has some other other plausible properties, e. g. small gross-error sensitivity, high breakdown point, etc. Notice how the construction of the estimator has been performed or, in other words, what is the usual structure of the theoretical assertions: We keep fix the sequence of regression models, i. e. we keep fix the sequence of carriers $\{X_i^0(\omega)\}_{i=1}^\infty$, response variables $\{Y_i^0(\omega)\}_{i=1}^\infty$ (or equivalently the sequence of random fluctuations $\{e_i^0(\omega)\}_{i=1}^\infty$) and a fix β^0, and then we prove some properties of the estimator $\hat{\beta}(\mathcal{D}_n^0(\omega))$.

In what follows, in the case when we shall not want to specify the regression model $(\mathcal{D}_n(\omega), \beta)$, we shall write also $\hat{\beta}(\omega)$ for an estimator to emphasize only the dependence on ω and to distinguish it from the estimate $\hat{\beta}$, i. e. from the vector of numbers. Having defined the estimator $\hat{\beta}(\mathcal{D}_n^0(\omega))$ (please keep in mind that it is a function of the system of random variables $\{\mathcal{D}_n^0(\omega), \beta^0\}_{n=1}^{\infty}$) we consider as an estimate of β^0 the value $\hat{\beta}(\mathcal{D}_n^0(\omega_0)) = \hat{\beta}(\mathcal{D}_n)$ (see (4)). But it does not guarantee automatically that the estimate $\hat{\beta}$ will be also "reasonable". The matter will be discussed below, nevertheless, please notice that even we do not know whether the point ω_0 lies in the corresponding set of high probability.

Moreover, to describe the real behavior of the estimator it may happen that we need to assume that generally for any fix n we need another system $\{\mathcal{D}_k(\omega)\}_{k=1}^{\infty}$. Let us formalize this fact. In what follows the upper index will distinguishe also among the different estimators. Let us assume two different estimators $\hat{\beta}^h(\mathcal{D}_n^{\ell}(\omega))$, $h, \ell = 1, 2$ constructed in two different regression models with generally $\mathcal{D}_n^1(\omega) \neq \mathcal{D}_n^2(\omega)$ and $\beta^1 \neq \beta^2$. Of course, we assume that the both models will be used to explain the same data, i. e. that there are ω_1 and ω_2 so that $\mathcal{D}_n^{\ell}(\omega_{\ell}) = D_n$, $\ell = 1, 2$. Having applied the estimators $\hat{\beta}^1(\mathcal{D}_n^1(\omega))$ and $\hat{\beta}^2(\mathcal{D}_n^2(\omega))$ on the data D_n we may find that the norm $\|\hat{\beta}^1(D_n) - \hat{\beta}^2(D_n)\|$ is significantly large (see [2] or [19]). At a first moment it may seem surprising because both estimators have been constructed to be asymptotically consistent (and maybe even with high breakdown point). As we have seen from the Exhibit 2 the Least Median of Squares and the Least Trimmed Squares (both being consistent and with 50 % breakdownpoint) may be orthogonal each to other. One may then conclude that probably some assumptions under which the good properties of the estimators have been proved (the assumptions some of which we principally cannot check) were distorted, and hence we have no justification to apply these methods.

It may be true but another possibility of a "strange mutual" behaviour of the estimators may be the fact that we have $\hat{\beta}^1(D_n) = \hat{\beta}^1(\mathcal{D}_n^1(\omega_1))$ while $\hat{\beta}^2(D_n) = \hat{\beta}^2(\mathcal{D}_n^2(\omega_2))$, of course with $\mathcal{D}_n^1(\omega_1) = \mathcal{D}_n^2(\omega_2) = D_n$, however with $\mathcal{D}_n^1(\omega) \neq \mathcal{D}_n^2(\omega)$ and $\beta^1 \neq \beta^2$, and of course with generally $\hat{\beta}^1(\mathcal{D}_n^1(\omega)) \neq \hat{\beta}^2(\mathcal{D}_n^2(\omega))$. Then $\hat{\beta}^1(D_n)$ may be near to β^1 while $\hat{\beta}^2(D_n)$ is near to β^2, however β^1 being possibly far from β^2. In the words, the estimators assume *implicitly* different regression models $(\mathcal{D}_n^1(\omega), \beta^1)$ and $(\mathcal{D}_n^2(\omega), \beta^2)$, respectively. It does not mean that $\hat{\beta}^1(\omega)$ or $\hat{\beta}^2(\omega)$ are not asymptotically consistent. We may quite well have $\hat{\beta}^1(\mathcal{D}_n^2(\omega)) \underset{n\to\infty}{\longrightarrow} \beta^2$ in probability, as well as $\hat{\beta}^2(\mathcal{D}_n^1(\omega)) \underset{n\to\infty}{\longrightarrow} \beta^1$ in probability, nevertheless for our data they may behave as it was described above, i. e. estimator $\hat{\beta}^1(\omega)$ may prefer to assume that the data have been generated by the model $(\mathcal{D}_n^1(\omega), \beta^1)$ while the estimator $\hat{\beta}^2(\omega)$ prefers $(\mathcal{D}_n^2(\omega), \beta^2)$. At this moment one may object that with increasing size of sample both estimators should start to assume simultaneously either $(\mathcal{D}_n^1(\omega), \beta^1)$ or $(\mathcal{D}_n^2(\omega), \beta^2)$. Firstly, such assertion is of a little help for processing a unique set of data of a fix size. Secondly, even if we accept for a while the questionable assumption that we may increase the number of observations as much as we like, even then we should be more careful. In fact we may claim (in the case of strong consistency) precisely following:

If $\hat{\beta}^h(\omega)$ is strongly consistent in the linear regression model $\left\{(\mathcal{D}_n^{\ell}(\omega), \beta^{\ell})\right\}_{n=1}^{\infty}$ then

$$\exists (B_{h,\ell} \in \mathcal{B}, P(B_{h,\ell}) = 1) \quad \forall (\tilde{\omega} \in B_{h,\ell}, \varepsilon > 0)$$

$$\exists \left(m(\tilde{\omega}, \varepsilon, \hat{\beta}^h(\omega), \{\mathcal{D}_k^{\ell}(\omega)\}_{k=1}^{\infty}) \in N\right) \forall (n \in N, n \geq m(\tilde{\omega}, \varepsilon, \hat{\beta}^h(\omega), \{\mathcal{D}_k^{\ell}(\omega)\}_{k=1}^{\infty}))$$

$$\|\hat{\beta}^h(\mathcal{D}^\ell_n(\bar{\omega})) - \beta^\ell\| < \varepsilon.$$

And only an uncritical belief in the traditional paradigma may prevent us to see that for any fix, finite size of sample it does not help too much. It is easy to imagine that for a sequence of samples $\{D_n\}_{n=1}^{\infty}$ we may have a system $\left\{\{\mathcal{D}^\ell_k(\omega)\}_{k=1}^{\infty}\right\}_{\ell\in\mathcal{L}}$ such that for any $n \in N$ there is $\omega_{n\ell}$ such that $D_n = \mathcal{D}^\ell_n(\omega_{n\ell})$. Now the estimator $\hat{\beta}^h(\omega)$ may assume implicitly for any $n \in N$ "behind the data" different system of random variables $\mathcal{D}^\ell_n(\omega)$, i. e. ℓ may be a function of h and of n, i. e. $\ell = f_h(n)$. It may be caused e. g. by small or mediate size of sample or by varying level of contamination etc. The "selection" of the system of random variables which might generate our data is of course (for many types of estimators) independent of our will or assumptions, it just reflects an "interpretation" of the data by the estimator. So we have $D_n = \mathcal{D}_n^{f_h(n)}(\omega_{nf_h(n)})$, hopefully with $\omega_{nf_h(n)} \in B_{h,f_h(n)}$ (of course we assume that $D_n = D_{n-1} \cup \{y_n, x_{n1}, x_{n2}, \dots, x_{np}\}$) and for some $\varepsilon > 0$ it may happen that

$$m(\omega_{n\,f_1(n)}, \varepsilon, \hat{\beta}^1(\omega), \left\{\mathcal{D}_k^{f_1(n)}(\omega)\right\}_{k=1}^{\infty}) < n < m(\omega_{n\,f_2(n)}, \varepsilon, \hat{\beta}^1(\omega), \left\{\mathcal{D}_k^{f_2(n)}(\omega)\right\}_{k=1}^{\infty})$$

and

$$m(\omega_{n\,f_2(n)}, \varepsilon, \hat{\beta}^2(\omega), \left\{\mathcal{D}_k^{f_2(n)}(\omega)\right\}_{k=1}^{\infty}) < n < m(\omega_{n\,f_1(n)}, \varepsilon, \hat{\beta}^2(\omega), \left\{\mathcal{D}_k^{f_1(n)}(\omega)\right\}_{k=1}^{\infty}),$$

i. e. we have for all n

$$\|\hat{\beta}^1(D_n) - \beta^1\| < \varepsilon \qquad \text{aswellas} \qquad \|\hat{\beta}^2(D_n) - \beta^2\| < \varepsilon.$$

In the case of weak consistency the situation may seem at the first moment more transparent (because of the uniformity of the convergence): If $\hat{\beta}^h(\omega)$ is weakly consistent in the linear regression model $\left\{(\mathcal{D}^\ell_n(\omega), \beta^\ell)\right\}_{n=1}^{\infty}$ then

$$\forall(\varepsilon > 0, \delta > 0) \quad \exists\,(\, m(\varepsilon, \delta, \hat{\beta}^h(\omega), \{\mathcal{D}^\ell_k(\omega)\}_{k=1}^{\infty}) \in N\,)$$

$$\forall\,\left(n \in N, n > m(\varepsilon, \delta, \hat{\beta}^h(\omega), \{\mathcal{D}^\ell_k(\omega)\}_{k=1}^{\infty})\right)$$

$$P(\left\{\omega \in \Omega : \|\hat{\beta}^h(\mathcal{D}^\ell_n(\omega)) - \beta^\ell\| > \delta\right\}) < \varepsilon,$$

but we may still have

$$m(\varepsilon, \delta, \hat{\beta}^1(\omega), \left\{\mathcal{D}_k^{f_1(n)}(\omega)\right\}_{k=1}^{\infty}) < n < m(\varepsilon, \delta, \hat{\beta}^1(\omega), \left\{\mathcal{D}_k^{f_2(n)}(\omega)\right\}_{k=1}^{\infty})$$

and

$$m(\varepsilon, \delta, \hat{\beta}^2(\omega), \left\{\mathcal{D}_k^{f_2(n)}(\omega)\right\}_{k=1}^{\infty}) < n < m(\varepsilon, \delta, \hat{\beta}^2(\omega), \left\{\mathcal{D}_k^{f_1(n)}(\omega)\right\}_{k=1}^{\infty}).$$

At this moment one may object once again. The assumption that it may happen that for different $n \in N$ the data D_n are represented by different systems $\{\mathcal{D}_n^{f_h(n)}(\omega)\}$ does not agree with our natural feeling for, or belief into, the "stationarity" of data. But it is not inevitably the problem of the stationarity of data and its reflection in a mathematical model. The problem may be that the estimator may assume implicitly rather different representations for the two samples of data, one of which is a subsample of the other. And the experiences with robust (especially with high breakdown point) estimators confirm that. On the other hand, the matter is not so much surprising, we may just imagine an ancillary parameter which is not explicitly given in the problem, which however "selects" the family of distributions which will be taken into account.

It seems clear that the described phenomenon cannot appear for maximum likelihood estimators because in this case the (type of) distribution is imposed into the model as the assumption (e. g. the least squares estimation with the implicit assumption of normality of random fluctuations). However it is true only for the estimation in a "fully" parametric model. In the regression it would be true only if we also assume one given family of distribution for the carriers. On the other hand, it was just such assumption which proved to be sometimes a weak point of the classic statistics, and it was an inspiration for developing the robust statistics. However weakening the assumption about the type of distribution of random fluctuations may cause the effect which was described above. Naturally, we hope that it is not generally the rule.

Maybe that we shall see the problem better from another description of situation by means of empirical d. f.. We shall use a little changed notation which is more usual in such framework.

Let us denote for any $\beta \in R^p$ by $F_n(t,\beta)$ the empirical d. f. of the empirical residuals $r_i(\beta) = y_i - x_i^T\beta$, $i = 1, 2, \ldots, n$, i. e.

$$F_n(t,\beta) = \frac{1}{n}\sum_{i=1}^{n} I_{\{z \in R\,:\, y_i - x_i^T\beta < z\}}(t).$$

Please do not confuse $F_n(t,\beta)$ with $F_n(\omega,t,\beta)$ which is given as

$$F_n(\omega,t,\beta) = n^{-1}\sum_{i=1}^{n} I_{\left\{\omega^* \in \Omega,\ z \in R\,:\, Y_i(\omega^*) - X_i^T(\omega^*)\beta < z\right\}}(\omega, t)$$

and which is sometimes also called the empirical distribution function. Of course, $F_n(t,\beta) = F_n(\omega_0,t,\beta)$, for some ω_0, however the difference is that $F_n(t,\beta)$ is known while $F_n(\omega,t,\beta)$ is unknown except at the point ω_0. Now, considering $\beta^* \neq \beta^{**} \in R^p$, we may find absolutely continuous d. f. F_* and F_{**} so that using e. g. the Prokhorov metric ν we obtain for some small positive δ

$$\nu(F_n(t,\beta^*),\ F_*) = \nu(F_n(t,\beta^{**}),\ F_{**}) \leq \delta.$$

Then we may claim that our data D_n were generated by a model

$$Y_i^* = X_i^{*T}\beta^* + e_i^* \quad i = 1, 2, \ldots$$

with $\{e_i^*\}_{i=1}^{\infty}$ being a sequence of independent random variables with d. f. F_*. But the same holds for β^{**} and F_{**}. Then we cannot be surprised that one estimator may give the estimate $\hat{\beta}^*$ near to β^* while other gives $\hat{\beta}^{**}$ which is not far from β^{**}. And it is easy to imagine that F_* and F_{**} can be selected so that even the values of likelihoods (or values of some other statistics) at our data D_n (more precisely at residuals $r_i(\beta^*)$ and $r_i(\beta^{**})$, $i = 1, 2, \ldots, n$) can be the same. And it is clear that the size of sample did not play any role in the explanation via empirical d. f. being valid for data of any fix size.

So (returning to the earlier full notation) we may guess that for example it may be interesting to study conditions under which the estimator assumes implicitly for the all n the same (or nearly the same) sequence $\{\mathcal{D}_n(\omega)\}_{n=1}^{\infty}$ nevertheless under which it does not suffer by low robustness (as the maximum likelihood estimators frequently do).

Maybe that the minimal distance estimators may seem to be promising solution of the problem. Nevertheless, the technical difficulties which we encounter when trying to find an absolutely continuous distribution which is the nearest one to the empirical distribution (of residuals), force us to take some measures which may at the end give the estimators which suffer by the same disadvantage. Remember that in many cases the minimal distance methods yield the estimators with the redescending influence function (see [22]), and it is known that such estimators exhibit also the property of "diversity of the estimates", see [25].

IDEAS OF SELECTION OF APPROPRIATE ESTIMATE

There are at least two straightforward ideas which have been studied up to now and which may help to cope with (not to solve) the problem of diversity of the estimates. Both of them may be used for any type of estimator, however for one of them we have at our disposal applicable results only for the M-estimators.

The first idea might be called the *subsample stability.* The idea is simple. If our estimator is "appropriate" for the given data, and is able to "recognize the true model" then it should be able to do the same for the (reasonable) subsample of data. In other words, the estimates for the full data and for a subsample should be similar. Let us denote by $\hat{\beta}^{(n,I_{k_n})}$ the estimate which we obtain for data (y_i, x_i), $i \in \{1, 2, \ldots, n\} - I_{k_n}$, where $I_{k_n} = \{i_1, i_2, \ldots, i_{k_n}\}$ with $1 \leq i_i < i_2 < \cdots < i_{k_n} \leq n$. Then the idea of subsample stability says that if we find that the norm $\|\hat{\beta}^{(n,I_{k_n})} - \hat{\beta}^{(n)}\|$ is significantly large we should reject $\hat{\beta}^{(n)}$ as the appropriate estimate for our data. Under some regularity conditions on the criterial function ψ (of the M-estimator) an asymptotic formula for $\hat{\beta}^{(n,I_{k_n})}(\omega) - \hat{\beta}^{(n)}(\omega)$ is known for the case when either k_n is fix, either $\lim_{n\to} k_n \cdot n^{-\tau} = \lambda$ (for some $\tau \in (0,1)$ and $\lambda \in R$) or $\lim k_n \cdot n^{-1} = \lambda$ (for some $\lambda \in \left(0, \frac{1}{2}\right]$). The simplest one is the formula for the second case (the other ones may be found in [26]):

$$n\, k_n^{-\frac{1}{2}} \left(\hat{\beta}^{(n,I_{k_n})}(\omega) - \hat{\beta}^{(n)}(\omega)\right) = k_n^{-\frac{1}{2}} \gamma^{-1} Q^{-1} \sum_{i \in I_{k_n}} X_i^{\mathrm{T}} \psi(e_i(\omega)\, \sigma^{-1}) + o_p(1)$$

where $\gamma = \sigma^{-1} E\, \psi'(e_1\, \sigma^{-1}) + \sum_{k=1}^{s} f(r_k \sigma)\, [\psi(r_k+) - \psi(r_k-)]$, $r_1, r_2, \ldots, r_s$ are the points of jumps of the function ψ and $Q = \lim_{n\to\infty} \frac{1}{n}[X^{(n)}]^{\mathrm{T}} X^{(n)}$ ($X^{(n)}$ - design matrix). Using the Central Limit Theorem we may then find an approximations for the critical region for $n\, k^{-\frac{1}{2}} \left(\hat{\beta}^{(n,i_{k_n})}(\omega) - \hat{\beta}^{(n)}(\omega)\right)$. (Similar problems have been studied in [24]).

The second idea is also simple. If the estimate is appropriate for given data when the "empirical density" of the residuals in two complementary subsamples of data should be similar. It might be called *distributional homogeneity.* Let us put

$$H_n(Y, \beta) = n \int_{-a_n}^{a_n} \left[f_{[\frac{n}{2}]}^{\frac{1}{2}}(z, Y^{(1)}, \beta) - f_{[\frac{n}{2}]}^{\frac{1}{2}}(z, Y^{(2)}, \beta)\right]^2 f_n(y, Y, \beta)\, \mathrm{d}z$$

where $Y^{(1)} = \left(Y_1, \ldots, Y_{[\frac{n}{2}]}\right)$, $Y^{(2)} = \left(Y_{[\frac{n}{2}]+1}, \ldots, Y_n\right)$, $Y = (Y_1, Y_2, \ldots, Y_n)$, $\{a_n\}_{n=1}^{\infty} \nearrow \infty$ and $f_{[\frac{n}{2}]}(z, Y^{(1)}, \beta)$ is the kernel estimate

$$f_{[\frac{n}{2}]}(z, Y^{(1)}, \beta) = \frac{2}{n\, c_n} \sum_{i=1}^{[n/2]} w\left(c_n^{-1}[z - (Y_i - X_i^{\mathrm{T}} \beta)]\right)$$

etc. Then, again under some regularity conditions, the asymptotic distribution of the statistic

$$\mathcal{H}_n(Y,\beta^0) = \Delta_n^{-1}\{H_n(Y,\beta^0) - m_n\}$$

is $N(0,1)$ where

$$m_n = \frac{1}{2}c_n^{-1}\int_{-\infty}^{\infty} w^2(t)\,\mathrm{d}t$$

and

$$\Delta_n^2 = \frac{1}{2}c_n^{-1}\int_{-\infty}^{\infty} f^2(t)\,\mathrm{d}t\int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\infty} w(s)\,w(s+z)\,\mathrm{d}s\right\}^2\mathrm{d}z$$

where f is the density of the distribution function F of e_1.

It is clear that the statistic $H_n(Y,\beta)$ is the weighted Hellinger distance of the corresponding kernel estimates of the residual density. Of course, it is not very simple to compute its value, however using e. g. Romberg algorithm and some not very slow computer it is applicable for small and moderate samples. Details and a numerical example may be found e. g. in [20].

CONCLUSIONS

The discussion which was given above may tempt us to say that the (high) robustness does not guarantee anything. But it would be false conclusion.

First of all, it guarantee that the method does not suffer by the (large) sensitivity to the small changes of the distribution of the errors in the regression model. On the other hand, taking into account the examples given above we have to admit that it does not guarantee automatically anything. *But after all, if somebody would claim that some method of processing data is successfully applicable on any data, would we trust into such assertion?* (However, the belief into the miracles has probably the deeper roots than one may expect, see [21], p. 111.) It is clear from the presented examples that we have essentially two possibilities:

At first, to evaluate an estimator of the regression model and then we may test whether this estimate is appropriate for given data. Of course, some our ideas about the distribution of the residuals are anycase included in such testing, i. e. we have to express our *naturally subjective* belief that data are probably distributed so and so. One may object that the belief into the type of the data distribution may be *objectively* confirmed by some test which does not require any a priori assumptions. Unfortunately such idea is only an illusion, see [16].

Secondly, we may try to incorporate the idea of accommodation of the estimate to the data directly into the estimating method, i. e. to construct such methods which have built-in e. g. the idea of subsample stability. Similarly we could define a weighted-Hellinger-type minimal distance estimator. Leaving aside that the practical applicability of such estimator is problematic it may suffer by some other disadvantages (e. g. it is not asymptotically efficient).

Similarly, we may try to estimate, using some preliminary estimate, the density of residuals and then to find the maximum likelihood estimates of the regression coefficients. Such estimator has been studied in [23] and it proved to be theoretically plausible (asymptotically normal and efficient). Of course, to apply it, may be a terrible task (leaving aside that we may expect a good performance only for the large samples).

It may seem that in two latter cases, due to the adaption, we have reached the *objectivity* in the sense that the result is independent from any subject. Of course, it is again only the illusion.

So we may conclude: When estimating regression model we should apply several methods. In the case that the results are not significantly different ([19]) we may select one of them, probably according to some a posterior criterium which may (or even should) be inspired by the expected interpretation of the model which anyway incorporates some our opinions, some steps of belief, etc. The criterium may, of course, utilize also the ideas related to the *subsample stability*, to the *distributional homogeneity*, etc.

In the opposite case the possibilities explained above may give some hint which model may be acceptable for given data. None of them however can decide alone without a discussion with an expert in the region which the data came from. Any case, results of analysis of such data should not be the basis for a decision with possibly considerably important consequences.

References

[1] Boček, P., Lachout, P. (1994): Linear programming approach to LMS-estimation. To appear in *Memorial volume of Comput. Statist. & Data Analysis devoted to T. Havránek.*

[2] Boček, P., Víšek, J. Á. (1994): On one-step M-estimates. Submitted to *Transactions of the Fifth Prague Symposium on Asymptotic Statistics.*

[3] Daniel, C. (1976): *Applications of Statistics to Industrial Experimentation.* New York: J.Wiley & Sons.

[4] Daniel, C., Wood, F. S. (1980). *Fitting Equations to Data.* New York: J.Wiley & Sons.

[5] Donoho, D.L., Huber, P.J.(1983): The notion of breakdown point. *Festschrift for Erich L. Lehmann. Eds. P.J. Bickel, K.A. Doksum and J.L. Lehmann Jr.. Wadsworth, Belmont, California, pp. 157 - 184.*

[6] Hampel, F. R. (1974): The influence curve and its role in robust estimation. *Journal of the American Statistical Association 69, No. 364, 383–393.*

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

[8] Hettmansperger, T. P., Sheather, S. J. (1992): A Cautionary Note on the Method of Least Median Squares. *The American Statistician 46, 79–83.*

[9] Huber, P. J. (1964): Robust estimation of a location parameter. *Ann. Math. Statist. 35, pp. 73–101.*

[10] : Joss, J., Marazzi, A. (1990): Probabilistic algorithms for LMS regression. *Computational Statistics & Data Analysis 9, 123–134.*

[11] : Kovanic, P. (1986): A new theoretical and algorithmical basis for estimation, identification and control. *Automatica, 22, 657 - 674.*

[12] : Kovanic, P. (1993): Gnostical model of economics. *Proceedings of the 5-th Moravo-Silesian Symposium on Modelling and Simulation of Systems, House of Technology of Ostrava, Czech and Slovak Simulation Society and SCS, Dept. of Computer Science of Techn. Univ. Ostrava, eds. J. Štefan, 97 – 102.*

[13] Martin, R.D., Yohai, V.J., Zamar, R.H. (1989): Min-max bias robust regression. *Ann Statist. 17, 1608 - 1630.*

[14] Mason, R. L., Gunst, R. F., Hess, J. L. (1989): *Statistical Design and Analysis of Experiments.* New York: J.Wiley & Sons.

[15] Novovičová, J. (1990): M-estimators and gnostical estimators for identification of regression model. *Automatica 26, May, 1990.*

[16] Prigogine, I., Stengers, I. (1977): La Nouvelle Alliance. *SCIENTIA, 1977, issues 5-12.*

[17] Rousseeuw, J. P. (1994): Unconventional features of positive-breakdown estimation. To appear in *Statist. Probab. Letters.*

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

[19] Rubio, A., Aguilar, L., Víšek, J. Á. (1992): Testing for difference between models. *Computational Statistics 8, 57 - 70.*

[20] Rubio, A. M., Víšek, J.Á. (1993 a): Diagnostics of regression model: Test of goodness-of-fit. Submitted to the *Transactions of the Fifth Prague Symposium on Asymptotic Statistics.*

[21] : Štefan, J. (1993): Contribution to the evaluation of small data files. *Proceedings of the 5-th Moravo-Silesian Symposium on Modelling and Simulation of Systems, House of Technology of Ostrava, Czech & Slovak Simulation Society and SCS, Dept. of Computer Science of Techn. Univ. Ostrava, eds. J. Štefan, 97 – 102.*

[22] Vajda, I. (1984): Asymptotic efficiency and robustness of D-estimators. *Kybernetika, Volume 20 (1984), No. 5, pp. 358–375.*

[23] Víšek, J. Á. (1992 e): Adaptive maximum-likelihood-like estimation in linear model. Part I. Consistency. *Kybernetika 28 (1992), 357–382* Part II. Asymptotic normality. *Kybernetika 28 (1992), 454–471.*

[24] Víšek, J. Á. (1992 a): Stability of regression models estimates with respect to subsamples. *Computational Statistics 7 (1992), 183 - 203.*

[25] Víšek, J. Á. (1994 a): A cautionary note on the method of Least Median of Squares reconsidered. Submitted to *Computational Statistics.*

[26] Víšek, J. Á. (1994 b): Scale invariant sensitivity analysis of non-linear regression estimates. Submitted to *Annals of Statistical Mathematics.*

Department of Stochastic Informatics, Institute of Information Theory and Automation, Academy of Sciences the Czech Republic, Pod vodárenskou věží 4, 182 08 Prague, Czech Republic, Phone: 42-2-6605-2019 (or 2466) Fax: 42-2-0774-452, e-mail: VISEK@CSPGAS11.BITNET