An introduction to the method of the robust regression of Rousseeuw

Tsvetan B. Georgiev
Institute of Astronomy and Rozhen NAO, Bulgarian Academy of Sciences
tegorg@astro.bas.bg
(Lecture at the Astronomical Seminar in Sofia on 17.10.2007.
Accepted on 12.12.2007)

Abstract. The method of the robust regression of Rousseeuw (1984), based on the method of the least sum of the trimmed squares of deviations (LTS), is described. The ordinary method of least squares (OLS) measures the scatter of the deviations summing all N squares of deviations. The LTS sums only the left half of the ordered squares of the deviations, concerning at least N/2+1 data. Large deviations may be presented in the right part of the order, but LTS ignores them. Thus LTS can withstand asymptotically 50% fraction of large deviating data and still remain robust. From this point of view the OLS method has zero robustness. The robust regression method tests and qualifies through LTS each available pattern of possible solution: each point in 1D case, each line through pair of points in 2D case, each plane through triad of points in 3D case, etc. The pattern with the shortest LTS scatter is considered as 1st approximation of the solution. Usually the 2nd approximation is carried out by the OLS, after removal of the outliers in respect to the 1st approximation. In the present work seven types of utilizable deviations of a point from the line pattern are tested in the 2D case: vertical, horizontal, orthogonal, quadratic mean (diagonal), geometric mean, harmonic mean and arithmetic mean. It is shown that the use of geometrical mean deviation leads to well reproducing of the bisector of the direct and reverse regression in both ordinary and robust methods. However, in the general case, when the standard statistical assumptions are violated strongly, and (or) when the regression slope A is $A \gg 1$ or $A \ll 1$, the orthogonal deviations are the natural tool for both regressions. In the multidimensional cases, depending of the type of used deviations, the final application of the OLS may be difficult or impossible. Therefore in the present work the robust regression method is complemented with a 2nd approximation in the spirit of the robust regression too. It is based on extraction of several (up to 10-20) best patterns of solutions (points in 1D case, pairs of points in 2D case, triads of points in 2D case) and creation of additional intermediate patterns (middle of points, bisectors of lines and planes). These additional patterns are tested again by the LTS in a search for a better solution. Thus the inner accuracy of the method may increase 3-5 times. An empirical approach to the slope error estimation by means of the cumulative error function, built on the trimmed best patterns of the solution, is also introduced. The robust regression is applied on examples for 1D mode estimations and 2D regressions without or with intercepts. Main sequence fitting on color-magnitude diagram with $14 - 19\%$ fraction of strongly deviated stars is demonstrated.
Key words: data analysis - methods, statistical - methods; PACS 95.75.z, 95.75.Pq

Въведение към метода на устойчивата регресия на Русю

Цветан Б. Георгиев

Описана е устойчивата (робастата) регресия на Русю (1984), базирана на метода на най-малката сума на отбраните квадрати на отклоненията (МОК). Ординарният (обичайният) метод на най-малките квадрати (МНК) измерва разброса на отклоненията сумирайки всичките N квадрати на отклоненията. МОК сумира само лявата половина на подреденитае квадрати на отклоненията, обхващайки поне N/2+1 броя. Големи отклонения могат да присъстват в дясната част на редицата, но МОК ги игнорира. Така МОК понеся асимптотично 50% фракция от силни отклонения, запазвайки устойчивостта си. От тази гледна точка МНК има нулева устойчивост. Методът на устойчивата регресия проверява и квалифицира чрез МОК всеки достъпен образец на възможно решение - всяка точка в едномерния случай, всяка права линия през двойка точки в двумерния случай, всяка равнина през тройка точки в тримерния случай и т.н. Образецът с най-малък МОК разброс се извлича като първо приближение на решението. Обикновено

второто приближение се прави чрез МНК след отстраняване на големите отклонения спрямо първото приближение. В тази работа при двумерния случай са тествани седем вида използваеми отклонения на точки от линията-образец - вертикално, хоризонтално, ортогонално, средноквадратично (диагонално), средногеометрично, среднохармонично и средноаритметично. Показано е, че използването на средногеометричното отклонение води до добро възпроизвеждане на ъглополовящата на правата и обратна регресия,както при обикновения, така и при устойчивия метод. Обаче, в общия случай, когато стандартните статистически изисквания са силно нарушени и/или когато регресионният наклон е $A \gg 1$ или $A \ll 1$, ортогоналните отклонения са естественото средство и за двата вида регресии. В многомерния случай, в зависимост от използваното отклонение, финалното прилагане на МНК може да е трудно или невъзможно. Затова в тази работа методът на устойчивата регресия е допълнен с второ приближение, в духа на метода на устойчивата регресия. То се базира на извличане на няколко (до 10-20) най-добри образци на решения (точки в едномерния случай, линии в двумерния, равнини в тримерния) и създаване на допълнителни промеждутъчни обраци (средини на точки, ъглополовящи на линии или равнини). Тези допълнителни образци се тестват отново чрез МОК в търсене на по-добро решение. Така вътрешната точност на метода може да се повищи 3-5 пъти. Предложен е и емперичен подход към оценяване на грешката на регресионния наклон чрез кумулативна функция на грешката, построена по отбраните най-добри образци на решението. Методът на устойчивата регресия е приложен за оценки на модите на едномерни разпределения и за двумерни регресия без и със свободни членове. Демонстрирано е фитиране на главната последователност на диаграма цвят-величина с около $14 - 19\%$ фракция от силно отклоняващи се звезди.

Introduction

The ordinary least square (OLS) regression is used widely in astronomy, biology, economics etc. It is commonly known that the OLS (Y—X) line is the best when important assumptions hold: (0) the true relation between the variables is linear; (1) the values of the independent variable (X) are measured without errors; (2) the observed values of the dependent variable (Y) are subject to errors with mutual independence, zero mean and finite variance (common for all observations, i.e. case of homoscedastic errors); (3) the X and Y data do not have intrinsic scatter. The standard OLS analysis is not strict when any of the accounted assumptions is not filled (see also Isobe et al. 1990 (hereafter IFAB90), Feigelson & Babu 1992 (hereafter FB92), Branham 2001, Kelly 2007).

The OLS conditions are frequently violated in astronomy, where the presence of heteroscedastic errors (observations with different individual errors) and intrinsic scatter both in X and Y data is usual. The choice of the independent variable frequently is not clear, too. Good examples are the dependences „mass - luminosity“ for stars or galaxies. They have intrinsic scatters in both X and Y data, caused by unaccounted factors (ages, metallicities, hidden masses). Other interesting example is the Hubble diagram, where observation errors of the redshifts and magnitudes of the galaxies are negligible in respect to the errors due to the uncertainties in the distances, the corrections for non-Hubble motions, the intrinsic scatter of the luminosities of the galaxies, etc. (IFAB90). In such cases different kinds of deviation of the observation point from the fitted line may be used (Sec.1,2). More difficult case is, for example, the fit of the main sequence on a color - magnitude diagram. The reason is the significant fraction of outlier points around the main sequence: background stars, evolved

stars etc. This problem is solved remarkably by the robust regression (RR), described here (Fig.7,9).

Since the OLS methods found in the standard statistics textbooks and respective software are not always satisfactory, the astronomers are forced to search for better tools. Regression methods in astronomy and different improvements of the OLS for the cases, when conditions (1) - (3) are not fulfilled, have been subjects of multitude papers (see also Akritas & Bershady 1996, Tremaine et al. 2002, Kaspi et al. 2005 and references therein). However, the results, called sometimes „robust methods“, do not include the cases when numerous strong outliers exist among the data.

Commonly, the OLS (including its improvements) is considered as perfect tool in presence of normal or almost normal errors in Y variable, but it is worthless in the presence of strongly outlying data. The latter fact is extremely important when the large errors exist in the independent (X) variable (Fig.5,7,9). Generally, the problem is also very serious (i) when the number of outliers is large, (ii) in the multi dimensional (MD) cases, when the visual control of outliers is almost impossible, (iii) in the image processing, when the program code must work surely and fast, etc. In such cases RR methods are necessary.

RR methods with superior performance over OLS in many situations exist, but they are not yet widely used. The possible reasons may be: (i) difficulties in the programming and long computing time; (ii) bad choice of the method in the first attempt of applying (e.g. choice of method which is not really robust), (iii) the belief that the classical methods have natural robustness; (iv) the absence of robust methods in many popular software packages (see also Wikipedia, 2007).

This work emphasizes on the robust non-OLS regression method based on robust estimator of the scatter of 1D random variable. It is introduced by Peter Rousseeuw (1984) and is called „least trimmed squares“ (LTS) estimator (Sec.3). This method can ignore numerous outliers, asymptotically up to 50% of all data, like the 1D median estimator. The LTS estimator recognizes correctly the mode of 1D random distribution, as well as the location (data center) in MD cases (Fig.6). In the regression applications LTS places the best line in 2D case (or the best pale in 3D case, etc.) among most populated ellipsoid of data distribution (Fig.7,9).

The RR, based on the LTS estimator, is widely discussed in the monograph of Rousseeuw & Leroy (1987, hereafter RL87). Two ideas build the base of the RR method. The first one is the examination of each available pattern of possible solution (single point in 1D case, line through pairs of points in 2D case, plane through triads of points in 3D case, etc) and revealing the superior pattern. The second idea is the examination of the patterns by extremely robust estimation of the scatter (as 1D system of deviations, Sec.2) through the LTS estimator (Sec.3,4).

The LTS is efficient and relatively simple robust estimator. More sophistical developments in this field can be found in the papers of Rousseeuw & Yohai (1984), Yohai (1987) and Rousseeuw & Van Driesen (1999). Contemporary information about robust estimators and regressions can be found also in Internet through the query „robust regression“ (Fox 2002, Chen 2007, Olive 2007).

The goal of this paper is to be an introduction into the basic ideas and recommendable solutions in the field of the extremely non-OLS methods, mainly about these based on the LTS estimator. Fortunately, the description and the illustration of the possibilities of the RR method may be carried out full enough in 1D and 2D cases, while the MD generalization is natural and simple.

Section 1 presents the work of seven methods for measurement of the deviation of a point from a line and underlines the potential of the RR over the OLS regression. Section 2 emphasizes on the superiority of the orthogonal deviations in the general case of the OLS and RRs, when the assumptions of the OLS method are strongly violated. Section 3 gives general description of the robust method and introduces two additions to it: 2nd approximation in the spirit of the RR and an empirical approach to coefficients error estimation. Section 4 presents comparison of five 1D estimators with account of the robustness parameter „breakdown value". Section 5 presents a few applications of the methods in 1D and 2D cases.

1 Seven types of deviations form the tested line and the potential of the RR in the MD case

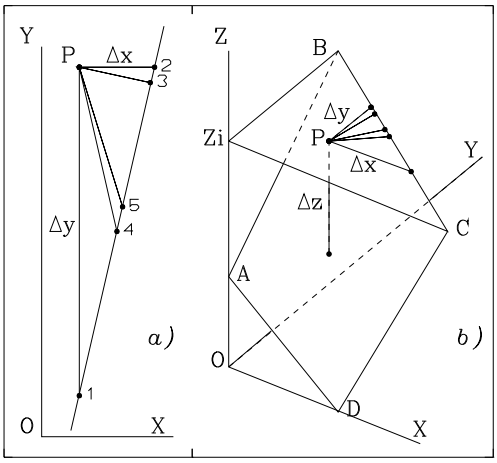
Since the problem is the placing of the regression solution (location, or line, or plane, etc) in the most populated region of data, the measuring of the deviations becomes important component of the method. Therefore, we must regard this problem before the problems of the RR.

In the particular case when the location (or the center of the distribution) is searched in 1D, 2D, etc. cases, the simple Euclidian distance is unique and enough (Fig.6). However, in the general 2D regression case, and more in the MD applications, many possibilities exist. We did not find recommendations in this field for the case of RR and were forced to study the problem.

Seven types of deviations in the 2D case were tested for the present work. Each of them may be used for deriving explicit OLS formulas for 2D regression coefficients, for regression error etc. (IFAB 90, FB92, Sec. 2). However, any MD OLS regression based on deviations that are not solely Δy or Δx require solving complicated systems with nonlinear equations. Contrariwise, the RR method involves only deriving the formula for computing the chosen type of deviation from the line, plane, hyperplane, etc. Further the scatter of all computed deviations of the chosen type will be estimated by means of the (1D) LTS method.

Figure 1a presents five types of deviations of a data point P from a given (or fitted) line in the 2D case. Let us first emphasize on the MD potential of the RR and then return to the important details about the use of the different deviations further.

Figure 1b shows a possible 3D implementation of 5 deviations, presented in Fig.1a. This may be useful for example in the problem for deriving the fundamental plane of elliptical or spiral galaxies. Let consider the mass parameter of the galaxy (velocity dispersion or HI line width, respectively) associated with Z-axis and let assume that it is known with high accuracy. Let consider also that the galaxy luminosity (e.g. X-axis) and galaxy size (e.g. Y-axis) are known with significant errors. This is a non-OLS situation, because the „independent" variables (X and Y) are subjected to errors, but the „dependent" variable (Z) is considered free of errors. However, applying the RR method we may use obligatory the plane BCZi, parallel to the plane OXY and passing



Фиг. 1. (a) Five methods for measuring the deviation of the current point P from the checked line: 1 - vertical, 2 - horizontal, 3 - orthogonal, 4 - diagonal and 5 - geometric mean (the bisectoral (6) and average (7) deviations are not shown); (b) Possible application of different deviations in a 3D non-OLS situation where X and Y data have significant errors, but Z data are known with high accuracy (see the text).

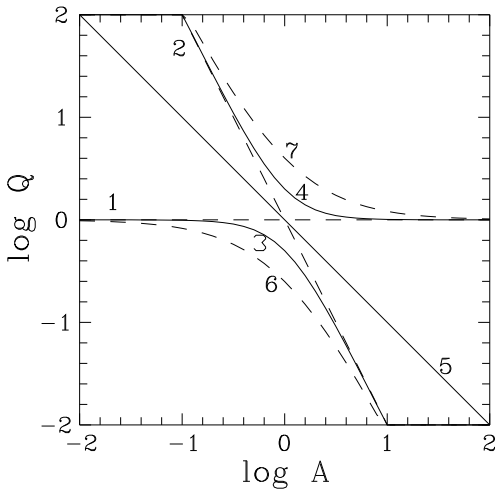
through the current point P (Fig.1b). Since the error of Z data is negligible, the deviation R from the examined plane ABCD must be measured as deviation from the line BC, in the plane BCZi. Thus any type of deviation, as in Fig.1a., may be used. Generally, different types of deviations including Δz may be used easy for RR in the 3D case, depending on the additional information about the errors in X, Y and Z directions.

However, the concrete value of slope of the fitted line depends on the used physical units of X and Y data.

Let turn back to Fig.1a and define seven kind of deviations. Fig.1a shows a XY plot of a line $Y = AX + B$ (with slope $A = 4$), one current point P and five segments that illustrate five methods for measuring of the deviations of this point from the line. The vertical distance Δy (1) and the horizontal distance Δx (2) are attributes of the direct (Y—X) and reverse (X—Y) OLS regressions. Let present the deviations by two methods: through Δy and Δx, as well as through Δy only, in the form $R = Q.\Delta y$, where Q is a coefficient depending on A. Therefore, we have $R = \Delta y$ with $Q = 1$ for (1), and $R = \Delta x = \Delta y/A$ with $Q = 1/A$ for (2). The orthogonal deviation (3) (introduced by the statistician K. Pearson in 1901) is measuring vertically to the line, i.e. $R = \Delta x \Delta y/(\Delta x^2 + \Delta y^2)^{1/2}$ with $Q = 1/(A^2 + 1)^{1/2}$. The diagonal deviation (4), proportional to quadratic mean, is $R = 0.5(\Delta x^2 + \Delta y^2)^{1/2}$ with $Q = 0.5(1/|A| + 1)^{1/2}$. The geometric mean (5), $R = (|\Delta x|.|\Delta y|)^{1/2}$, with $Q = 1/(|A|)^{1/2}$, is introduced by the astronomer G. Stromberg in 1940 as „impartial deviation“, and independently by the statisticians Kermasck & Haldane in 1950 as „reduced major-axis deviation“ (see IFAB90). The harmonic mean (6) ($1/R = 1/\Delta x + 1/\Delta y$, not shown in Fig.1a), that is the length of the bisector of the segments Δx and Δy, is $R = (\Delta x + \Delta y)/(\Delta x \Delta y)$ with $Q = 1/(|A| + 1)$.

The arithmetic mean (7) or the average (also not shown in Fig.1a), is $R = 0.5(\Delta x + \Delta y)$ with $Q = 0.5/(|A| + 1)$.

In the case of the 2D RR, when the assumptions of the OLS are strongly violated, every chosen type of deviation among (1) - (7) can be applied to measure the goodness of the tested pattern of line. Therefore, the dependence of the used deviation R on the slope A of the tested line is of interest. It is clear that when the slope of the line is close to unit, the deviations (3) - (7) tend to coincide. However, when the slope of the line is $A \gg 1$ (or $A \ll 1$) the deviations (3) and (6) tend to be proportional to Δx (or Δy) and the deviations (4) and (7) tend to be proportional Δy (or Δx).



Фиг. 2. Behavior of the deviations in respect to Δy, expressed by a Q coefficient, when the slope A of the tested line changes. The ordinary deviations Δy with $Q = 1$, and $\Delta x = \Delta y/A$ with $Q = 1/A$ are presented by dashed lines. The geometric mean deviation $R = (|\Delta x|.|\Delta y|)^{1/2} = \Delta y/|A|^{1/2}$ with $Q = 1/|A|^{1/2}$ is presented by the a solid diagonal line. The behaviors of other types of deviation, presented by solid or dashed curves, are noted in the plot.

These effects are illustrated in Fig.2 as dependences of the coefficients Q (defined above) on the slope A of the tested line. The Q coefficients for the deviation (3), (4), (6) and (7) have nonlinear behavior, i.e. the measuring method works non-equally with different slopes A. From this point of view Fig. 2 shows that in the case of the geometric mean (5) (as well as in the simplest cases (1) and (2)) the measuring method seems to work equally with different slopes A. So, the geometric mean (5) seems to play a special role among the deviations 3-7 and may be preferable for the RR.

Looking on the examples of IFAB90 and FB92 we can add also, that the OLS regression line, based on the geometric mean deviation (5), practically coincides with the bisector of (Y—X) and (X—Y). The experience from the preparation of the presented here examples supports strongly this conclusion for the practice of the RRs, too. Therefore, since the RR method is time con-

suming, but following the recommendations of IFAB90 and FB92 we seek the bisector line, we can fit the bisector well using the geometrical mean deviation (5). Note that in the MD case with $P > 2$ dimensions the concept of the bisector is not clear, but the geometric mean value has reasonable generalization: $R = (\Delta x_1 \Delta x_2 . . . . \Delta x_P)^{1/P}$.

From the RR point of view additional details about the deviations must be elucidated. Let note that using the deviations (3) - (7) the OLS regression does not change when X changes Y. Also, in the case (1), (2) or (5) the deviation and the regression slope depend equivariantly (reasonably) on linear transform of the X or/and Y data. However, in the cases (3), (4), (6) or (7) this quality of the regression is absent, because the heel of the deviation onto the tested line changes non equivariantly in respect to the surrounding data points. From this point of view the deviation (5) seems to be preferable again.

Note also, that the different deviations may be used in the RR for examination of available lines under different limitations. The deviation Δy (1) is not defined in the case of vertical line pattern and Δx (2) - in the case of horizontal line pattern. The deviations (4), (5) and (7) are not defined in both cases. However, the orthogonal (3) and bisectoral (6) deviations have limiting values R in both cases: for (3) it is the distance R to the vertical or horizontal line and for (6) it is the same distance, but multiplied by $2^{1/2}$.

Figure 2 suggests that the geometric mean seems to be the preferable deviation. However, in the next section we will show that when the assumptions of the OLS method are violated the orthogonal deviation must be preferable both for OLS and RRs.

2 What should be the preferable deviation in the astronomical OLS and robust regressions?

The considerations in Sec.1 forced us to turn back toward the OLS regression in 2D case. It is known that the direct and reverse OLS regressions have different slopes and the reverse regression is always steeper. We will look on the influence of the configuration of the data points on the bias of the OLS regression. Let us concentrate on the widely known and many times cited recommendations of IFAB90 and FB92.

The authors of the papers IFAB90 and FB92 investigated and commented the OLS use of deviations (1), (2), (3), (5), the bisector of the OLS regressions (Y—X) and (X—Y), and the arithmetic mean of the slopes of the OLS regressions (Y—X) and (X—Y). The last two types of deviations are not considered in this work because they can not be expressed explicitly, i.e. they are useless for the RR. Important related problems in the 2D case, such as formulas for estimations of the slope, the slope error and regression error are solved and discussed by IFAB90 and FB92 too.

We will regard two conclusions of these authors about the 2D regression, as follows. (i) The bisector of the OLS regressions (Y—X) and (X—Y) is recommended as the superior solution in the complicated cases; (ii) The orthogonal regression, based on the deviation (3) is not recommended, because (ii-a) in the numerical simulations its behavior is unstable and (ii-b) it is significantly biased from the bisector. However, is the bisector a naturally good idea in the 2D case? Why cannot the orthogonal regression be recommended? Note that the formula for the slope of the orthogonal regression coincides with the

expression giving the slope of the major axis of the ellipsoid of the data. This expression is introduced from mechanics into statistics through the method of the moments by K. Pearson. The key solving the problem is the influence of the configuration of the data points onto the bias of the regression slope, as can be seen in Fig.3.

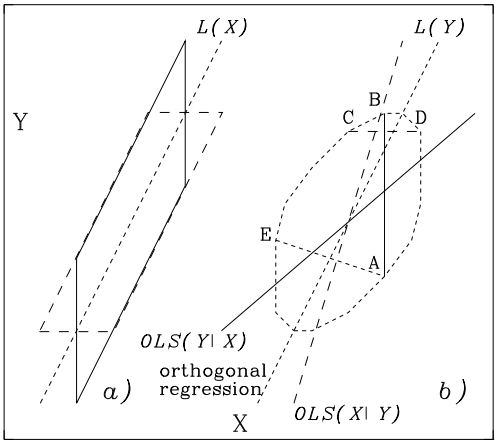
Remember that when we approximate an elliptical 2D distribution of the experimental points within expected regression slope $A \gg 1$ the direct regression was biased to the horizontal line strongly and the reverse regression was biased to the vertical line weakly. When $A \ll 1$ the situation reverses. Can both OLS regressions coincide? The answer is no. Even in the case $A = 1$ both regressions will have smallest biases (with asymptotically equal absolute values and opposite directions).

Figure 3a shows an expected line regression $Y = L(X)$ with slope $A = 2$ and two rhomboids around it. If the observational data have „good“ distribution in the solid line rhomboid (including sharp vertical bounds and enough populated upper and down corners) the direct OLS regression will reproduce $L(X)$ well. In such a situation the orthogonal OLS regression will be biased toward the vertical line (see below). However, if the corners are empty, the direct regression will be biased (stronger with increasing of A) toward the horizontal line. For reproducing the dependence $L(X)$ by the reverse OLS regression, the dashed line rhomboid must be „well“ populated, including its corners. In such a situation the orthogonal OLS regression will be biased toward the horizontal line (see below). If the corners of the dashed line rhomboid are empty, the reverse OLS regression will be biased (weaker with increasing of A) toward the vertical line. It is also clear that when all 4 corners are populated, or when all 4 corners are not populated, the direct and reverse OLS regression will be biased unavoidably, and in different ways, depending on A. Because of this the bisector of the direct and reverse regression will be biased in different way too. Therefore, the bisector regression could be recommendable only in the rare case of $A = 1$. Obviously, in the general case the bisector of the direct and reverse OLS regressions is not a good idea. - protivno na zakl. v prednija section

Figure 3b shows that in the general case, when the above mentioned „corners“ in the point distribution are empty, the orthogonal regression seems to be superior. Let consider this problem from a slightly different point of view.

The most frequent situation is approximately elliptical distribution of the data points, as it is shown in Fig.3b, with some „main sequence“. It is natural to suppose that the major axis of the ellipse must represent the expected dependence $L(X)$ in the best way. Note that the edge population of the ellipse applies strong and different leverage action on the direct and reverse OLS regression lines. The larger part of the segment AB lies below the line $L(X)$ and pulls the right edge of the direct regression (solid line) down toward the horizontal line. The opposite (down-left) part of the ellipse pulls up the left edge of the direct regression. Simultaneously, since $A > 1$ the reverse regression is less rotated. Not so large part of the segment CD lies to the left from $L(X)$ and pulls weakly the reverse regression toward the vertical line. In case of $A < 1$ the situation changes. The direct regression will be biased weakly, but the inverse regression will be biased strongly. It is clear again, that generally the bisector of the direct and reverse lines is not a good idea.

From this point of view it is obvious that the major axis of the elliptical distribution will be reproduced in the best way by regression, based on the



Фиг. 3. On the reproducing of an expected dependence $Y = L(X)$ with $A = 2$ and different distribution of the observation points. a) The solid line rhomboid presents the distribution of the observations needed for reproducing the true (unbiased) dependence with the direct OLS regression. The dashed line rhomboid shows the spread of the observations needed for deriving the unbiased reverse OLS regression. b) The ellipse shows an usual point distribution around the expected dependence when the orthogonal OLS regression (short dashed line) may be recommended as superior, independently on the slope value. The solid and long dashed lines show the distinctly biased direct and reverse OLS regressions.

orthogonal deviations, independently of the slope A. Note that in Fig.3b the line $L(X)$ passes always through the middle point of the segment AE (vertical to the line $L(X)$) and all such segments. Therefore, when don't have additional useful information, the orthogonal regression should be clearly preferable.

Figure 3b also shows, that when $A > 1$ the orthogonal regression will lie close to the reverse regression, and if $A < 1$, on the contrary, close to the direct regression. The same conclusion can be drown from Fig. 2. The examples in the papers IFAB90 (Fig.2, with $A \approx 3.5$) and FB92 (Fig.1, with $A \approx 0.5$) give an independent confirmation of this conclusion. So, the problem with the strong and widely cited recommendations of IFAB90 and FB92 (i) and (ii), marked in the beginning of this Section 2 is, we think, that these authors did not consider or simulate regressions with $A \gg 1$ or $A \ll 1$.

Let introduce here the example in FB92 about the Hubble diagram. These authors give explicit recommendations, as follows. If we are looking for best estimate of H_0, defined by Hubble's law $V = H_0 D + V_0$, where the entire scatter arises in the velocities V and none in the distances D, we must calculate OLS (V—D) regression. If we are looking for the best estimate of the age of the universe, which is proportional to $1/H_0$, we must calculate OLS (D—V) regression. If we are seeking the best estimate of some „structural" relationship between velocity and distance, making no judgement on whether velocity depends on distance or vice versa, we must derive the bisector of the OLS (V—D) and (D—V) regressions.

However, the considerations in this Section 2 give other recommendations, including for the case of the Hubble's diagram, as follows. If we can bound the distribution of the observation point to have 'good' distribution in a rhomboid,

as in Fig.3a, we must apply OLS regression, If we have the general case with elliptical distribution of the observation points, we must apply orthogonal regression. Generally, the bisector of the direct and reverse OLS regressions may be recommendable only if the expected regression slope is close to unity. These recommendations are valid for the RRs too.

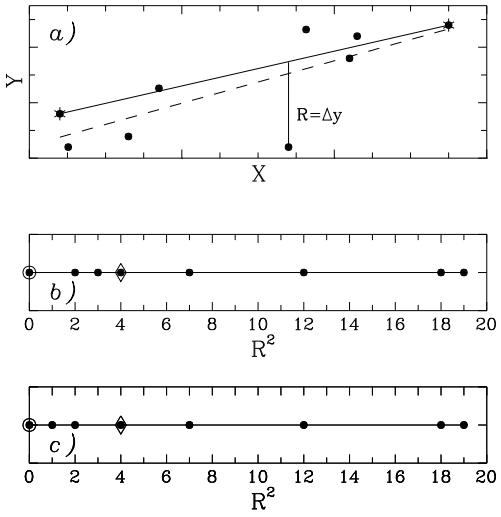
3 General presentation of the robust 1D location estimators and the robust 2D regression

The „least trimmed squares“ (LTS) estimator is introduced by Rousseeuw (1984). The predecessor of this method is the „least median of squares“ (LMS) estimator, proposed by Hampel (1975) and developed by Rousseeuw (1984). These methods are widely discussed in RL87 and the conclusion is that the LTS estimator is superior (Fig.4). Though, the examples in RL87 are based on the LMS method and here we carry out comparisons of both methods. The LMS and LTS are in principle 1D scatter estimators and the RR, described here, is a superstructure over LMS or LTS. The main advantage of this RR is that the fit covers the most populated part of the distribution of the data, independently on the deviations of strong outliers.

Let underline the features of the „standard“ OLS estimator. The principle of the OLS estimator, introduced by Legendre and Gauss in the beginning of 19th century, involves minimization of the sum of the squares R_I^2 of all deviations from the searched point, or line, or plane etc. The application of this principle in the 1D case leads to an estimation of the constant of the distribution (mean, center, location) through the arithmetic mean (average) value of all data. So, in principle the OLS is a 1D scatter estimator and any OLS regression is a superstructure over it. In 2D, 3D, 4D, etc. cases the regression analysis based on the OLS principle leads to a system of linear equations, which solution gives the expressions for coefficient estimations for line, plane, hyperplane, etc.

The principle of the LMS estimator involves minimization of the median of the ordered squares of the deviations R_K^2 (Hampel 1975). The right tail of this order may contain very large deviations, but LMS ignores them (Fig.4b). This method is a reasonable generalization of the idea of the simple 1D median. The LMS derives the median value $R_M = C'C"R_H$ as a robust measure of the scatter of the deviations. Conventionally, $H = N/2 + 1$ is the „half“ of the point number. Here C' is a coefficient used for consistency with the standard error of the Gaussian distribution. Since the half of the Gaussian distribution is bound by $\pm 0.6745\sigma$, the LMS coefficient is $C' = 1/0.6745 = 1.4826$. The other coefficient is a finite-sample correction factor. From numerical simulations LR87 recommend $C - 1 + 5/(N - P)$, where P is the dimensionality.

The principle of LTS estimator involves minimization of the sum of H trimmed squares of residuals R_K^2, and more specifically - the sum of the left-hand half of the ordered squares of the residuals (Rousseeuw 1984). The rest (i.e. the right-hand part of the ordered sequence of squares) may contain very large deviations, but is ignored entirely (see Fig.4c). The LTS estimates the standard error of the scatter as $R_T = C'.C"(\sum R_K^2)/(H - P))^{1/2}$. In this case the consistency with the Gaussian error distribution is achieved by means of the coefficient $C' = 2$. The factor $C"$ is the same as in the LMS case (see above). The value of R_T can be computed also with use of weighs of the data.



Фиг. 4. (a) An example of nine points, the OLS regression through these points (dashed line), a currently tested line, defined by the edge points of the sample (solid line) and one Δy deviation from the tested line. Cases (b) and (c) show examples of ordered squares of deviations from tested lines where the median value (marked with rhombs) in both cases is 4, but the sum of the trimmed squares is 9 in (b) and 7 in (c). The LMS cannot distinguish the cases (b) and (c), but LTS will choose (c).

Since the principle of the RR (based on LTS or LMS) is significantly different from the principle of the OLS regression, the RR solution cannot be computed by simple expressions. that is why the RR method involves qualifying each available pattern of solution - each point in 1D case, each pair of points in 2D case (defining line), each triad of points in 3D case (defining plane), etc. The method of qualifying is LTS (or LMS) estimation of the scatter of the deviations and the result is the pattern with shortest system of deviations. This is the 1st (main) approximation to the RR solution.

Figure 4 is an illustration of the application of LMS and LTS estimators. Figure 4a shows a line that is currently tested (solid line) and one Δy deviation from it. Figures 4b and 4c show examples of ordered sequences of squares of deviations from different tested lines. The deviations with zero value are double, because they correspond to the points defining the tested line. The cases (b) and (c) have equal LMS scatter estimations ($R_5^2 = 4$ in both cases) but different LTS estimations (the sums of the trimmed squares are 9 in (b), but 7 in (c).) So, the LTS recognizes surely that in case (c) the tested line belongs to more concentrated region of the data points and by this reason it is better that the line tested in case (b).

Generally, the 1st approximation of the RR as rough. It is possible also the existence of a hole in the point distribution just in its most populated part. Then the 2nd (final) approximation is obtained by removal of the outliers with respect to the 1st approximation (typically with deviations more than 2.5σ, RL87) and applying of the OLS on the remaining „good“ points. By this way the 2nd approximation gives the standard errors of the regression and

its coefficients. However, the RR allows various methods for measuring of the deviations (see Sec.1), for which the direct OLS solution may be very difficult or impossible. Therefore the 2nd approximation must be also of LTS type. The idea for the proposed here 2nd approximation comes from two historical robust estimators, described in Sec. 4.

A method for the 2nd (final) RR approximation, using again the LTS estimator, is proposed here for the first time (as the author believes). It is based on addition of patterns, produced as averages or bisectors from the pairwised best old patterns. Since the LTS is a natural qualifier of the patterns, numerous best patterns can be extracted instead of only one, the superior pattern. The extracted patterns can be used for enrichment of the region of the solution with new, possible better patterns. In 1D case the enrichment is based on addition of the averages of all pairs of the best points (Fig.5c). In the 2D, 3D etc. cases the enrichment involves addition of bisectors of all pairs of the best lines, bisectors of planes, etc. The added patterns can be tested again by the LTS and a new best pattern (if such is found) can be derived as the 2nd (final) RR approximation.

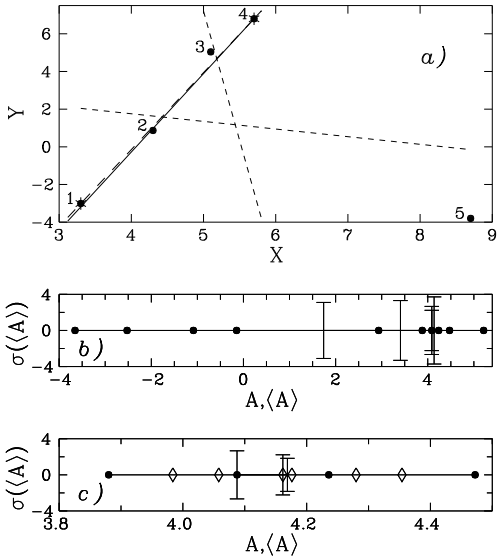
The recommended here number of extracted best patterns is $N^{1/2}+1$, or typically between 10 and 20. Then the number of the additional pairwized patterns will be between 10.9/2=45 and 20.19/2=190. The number of pairwised combinations of these additional patterns is negligible in comparison with the number of the original patterns in 2D and 3D cases (pairs of points, triads of points etc.). The computing time increases by a few percents, but the inner accuracy of the RR solution increases 3-5 times (see Fig.4 and Fig.6).

In principle the RR method does not give estimation of the slope error. The extraction of numerous best patterns allows also an empirical approach to slope error estimation. It is introduced here for the first time (as the author believes). This approach is based on ordering of the best patterns (from 1st approximation only) by their LTS scatter and producing of „error growth curves“ for the coefficients. Each such curve contains the standard error of the coefficients, derived from 2, 3, 4, etc. best patterns. In the 2D case the value of the error curve, corresponding to the best $N^{1/2}+1$ patterns occurs usually very close to the OLS slope error estimation (after removal of outliers). An example is shown for the last case in Fig.9 and Fig.10.

4 The measure of the robustness of the estimator and a comparison of five 1D estimators

Generally, the LTS is an 1D estimator. By this reason its character and its superiority among other such estimators in the 1D case is important.

Figure 5a shows an XY plot with four „good“ points (1-4), situated approximately along a lin. One „bad“ point or strong outlier (5) is situated in the right-down corner of the plot. The direct and reverse OLS regressions for all points are plotted with short dashed lines. Because of the outlier the OLS lines become strongly deviated from the dependence, hinted by the four good points. The influence of the outlier is stronger on the direct regression, because it lies out of the range of the good points more in X than in Y direction. In the case of a direct OLS regression this outlier cannot be recognized by simple check of the residual deviations because the good points (1) and (4) are situated farther from the line than point (5).



Фиг. 5. (a) A simple example of five experimental points including one strong outlier (5). The short- dashed lines show the OLS direct (shallower) and reverse (steeper) regressions over all 5 points. The long-dashed line and the solid line present the robust regressions in 1st and 2nd approximation, respectively. The asterisks show the points defining the 1st approximation. (b) Example of 1D distribution of 10 points, corresponding to the slopes of the lines through all pairs of points in (a). Five estimations of the mean value of this distribution are shown with "standard' error bars (from left to right): average, median, as well as the 1st approximation of the LMS, LTS and Shorth estimators. The estimation from LMS and LTS coincide (LTS has smaller error bar). (c) Large-scale plot of the region of the most concentrated 4 points in (a). Six new points (marked by rhombs) are added as averages of each pairs of the four points (dots). The 2nd approximations within the methods LMS, LTS and Shorth are shown again with error bars, as in (b).

In Fig.5a the 1st and 2nd approximations of the robust regression, derived by the LTS method (Sec.3), are plotted with dashed and solid lines, respectively. In this example the 1st approximation with the LTS or LMS estimators (within all types of deviations, presented in Sec. 2) recognizes the line through the pair of edge points (1,4) as the best fit. The 2nd approximation differs form the 1st approximation only negligibly. After removing the point (5), the direct and reverse OLS regression over points (1-4) practically coincide with the 2nd approximation of the RR.

Figure 5b shows the slopes of all 10 patterns of lines, passing through the pairwised 5 points in Fig.5a, as example of an 1D random variable. This sample will be used (i) for introducing of a parameter that measures the robustness of scatter estimator, (ii) for comparison of five estimators in 1D case and (iii) for 1D illustration of the introduced in this paper (Sec.3) 2nd approximation of the RR method (Fig. 5c).

Figure 5b shows 5 estimations of the mean value of the random variable (or of the „best“ regression slope): average, median, 1st approximation of the methods LMS and LTS, as well as the method „Shorth“ (see below). The

vertical segments are the corresponding error bars. Note that in respect to Fig.4b and 4c, illustrating the work of LMS and LTS as scatter estimators, the purpose of Fig.5b and 5c is illustration of LMS and LTS as estimatiors of the mean value of an 1D distribution.

The most populated part of the distribution, located in the right part of Fig.5b, contains 6 points. These points correspond to the slopes of the lines passing through the 4 "good"points in Fig.5a. The slopes of the other 4 lines, passing through the "bad"point (5) are located in the left part of the plot as outliers.

In Fig. 5b the median lies between the points (5) and (6), because the number of points, ten, is even. The LMS and LTS methods regard each point as a possible solution (mean value). For each point they compute the squares of the deviations of the other points from it and find the point with the shortest system of deviations as result. In 1st approximation LMS and LTS find as superior the point (7), placed in the most populated part of the distribution. Thus LMS and LTS methods estimate the mode value of the distribution. From this point of view the simple median is worse estimator, because its value is relatively far from the mode value.

In Fig. 5c the most populated part of the distribution, containing 4 original points and the 2nd approximations of the methods LMS, LTS and Shorth (see below), are shown in large scale. These 4 original points are used for producing of 4x3/2=6 new points, which are averages of the pairwised combinations of the original points. In the 2nd approximation LMS does not find better pattern but LTS does. It is one additional point, lying between points (7) and (8). Here the concentration of the patterns that can be examined as solutions is about twice as large. If 6 most concentrated points were usedfor duch purpose, the number of the additional patterns would be 6x5/2=15 and the solution would be about 3 times more accurate. The right-most bar in Fig. 5c shows the result of the 2nd approximation of the method Shorth (see bellow). So, the methods LTS and Shorth find surely the mode of the 1D random distribution, shown in Fig. 5b, giving also a robust estimation of the best regression slope for Fig. 5a. However, this approach to find the best regression slope is not good, because (i) it estimates the slope independently on the intercept and because (ii) it cannot be generalized for MD cases.

Let us introduce a robustness parameter. A conventional measure of the robustness of an estimator against strong outliers is the so-called „breakdown value". It is the proportion (or fraction) of contamination of „bad" data that the method can withstand and still maintain its robustness (RL87).

From this point of view the median value is extremely stable. In Fig.5b the median belongs to the most concentrated part of the distribution, containing 6 (or 50%+1) points. If we take 1 point from the left part of the distribution and place it very far (to the left or to the right) the median will change weakly and will remain in the most concentrated part of the distribution. However, if we take a point from the right of the median and place it to the left of the range of the points, the median will jump significantly to the left. Then the fraction of the points in the most concentrated right part of the distribution becomes 50%. Therefore, for large N the asymptotic robustness of the median estimator is 0.5 or 50% of all points. The robustness should be not larger than 0.5, because then the „good" part of data should be not easily defined.

In Fig. 5b the average value (i.e. the OLS estimation) is the most affected by the outliers and it lies out of the most populated region of the distribution.

It is easy to see that only one distant enough outlier may „breakdown“ the average value arbitrarily, even within great number N of the observed points. Therefore, the robustness (or the breakdown value) of the OLS estimator, 1/N, tends asymptotically to 0 when N increases to infinity. By the same reason the robustness of the OLS regression is also 0.

Let us show that the robustness parameter can have an intermediate value, such as 29%.

There are two non-OLS estimators, published about 50 years ago, and based on the median estimator. (i) An old method for robust estimation of 1D location is based on the median of all pairwise means in the sample of N points (RL87, p.164). The same method can be applied in the problem of MD location (e.g. for 2D or 3D centering of stellar systems). (ii) The first non-OLS robust regression estimator for N points in 2D case is the „median of lines“, defined by pairwised points. The coefficients of the best line are derived as medians of the corresponding coefficients of all used lines (RL87, p.67). The median value in Fig. 5b illustrates just this case. So, in (i) and (ii) the simple median estimation is used. Let us estimate the robustness of this method, but from the point of view considering the number of the original data points. The number of the pairs of points, i.e. the combinations of points in both cases is $N_C = N(N-1)/2$. If the number of outliers is K, the condition for 0.5 robustness can be written as $(N-K)(N-K-1)/2 > 0.5N(N-1)/2$. Then for large N we derive $K/N < 1-(1/2)^{1/2} \approx 0.293 \approx 29\%$.

The LMS and LTS methods, having 50% robustness, contain reasonable generalization of the "median idea"of the above mentioned methods (i) and (ii),the LMS and LTS methods derive the regression coefficients simultaneously. The idea for the creation of new patterns for the proposed here 2nd approximation of the RR (Sec.3) comes also from these two historical examples.

Now let compare the five estimators of 1D location, shown in Fig. 5b.

The estimator „average“ (with zero asymptotic robustness) can be considered as direct product of the OLS principle. It has well known generalizations for MD location (averages by all axes or directions) and MD regression (solution of system of linear equations). The standard errors of the regression and the slope(s) are also well defined and easily computed. However, IFAB90 claimed that the standard formulas for these errors are not entirely correct and proposed more accurate formulas in the 2D case.

The estimator „median“ (with 50% asymptotic robustness) is introduced by Laplace in the middle of 19th century(RL87). This „heuristic“ method belongs to the so-called „range statistics“, whose properties are very difficult to be investigated analytically. Conventionally, the „standard error“ of the median estimation is defined again by a median. It is the median of the absolute values of all deviations from the median estimation of the mean value (or square root of the median of the squares of these deviations), multiplied by 1.4826 (see Sec.3).

The full generalization of the median method for MD regression with robustness of 50% is given by Siegel (1982). In the 2D case with N points the derivation of the line coefficients A and B becomes independent, as follows. In the first stage each point (e.g. point I) is fixed and pairwised with every other point J and the median, e.g. A_I, is derived. In the second stage the median of all such N medians is deriving as a final solution (RL87, p.15). Thus the number of checked combinations in 2D or 3D cases are respectively 4 or 18

times more than in the methods LMS or LTS (but the check is significantly simpler). Though, this method is not very attractive, because (i) the simple median is not the best estimator (see Fig.5b), (ii) the coefficients are derived independently and (iii) the computing time for this method is too long.

The LMS and LTS estimators have the highest robustness of slope estimation and regression error estimation, asymptotically 50% (see the theorems in RL87). The „standard deviation“ of the LMS is computed just as in the case of median estimator (see above). The use of squares of deviation instead of absolute values of deviations allows avoid some peculiar situations discussed in the theory, but this is not so important in the practice of LMS (RL87). The standard errors of the LMS and LTS methods can be computed as it is shown in Sec.3. In the examples regarded in the preparation of this work the standard error of the LTS happens to be slightly smaller than the standard error of the LTS. A special feature of the LMS is that its asymptotic efficiency (as for all median based methods) is characterized with abnormal low convergence rate. This rate is proportional to $N^{1/3}$, while the same rate for the OLS and LTS estimators is $N^{1/2}$ (RL87). This is also not important in practice, however in comparison with the LMS, the LTS estimator is better (see Figs. 4,5,6). The MD generalizations of the methods LTS and LMS are natural.

Another special method with unknown origin, which can derive robustly the mode of an 1D location is so-called „Shorth“ method (RL87, p.164). It is defined as arithmetic mean of the shortest subsample with half number ($H = N/2 + 1$) in the ordered data. The standard deviation of the Shorth estimator is computed in the examples here as the half size of this shortest interval, multiplied by the coefficient $C = C'C$", as in the case of median estimator (see above). The Shorth estimator is very alike to the LMS estimator, however, the computations are more complicated and more time consuming than for LMS or LTS. In a case of a distribution with strong asymmetry the Shorth estimator may be applied twice, using the shortest interval, found in the 1st approximation as a field for a new search, now for the shortest interval, containing 1/4 of all data. Examples for 2nd approximation of Shorth estimator are given in Figs. 5c, 6ab, and 7b.

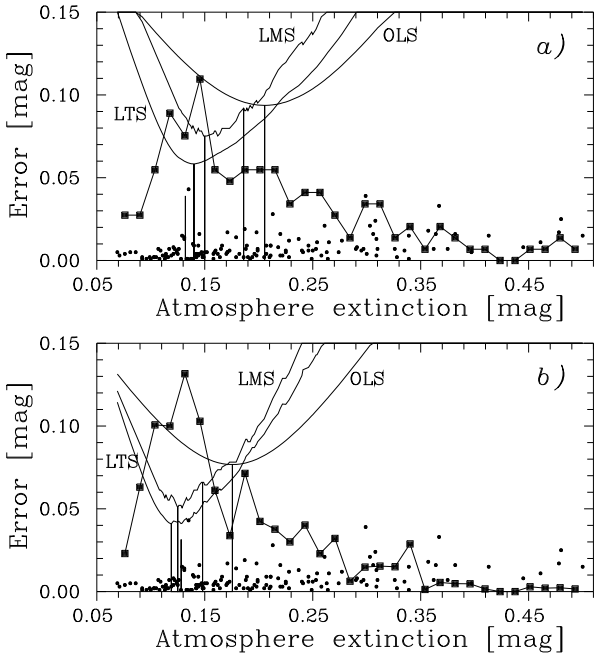
Since the fast median smoothing is widely applied in the astronomical image processing, the implementation of the Shorth method there is an attractive possibility. The method of „hard median filtering“ or „mode filtering“, based on the Shorth estimator, has been developed by the author (Georgiev 2002). The mode filtering removes impulse noise from the image more efficiently than the usual median filtering (i.e. with use of smaller filtering window), but it is a few times slower.

The Shorth estimator is worth noting in the framework of the RRs, but it has a remarkable generalization for the case of MD location. This generalization can recognize the most populated ellipsoid among the data. The principle is „searching for the minimal volume ellipsoid covering (at least) the half of the observing points“ (RL87, p.258). A direct application of this principle involves giant computation time. However, Rousseeuw & Van Driessen (1999) developed a fast method for computations, including even the next level of generalization, called „minimum covariance determinant estimator“.

5 Examples

5.1 Atmosphere extinction mode of the clear Rozhen sky

Figure 6 shows an application of the five location estimators, discussed in Sec.4. It presents with small dots the distribution of the atmosphere extinction in Rozhen NAO in the V band from 146 measurements (courtesy of Dinko Dimitrov, 2007). Only 8 measurements have standard errors ¿ 0.02 mag. Figure 6a presents the results with data weights 1/146 and Fig.6b - with data weights reciprocal to the errors of the observations. Filled squares present the histograms, polynomially smoothed by 5 neighbours. In this example the 1D distribution of the data has a strong asymmetry - a small, well-defined maximum plus long and a heavy right tail.



Фиг. 6. Five estimations of the mean value (the location parameter) in the 1D case. Small dots present 146 measurements of the atmosphere extinction in Rozhen NAO (Dimitrov, 2007). Filled squares present the histogram of the distribution. The shortest vertical segment shows the position and the „standard error“”of the Shorth estimation. Other segments show (from right to left) average, median, LMS and LTS estimations. The „error curves“ show (from up to down) the behavior of the average, LMS and LTS estimators. (a) The individual weights of the data are 1/146. (b) The weights of data points are reversely to their standard errors.

The results from five applied estimators are shown with vertical segments, proportional to the corresponding standard errors. From right to the left they are average, median, LMS, LTS and Shorth. The mode region of this distribution shows high data point concentration and only the 1st approximations

of LMS, LTS and Shorth estimators are presented here. Note again, that the simple median estimator should not be preferable.

This example is mainly for a demonstration of the error curves of the 1D estimators. The error curve is the sequence of standard deviations when the estimator „scans“ and tests the ordered points, searching for the superior point as estimation of the „center“ of the distribution. The scanning mode is natural way for applying the LTS or LMS estimators, while the average, median and Shorth estimators give explicit results. Though, in Fig.6a,b the error curves of the average estimator are especially computed and plotted for a illustration (the highest and shallowest curves). The error curves of the LMS estimator, based on the ordered squares of deviations (or on the ordered absolute values of the deviations) from the tested point are the next curves, deeper and not smooth. The error curves of the LTS estimator are the deepest. The error curves of the Shorth estimator (not shown) are very close to the curves of the LMS estimator. Note that the average estimation, as well as the simple median estimation (derived just after ordering of the data) do not give a good estimation of the mode value of the distribution, but LMS, LTS and Shorth do. Note also, that the mode estimations with LTS correspond to well defined minimums of relatively smooth error curves.

This application leads to a mode estimation of the Rozhen atmosphere extinction in the V band (for perfect atmosphere conditions) of 0.140 ± 0.058 mag in Fig.6a and 0.119 ± 0.041 mag in Fig.6b. Since the use of the weight reversely to the error estimation of the individual data is not especially established, we recommend the first extinction estimation.

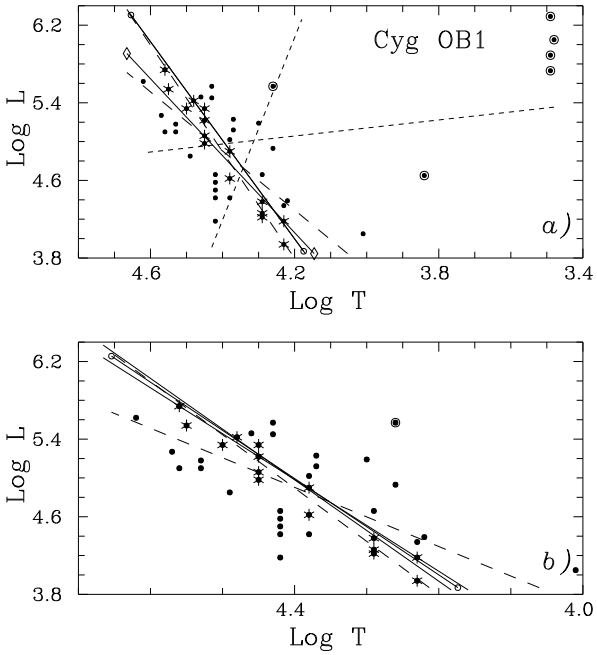
5.2 Fitting the main sequence of the Hertzsprung-Russell diagram

Here we show the applications of the LMS amd LTS regression methods in the 2D case, reproducing the remarkable example of RL87 (p.27).

Figure 7 shows the diagram of the effective temperature T and luminosity L (in solar units) for 47 stars of the stellar association Cyg OB1. The purpose here is to fit linearly the main sequence in a presence of evolved stars, which introduce significant intrinsic scatter in the abscissa direction. After applying the RR we find 7 such stars (14% of the data) as strong outliers.

Figure 7 shows that the difference between the slopes of direct and reverse LMS RRs (a) is significantly larger than for LTS RRs in (b) (solid lines). This is another evidence of the superiority of LTS over LMS. Therefore, the LMS method can indeed be excluded from the applications. Also, since the regression slope is $A = 4 - 5$, the orthogonal RRs (presented by solid lines and edged by circles) in both cases practically coincide with the reverse RRs (Δx based, or of type (2) in Sec. 2). The orthogonal RRs are also very close to the reverse OLS regressions (the steepest long dashed lines), just in concordance with the conclusions for the orthogonal regressions in Sec. 3. In both cases in Fig. 8 the RR, based on the geometric mean deviation $R = (\Delta x.\Delta y)^{1/2}$ (not shown) lies between the direct and the reverse RRs, being close again to the reverse RR. The behavior of the orthogonal and geometric mean RRs correspond well to the expectations from Fig. 2 in case of $A > 1$.

The solution of RL87 with the direct LMS RR, in 1st approximation, is $Y = 3.90X - 12.90$. Our solution with the direct LMS RR, in 2nd approximation, is $Y(\pm 0.43) = 3.94(\pm 0.61)X - 12.60(\pm 2.78)$ and with the orthogonal LMS RR in 2nd approximation is $Y(\pm 0.48) = 5.08(\pm 0.75)X - 12.60(\pm 3.30)$. (The value



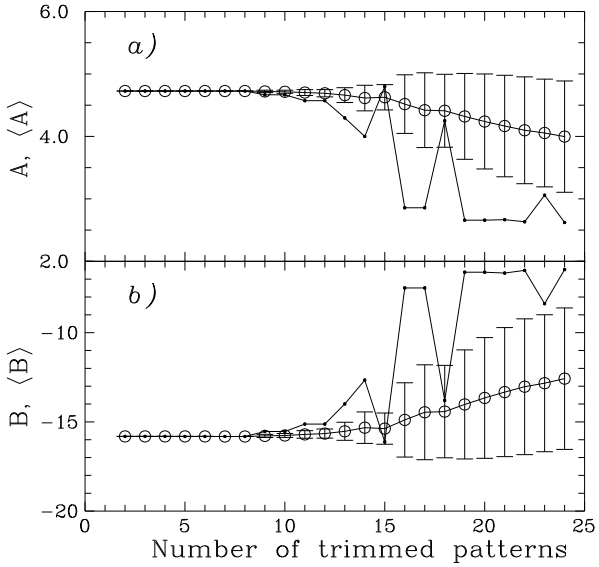
Фиг. 7. An application of the RRs for a linear fit of the main sequence on a Hertzsprung-Russell diagram of the stellar association CYG OB1 (data from RL87). Small dots present 47 stars. Short-dashed lines in (a) present the direct and reverse OLS regressions, which are not meaningful here. The solid lines present the direct and reverse RRs based on LMS in (a) and on LTS in (b). The solid lines with circled edges present the orthogonal RRs. Circled dots present 7 outliers, found as deviating more than 2.5 sigma form the orthogonal RRs. The long dashed lines present the direct and reverse OLS regression, found after the removal of outliers. The asterisks show "the best"5 points, used for 2nd approximation of the methods. The solid line that is edged by rhombs in (a) is our direct LMS RR, coinciding with the unique solution of the problem, given in RL87 (only this RR is presented in RL87). For a better distinction of the LTS RRs the case (b) is plotted with rescaled abscissa.

just after Y corresponds to the standard error of the regression estimation of Y).

Our recommended fit, derived with orthogonal LTS RR with 2nd approximation, is $Y(\pm 0.41) = 5.14(\pm 0.29)X - 17.61(\pm 1.29)$. The coefficient error estimations are carried out on the base of cumulative error curves, as in the next example, and since here $N^{1/2} + 1 = 7$, the errors correspond to the 7th position of the error curves (see Fig.8).

Figure 8 illustrates the empirical approach to coefficient error estimation, proposed in this work, because the RR do not originally have such a possibility. This approach is based on cumulative curves of the averages and corrsponding standard deviations of the coefficients A and B, derived on the trimmed best patterns.

The abscissa of Fig. 8 is the number of the pattern between 2 and 66, where the patterns are ranged by their LTS scatters. The fluctuating lines in (a) and (b) show the behaviour of the coefficients A and B and the smooth curves show the behaviour of the cumulative mean values of the coefficients.



Фиг. 8. The behaviour of the estimated value (fluctuating curve) and the cumulative average (smooth curve) of the coefficients A (a), B (b), as well as the cumulative functions of their standard deviations (c) versus the number of the trimmed best pattern for the example, shown in Fig.7. The vertical line shows the position of the pattern No.12 used here for estimations of the coefficient errors.

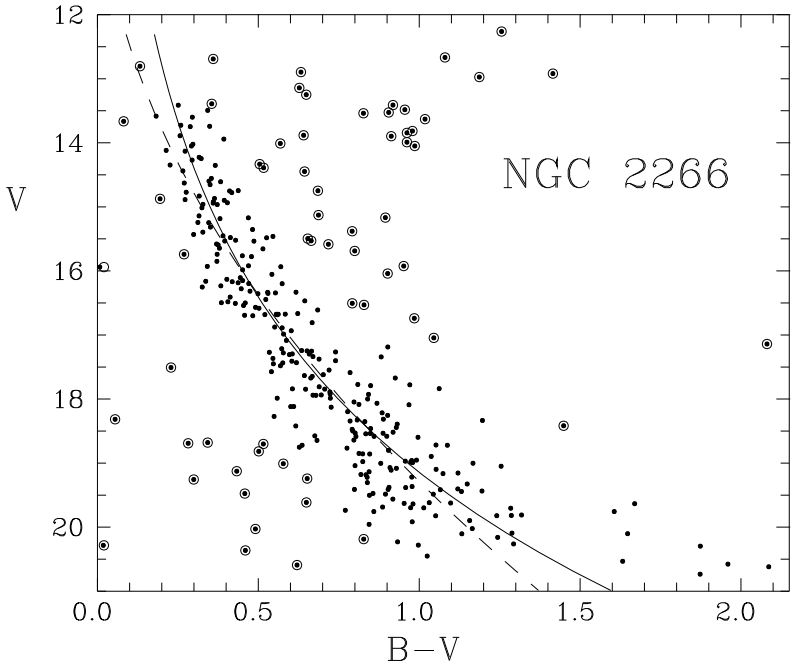
The value above the K-th pattern corresponds to the average value of the coefficient, derived from the patterns with numbers 1 - K. Fig.8c shows the behaviour of the cumulative curves of the standard deviations of the average values of the coefficients. The vertical line shows the pattern with number $N^{1/2} + 1 = 12$ (here N=131) which is proposed to be used for estimation of the coefficient errors.

The proposed method for coefficient error estimation is not connected with any statistical assumptions and therefore it cannot be universal. However, in many similar cases, e.g. many color-magnitude diagrams, it could be used as an additional tool for comparing the results.

5.3 Fitting the main sequence of a color-magnitude diagram

Figure 1 concerns V magnitudes and (B-V) colour indeces of N=331 bright stars in the field of the open star cluster NGC 2266 (courtesy Maciejewski 2007). It is known that the fit of the main sequence in such a CMD difficult task (see the survey of Maciejewski & Niedzielski 2007), but this is an good example for illustration of the strength of the RR.

In the regarded example of CMD background stars and evolved stars introduce significant intrinsic scatter, mainly in the X direction of the CMD. Observing errors both in X and Y data exist too, but they can be consider negligible in respect to the X intrinsic scatter. By these reasons the reverse RR based on minimization of the horizontal deviations (Δx), is appropriated as the adequate model here.



Фиг. 9. An application of the RRs for nonlinear fits of the main sequence on a CMD with LTS based RRs (data of Maciejewsky, 2007). The dashed and solid curves represent the results after previous square root or logarithmic transform of X-data. Circled dots present 62 outliers, found as deviating more than 2.5 sigma form the solid line.

The purpose here is deriving the 2D reverse (minimizing Δx) RRs of the type Y = AX + B. Here Y is V-magnitude, X is colour-indexes, A is regression slope and B is regression intercept. We apply previously square root or logarithmic transform of X data. THese transforms may be presentet as $(B-V) = \alpha V^2 + \beta V + \gamma$ or $(B-V) = \alpha.10^{\beta.V}$. The results are shown in Fig.9 by dashed or solid line, respectively.

6 Concluding remarks

The presented examples show that the RR method, based on the LTS, is a powerful tool for fit a empirical dependences with large numbers of outliers. The results of the present work gives evidences, that when the usual statistical assumptions are strongly violated the LTS of the orthogonal deviations should be the best choice.

The RR method is in principle computer time consuming. The number of the tested patterns (combinations of pairs, triads etc of points) N_C in MD case is given by the Newtonian binomial coefficients: $N_C = N$ in 1D case, $N_C = N(N-1)/2$ in 2D case, $N_C = N(N-1)(N-2)/6$ in 3D case etc. Since N_C is proportional to N^P, where P is the dimensionality, the number of tested patterns might be too large. If $N = 100$. in 2D case $N_C \approx 5000$, if $N = 1000$ $N_C \approx 500000$.

The simplest algorithm of the LTS method in 1-st approximation, applicable for deriving a RR line in 2D space, can be presented as follows.

Suppose that we are searching for the regression line $Y = AX + B$ through N data points (x_i, y_i). We test the lines passing through all pairs of points as patterns of possible solution (fit). For every such line we derive and order the squares of the deviations of the points from the line. Then we find the median of the ordered squares of deviations (the value of the deviation No. H, where $H = N/2 + 1$), as well as the (eventually weighted) sum of the trimmed squares of the deviation (from 1 to H). We choose the line with the smallest sum of trimmed squares as the best one. We can multiply this sum with suitable coefficient to get a result compatible with the Gaussian standard error. In a 3D case, in a search for a plane $Z = AX + BY + C$, we apply the same procedure, using the planes passing through all triads of points as tested patterns.

The 2nd approximation of the fit, if need, can be envisaged in the above mentioned algorithm and numerous best lines (or planes), typically about 20, can be extracted instead of only one, the superior. The bisectors of the extracted lines (or planes) can be tested in the same way in a search for a better fit. More information about the 2nd approximation and error estimation will be given in the future (Georgiev 2008).

In the end, let compare the OLS and RR methods. The features of both methods are collected in Table 1.

Таблица 1. A parallel between the features of the OLS and RR methods

Particularity	**OLS regression**	**LTS robust regression**
Purpose	Fitting with objective	Fitting with obvious
	(proved) validity	(subjective) validity
Demands about data	Absence of outliers	Presence of "main sequence"
Demands about errors	Normal distribution etc.	None
Principle	Minimizing the sum of all squares of deviations	Minimizing the sum of the trimmed squares of deviations
Method	Applying of OLS formulas	Testing available patterns by LTS
Tool	Using Δy deviation	Using many kinds of deviation
Outlier detection	In some cases, not surely	In all cases,robustly
Robustness	Asymptotically 0 (min.)	Asymptotically 0.5 (max.) (Sec.4)
Internal accuracy	Originally high	Originally low. Possible increasing: Adding intermediate patterns (Sec.3)
Slope error estimation	Applying OLS formula	Originally none. Possible solution: Cumulative error curve (Sec.3)
MD generalization	By default	By default
Polynomial applying	By default	In some cases (to be elucidated)
Computing time	$\propto N$	$\propto N^4$

Acknowledgements. The author is thankful to Dr. G. Maciejewski for the discussions and to Dr. K. Stavrev and Dr. R. Bachev for the critical reading of the manuscript and many important remarks.

The presented work is supported by the Grant BY-01-06 of the Ministry of Education and Sciences of Bulgaria.

References

Akritas M.G., Bershady M.A., 1996, ApJ 470, 706-714
Branham R.L., 2001, NewAR 45,649
Chen C., 2007, SAS Institute Inc., Paper 265-27, www2.sas.com/proceedings/sugi27/p265-27.pdf
Dimitrov D., 2007, private communication
Feigelson E.D., Babu G.J.,1992 ApJ 397, 55 (FB90)
Fox J., 2002, cran.r-project.org/doc/Fox-Companion/appendix-robust-regression.pdf
Georgiev T., 2002, Bull. Spec. Astrophys. Obs. 53, 114
Georgiev T., 2008, in preparation
Hampel F.R., 1975, Bull. Internat. Stat. Institute 46, 375
Hodges J.L., Jr., Lehmann E.L., 1963, Ann. Math. Stat. 34, 598
Isobe T., Feigelson E.D., Nelson P.J., 1986 ApJ 306, 490
Isobe T., Feigelson E.D., Akritas M.G., Babu G.J., 1990 ApJ 364, 104 (IFAB90)
Kaspi S. et al. 2005, ApJ 639, 61
Kelly B.C., 2007, AJ 665, 1489
Macejewski G., 2007, private communication
Macejewski, G., Niedzielski, A., 2007, A&A, 469, 1065
Olive D., 2007, Applied Robust Statistics, Preprint M-02-06, www.math.siu.edu/olive/olbookp.html
Rousseeuw P.J., 1984, J. Am. Stat. Assoc. 79, 871
Rousseeuw P.J., Leroy A.M., 1987, Robust Regression and Outlier Detection, John Willy & Sons (RL87)
Rousseeuw P.J., Van Driessen K., 1999, Technometrics 41, 212
Rousseeuw P.J., Yohai V.J., 1984, in Robust and Nonlinear Time Series Analysis, edited by J. Franke, W. Hardle and R.D.Martin, Lecture Notes in Statistics, 26, Springer, p. 256.
Siegel A.F., 1982, Biometrica, 69, 242
Tremaine S. et al, 2002, ApJ 574, 740
Vansina F., De Greve J.P., 1982, Astrophys. Space Sci. 87, 377
Wikipedia 2007 - en.wikipedia.org/Robust-regression
Yohai V.J. 1987, Annals of Statistics 15, 642