Bulgarian Academy of Sciences. Space Research and Technology Institute.
Aerospace Research in Bulgaria. 26, 2014, Sofia

ROBUST SMOOTHING OF TIME SERIES BY SIMPLE FAST ALGORITHM. TRACING OF TREND IN STELLAR FLICKERING AND CONTINUUM OF STELLAR SPECTRUM

Tsvetan Georgiev

New Bulgarian University,
Institute of Astronomy and NAO, 1784 Sofia
e-mail: tsgeorgiev@nbu.bg

Abstract

The method of the moving data window has been widely used for tracing the behavior of time series on a large scale where the estimation of the central point of the window is based on the Method of the Least Squares (MLS). However, the ordinary MLS minimizes the scatter of all n squares of the deviations and it is extremely sensitive to strong outliers. One alternative is the Method of the Least Trimmed Squares (MLTS) of Rousseeuw that minimizes only the left part of the squares of the deviations, ordered increasingly, including at least h = n/2+1 data points. Strong outliers may be present in the right part of this order, but the MLTS ignores them. Thus the MLST has an asymptotic robustness of 50% against strong outliers in the data, while the robustness of the MLS is definitely 0%. Apart from that the MLS ordinary regression is derived by direct formulas with respect to the coefficients while the MLTS robust regression is derived by testing all the available patterns of possible solutions: single data points in 1D case, lines through pairs of points in 2D case, planes through triplets of points in 3D case, etc. The pattern that has the shortest MLTS scatter is revealed as a solution. The main disadvantage of the MLTS is that in 2D, 3D, etc. it needs huge computing time in order to check all the available patterns. It may take a few million – billion times longer than it takes for the calculation of the respective ordinary regression. This work presents (i) a simple fast algorithm for the MLTS that omits progressively numerous patterns and may reduce the computing time a few thousand – million times. It presents also (ii) the capability of the MLTS applied in processing time series, especially with respect to the task of tracing stellar light curves in the presence of flares and tracing continuum stellar spectra in the presence of many spectral lines. Here we deal with equally spaced time series, but the method can be applied for all cases as a general solution.

1. Smoothing by the Method of the Moving Polynomial (MMP)

A time series is an ordered discrete sequence of values that are dependent on time or other argument. Examples of a time series are a stellar variability curve, a stellar spectrum, a photometric section of a galaxy image, an index of a geophysical activity, etc. Usually the procedures applied on a time series aims (i) to decompose it into a trend and short scale variations, or (ii) to forecast some intermediate or further values of the time series. In many real cases the noise contamination disturbs the time series and at least a preliminary smoothing for suppressing the noise is needed. Theory and recommendations for time series smoothing are given in many books [1,2,3,4], as well as in many contemporary manual.

A common and widely used method for suppressing the noise is based on the so called moving data window. Let us define a time series z_k, $k = 1, \ldots, n$ and a data window of size w. We suppose w is an odd number and $1 << w << n$. Then the smoothing method works, as follows. The center of the data window moves along the time series step by step along the input time series. For every fixed position of the window, centered on the data point k, a numerical method uses the data in the window and estimates a "better" value $<z_k>$, corresponding to z_k . In the output time series $<z_k>$ replaces z_k . Usually the estimation of z_k is based on the Method of the Least Squares (MLS) and the estimation $<z_k>$ is the central value of a regression polynomial of a low degree p, $F_p(z)$, describing the large scale trend of the window data: $<z_k> = F_p(z_k)$. In the simplest case, $p=0$, the estimation $<z_k>$ is the average of the data in the window. In other cases, a regression line, $p=1$, a regression quadratic polynomial, $p=2$, etc., are used in the sense of an average line, an average quadratic polynomial, etc. Let us call this common method "Method of the Moving Polynomial" (MMP).

The MMP based on the MLS produces an output (a result) time series that has been smoothened at a scale shorter than the window size. At a fixed window size w, when the polynomial degree p increases, the smoothing effect decreases, i.e. more details stands out in the output (result) time series. Otherwise, at a fixed polynomial degree p, when the window size w increases, the smoothing effect increases too.

In the case of equally spaced data, explored here, a significant simplification of the calculations of the MLS estimation of the current value (the central value of the window) exists: (i) the regression coefficients ahead the odd powered polynomial terms are definitely zero and (ii) the MLS procedure may be changed by convolution of the time series with

preliminary calculated kernel coefficients [5]. Formulas for deriving the coefficients in 1D and 2D cases for p=2 and p=4, as well as an application for smoothing of digital image of a galaxy has been published [6]. So, excluding the average (with $p = 0$) that causes too strong smoothness, the simplest tracing of a complicated time series may be based on the MLS parabola (with $p = 2$): $\langle z \rangle = b_0 + b_2.t^2$.

Theoretically, the MLS is applicable over a system of statistical assumptions. The main of them is that the observed values of the dependent variable (z) are subject to errors with zero mean and a finite variance, common for all observations. On the contrary, if only one strong outlier is present among the data, the MLS is practically useless. The problem is very serious (i) when the number of outliers is large, e. g. 40% of the data, (ii) when the number of the arguments (independent variables) is larger than unit when the visual control is almost impossible, and (iii) in time series processing or image processing, when the program code should be able to ignore automatically numerous outliers.

Furthermore, the MMP (based on the MLS), being a linear transformation of the time series, saves the "total energy" (entropy, self-information) of the data. For this reason the strong impulses in the data spread and disturb the behavior of the time series at scales compatible with a double-sized window. The high sensitivity to impulse noise is the most fundamental disadvantage of the MLS. Though, the MLS is the best (i) when the supposed intrinsic behavior of the time series at large scales is simple (naturally smooth) and (ii) when the noise distribution is close to the normal distribution. Otherwise, a method that is robust against numerous strong outliers is urgently needed.

A wide spread robust method gives an estimation of z_k as the median of the values in the moving data window [7]. The median is a robust estimation of the population mean with an asymptotic robustness of 50 % against outliers. Again, when the window size increases, the smoothing effect increases too. However, the median method saves sharp edges and produces result time series which are jagged at the shortest scale. For these reasons an additional smoothing by the MMP after the median smoothing is recommended. Unfortunately, simple methods for building median line, median plane or median are not certain.

It is very attractive to have a smoothing method that combines the flexibility of the MLS and the robustness of the median smoothing. Moreover, while the MLS estimates average means, average lines, average

planes, average polynomials, etc., this method has to estimate mode means, mode lines, mode planes, mode polynomials, etc.

The application of the MMP presented here is based on an extremely robust method, described in Section 2. Because of its specific character the MLTS may take millions – billions times longer in respect ot the MLS and for this reason the MLTS is not widely spread. Therefore in Section 3 we present a simple fast algorithm for applying of the MLTS that may reduce the computing time thousands – millions of times. In Section 4 we apply the MMP based on the MLTS to trace the stellar light curves with flares and in Section 5 we apply this approach to trace the continuum in the stellar spectrum with many lines.

2. The Method of the Least Trimmed Squares (MLTS)

The ordinary method of the least squares (MLS) is based on the principle of the least squares, introduced by Legendre and Gauss at the end of XVIII century. The MLS estimator minimizes the sum of all n squares of deviations. Its two most important particularities are: (i) the estimations are presented by formulas for direct calculation of the coefficients and their standard errors (advantage) and (ii) the estimations have zero robustness against outliers (disadvantage).

Different improvements of the MLS, aiming robustness against impulse noise, are proposed in the scientific literature, but we concentrate on the extremely robust method based on another principle. It has been introduced by Peter Rousseeuw in 1984 [8] and it is known as "Method of the Least Trimmed Squares" (MLTS). The principle of this method that changes the principle of the MLS is: the best estimation minimizes the sum of the left half of the squares of the deviations ordered in an ascending order (ordered by the increasing), no less then $h = n/2+1$ for n data points.

The MLTS differs very significantly from the MLS in two respects: (i) The estimations are not to be presented by formulas for direct calculation of the coefficients and the standard deviations of the coefficients. For this reason any estimation should be made testing numerous patterns and this can be extremely time-consuming; (ii) The MLS estimation has an asymptotic robustness of 50% against outliers. For this reason practically up to 40% of the outliers do not change the estimation. Beside this, while h increases, the robustness of the MLTS decreases. In the case of $h = n$ the estimation through the MLTS coincides with the estimation through the MLS. However, if $h < n/2+1$, the MLTS may recognize wrongly a small

part of the distribution as a keeper of the mode value. In the present paper we explore only the number $h = n/2+1$.

The MLTS is widely discussed and illustrated by Rousseeuw & Leroy in 1987 [9]. Some astronomical applications have been presented as illustrations of the power of this method by Georgiev in 2008 [10].

The simplest application of the MLTS is the estimation of the mode mean of a 1D population. Let us take the sample z_j, $j = 1, \ldots, n$ into account. Then the MLTS works, as follows.

0. It takes into consideration consequently each value z_j, regarding it as a possible mode estimation. (The number of all checked points is $N = n$.)

1. It derives for every z_j all the n squares of the deviations $\Delta z_{jk}^2 = (z_j - z_k)^2$, $k = 1,\ldots, n$.

2. It sorts the values Δz_{jk}^2 increasingly and trims the first $h= n/2+1$ of them, ignoring the others.

3. It calculates the sum S_j of the trimmed squares of deviations and uses this sum as a label of the goodness of the data z_j as an estimation of the sample mean value;

4. It announces the value of z_j. which has the shortest sum S_j to be the estimation of the mode of the 1D population;

5. It announces the value $s = 2\times[S_j/(h-1)]^{1/2}$ as an estimation of the standard deviation of the population. Multiplying by 2 is necessary for compatibility with the standard deviation estimation, that is based on half the deviations, with such an estimation by the MLS, that is based on all the deviations.

Figure 1 show an example composed of 138 measurements of the atmosphere extinction of the *Rozhen* NAO with a standard error of a single value of about 0.01 mag (about 1%) (courtesy of Dimitrov [11]; see for details Fig. 5 in [10]). Three estimations of the population mean are shown as average, median and mode. Note that the derivation of the MLTS mode is based on a cleat mathematical principle and it does not need a histogram presentation of the data.

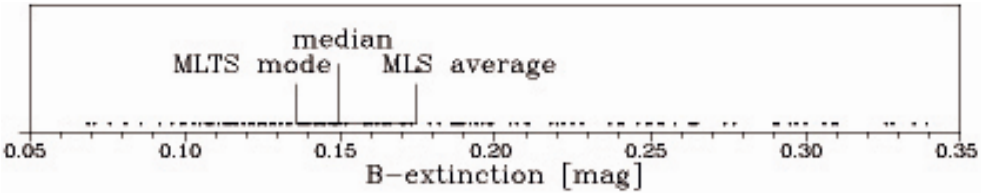

Fig. 1. Comparison of the positions MLTS mode, median and MLS average on a random value with a heavy right tail: atmosphere B-extinction over the Rozhen NAO. The mode estimation by the MLTS does not need visual (histogram) presentation of the data

Further, searching for the mode in 2D, 3D, etc., discrete distributions, MLTS checks every point (vector) r_j , as a possible mode estimation in 2D, 3D, etc. space. MLTS applies the same scheme, as in the 1D case, using the respective squares of deviations $\Delta r_{jk}^2 = |r_j - r_k|^2$, $k = 1,\ldots, n$; The number of the checks in these applications is always $N = n$.

The MLTS is designed mainly to derive the robust (mode) regression line $<z> = b_0 + b_1.t$. In this case MLTS checks the lines through all pairs of points as a possible solution:

0. It derives the parameters b_0 and b_1 of the line $z = b_0 + b_1.t$ through every pair of points.

The number of checked pairs (combinations) is $N = n.(n-1)/2$.

1. It derives all n squares of deviations Δz_{jk}^2 (for each pint k, $k=1,\ldots, n$, of the sample) with respect to every checked line j, $j=1, \ldots, N$.

Furthermore the MLTS follows the steps 2 – 5 in the previous example and derives the line that is best among the available line patterns.

Searching for 2nd degree (mode) regression curve (or mode regression plane $<z> = a.x + b.y + c$), the MLTS follows the same scheme, checking every triad of points. Than the number of combinations is $N= n.(n-1).(n-2)/6$. In the case of 3 arguments MLTS checks every four points and the number of the combinations is $N = n.(n-1).(n-2).(n-3)/24$, etc.

In this work we show applications of fitting or smoothing of time series or data rows using four kinds of low degree polynomials:

(2.1*a*) $<z> = b_0 + b_1.t$

(2.1*b*) $<z> = b_0 + b_2.t^2$

(2.2) $<z> = b_0 + b_1.t + b_2.t^2$

(2.3) $<z> = b_0 + b_1.t + b_2.t^2 + b_3.t^3$

Figure 2 shows examples with light curves (LCs) of the variable stars V 425 Cas and KR Aur that contain irregular fast light variations (flickering). The LC are obtained with the 2 m telescope of the Rozhen NAO, [12] and [13], with 162 and 64 data points, respectively . The levels of the MLS average, median and MLTS, as well as the regression polynomials of 2nd and 3rd degree, derived by the MLS and MLTS, fit all data.

In Fig.2*a* the general trend of the data follows the shape of a 2nd degree polynomial. By this reason both polynomials of the type (2.2) are closely situated. In this case the MLTS does not show some advantages. However, in Fig.2*b* the general (calm) trend follows an approximately horizontal line and both 3rd degree polynomials of the type (2.3) are essentially different. The MLS polynomial is deviated by a large flare, while the MLTS polynomial recognizes and elucidates the horizontal trend, ignoring the flare. In this case the MLTS shows clearly its robustness against outliers.

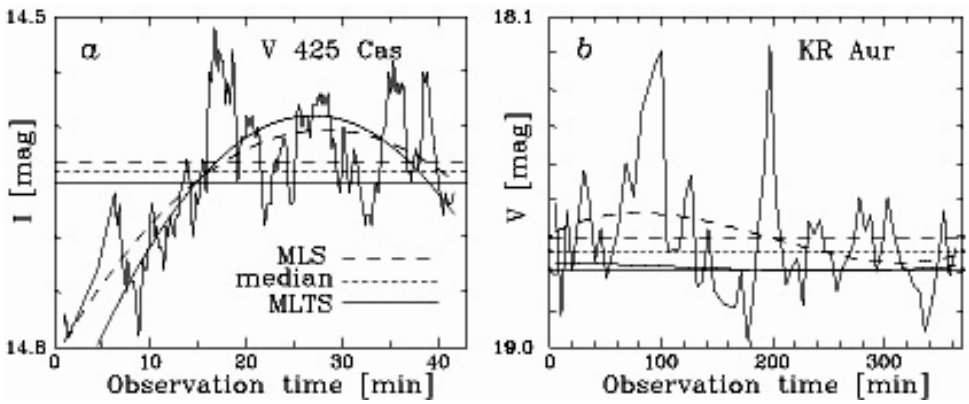

Fig. 2. Fitting of light curves of flickering stars with comparison of the MLTS mode, median and MLS average (horizontal lines), as well as of polynomial curves of 2nd degree (2.2) in the case a and 3rd degree (2.3) in the case b. In the case b The MLTS mode line or the MLTS 3rd degree polynomial may be used for detach of a residual curve and an estimation of the energy of the flares

3. Simple fast method for the application of the MMP though MLTS

Searching for the best polynomial, the MLTS must check a large number of combinations. This number increases fast with the increase in the

polynomial degree: $N \sim n^2$ for the 1st degree (2.1*a*) or (2.1*b*), $N \sim n^3$ for the 2nd degree (2.2), $N \sim n^4$ for the 3rd degree (2.3) etc. That is why the building of 3rd degree polynomial by the MLTS over 100 points needs to test $\approx 15.7 \times 10^6$ combinations, but over 1000 points it needs to check 44.2×10^9 combinations (Fig.3*a*, upper dashed line). Such tasks may take decades of computation time.

However, such a consecutive test of millions – billions combinations is not necessary. The practice shows that the number of the combinations that have to be checked in order to obtain an optimal result may be reduced thousands – millions times. In particular, a random number generator may be used to trim a high enough number of arbitrary combinations, but the simplest way is omitting numerous combinations, by attributing smaller importance to them. The simple fast method, described below, is based on omitting the neighboring combinations and it needs data that is preliminary sorted in an ascending order by argument *t*. When the number of arguments is larger than one, the data must be sorted in an ascending order by the first argument, and if need by the second, etc., arguments.

Let us concentrate on the simplest case (2.1*a*) or (2.1*b*) with a full number of combinations (pairs of points) $N = n(n\text{-}1)/2$. All such combinations may be counted and tested by the following C-code

(3.1)
```
N=0;  for (i=0;  i<n-1; i++) for (j=i+1;  j<n; j++) {  N++;
      /* Here is the place of the code that tests and labels the line patterns */  }
```

However, the neighboring pairs of points, numbered as (*i*,j), like (0,1) (1,2), (2,3), etc., (0,2), (2,4), (4,6), etc., or, generally, (0,0+*m*), (0+ *m*, 0+2 *m*), (0+2*m*, 0+3*m*), etc., may be omitted as close neighbors and less useful. Generally, beginning with the point numbered 0 and using only the pairs of points that have difference divisible bay *m* between their numbers, we may thin out the number of combinations about m^2 times. We could use all the points, testing also for the cases (*k*, *k* +*m*), (*k* + *m*, *k* +2 *m*), (*k* +2*m*, +3*m*), etc., for *k* =0,…,*m*, i.e, about *m* more times . So, such thin out procedure must be applied by the C-code

(3.2)
```
M=0;  for (k=0;  k<m; k++) for (i=k;  i<n-m; i+=m) for (j=i+m;  j<n;
      j+=m) { M++;
      /*  The code that tests and labels the line patterns must be written here  */  }
```

Here M is the number of the used combinations. Thus the reduction gain becomes $(N/M) \sim m$ in the case of (2.1*a*) or (2.1*b*), $(N/M) \sim m^2$ in the case of (2.2) and $(N/M) \sim m^3$ in the case of (2.3).

For example, in the case of $n = 13$ points, numbered as 0, 1, 2, …, 12, $p=1$ and $m = 1$,we have to check the full number of combinations, $N = 78$. However, if we use $m = 3$, we have to check $M \approx N/3$ combinations. Really in respect to (3.2) the combinations are 21. These combinations are shown in Fig. 3.

(0,3)	(0,6)	(0,9)	(0,12)	(3,6)	(3,9)	(3,12)	(6,9)	(6,12)	(9,12)
(1,4)	(1,7)	(1,10)		(4,7)	(4,10)		(7,10)		
(2,5)	(2,8)	(2,11)		(5,8)	(5,11)		(8,11)		

Fig. 3. Inventory of the combination used with applying of the fast method for MLTS regression line (p = 1) on n=13 points with thin out step m =3. The number of these "good" combinations is M =21, while the number of all combinations is N = 78

Here we present a method for automatic progressive increase of the thin out step m in dependence on n. The increasing is shown in Fig.4*a*. The user must supply a suitable supporting number, f.e. $n_0 = 21$. In that case, if $n \le n_0$, the computer program will use all combinations, corresponding to n, as in the general case (3.1), with $m = 1$. If the number of points in the current application of the MLTS occurs $n > n_0$, a suitable thin out step of $m > 1$ will be derived and used, so that it reduces the number of the used combination as in (3.2).

The C-code, given below, shows the automatic derivation of m, in dependence of n and a supporting number supplied by a user n_0 , with a respective number of combinations N_0, derived by (3.1). This code increases the thin out step m (Fig. 4*a*) and defines the number of checked combinations M to be more and more large than N_0, but with enough slow increasing (Fig.*4,b*, thick graphs).

```
/* Here is the part of the program that calculates N0 from n0 though
(3.1) */
(3.3)   M=0;  m=1;  if(n>n0) {
        N=(n0-1)*n0/2.;  for (l=n0;  l<=n; l+=2)  {
        M=0;  for (k=0; k<=m; k++)  for (i=k; i<l-nd; i+=k)  for (j=i+k; j<l;
        j+=k) { M++;
```

```
if (M>N0 && l/k>2) { N0=M; m++; }      }     }
```

The result is the thin out step *m* that will be used for C-code (3.2), as well as the preliminary derived number *M* of the combinations to be used. In this approach the thin out step *m* increases slowly but progressively with the increasing number of points *n*. The reduction gain *N/M* increases rapidly.

Figure43*a* shows the increasing of the thin out step *m* number in dependence on *n*, after the code (3.3). Figure 4*b* presents the slow increase of the used combinations *M* (thick jagged curves) and the fast increase in gain (*N*/*M*) (jagged curves at the bottom down corner).

In the examples given in Fig.4 the user-supplied supporting number n_0, that starts the increase of *m* and the increase of *N/M,* is 75, 32 and 22, respectively. There in the case of *n*=1000 points the thin out step tends to *m*=100 (Fig.4*a*). In the same time in the cases of polynomials of 1st, 2nd and 3rd degree the gain *N*/*M* is about 100, 3000 and 110 000 times, respectively (Fig. 4*b*).

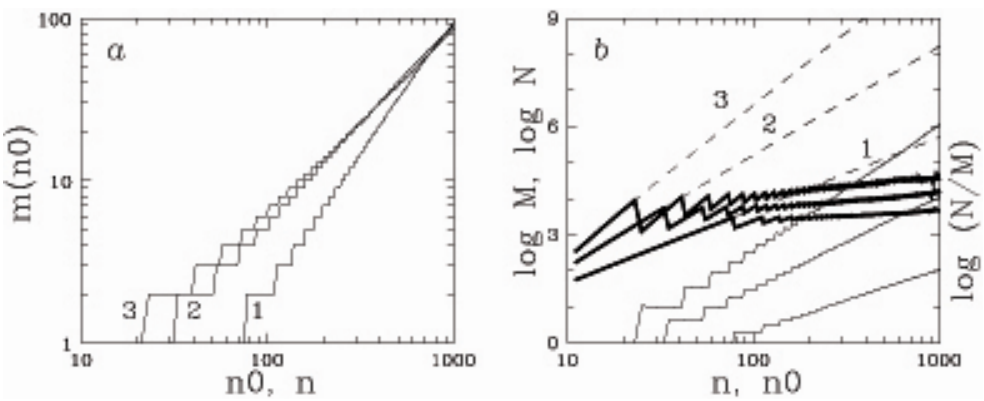

Fig. 4. a. Increasing of the thin out step m in dependence on the data number n and user supplied beginning number n_0, where the numbers 1, 2 and 3 correspond to the models (2.1), (2.2) and (2.3). b. Increasing of the full combination number N (dashed lines), used number of combinations M (thick graphs) the gain N/M (graphs in the right-down corner) in dependence on n. The numbers correspond, as in a, to polynomials of 1st (1), 2nd (2) or 3rd (3) degree

Figure 4 shows that the proposed fast method, based on omitting of combinations, makes the MLTS and the MMP based on the MLTS really useable. Some applications are given below.

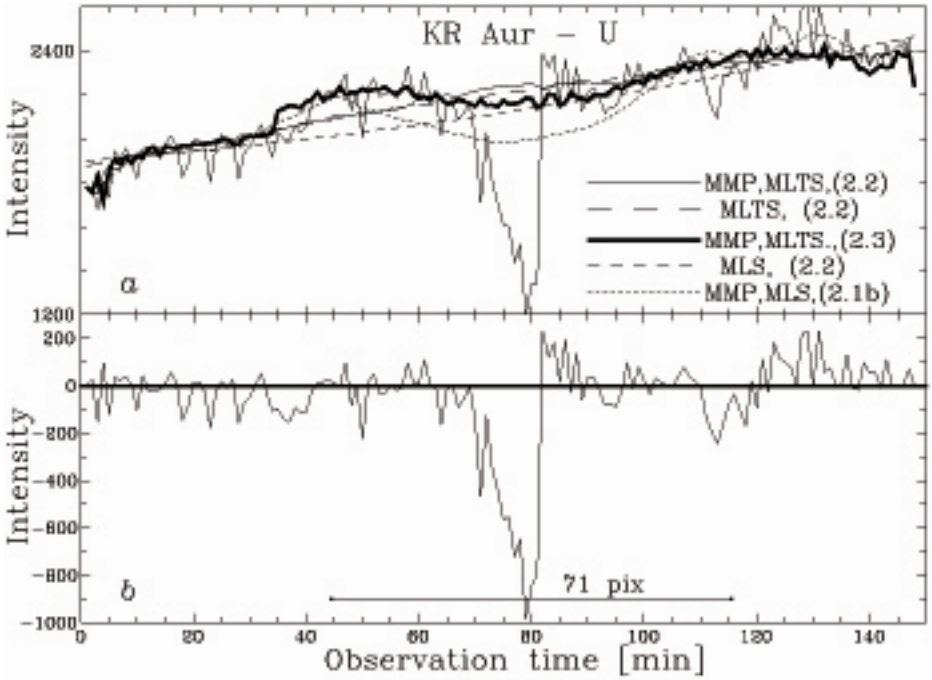

Fig. 5. a. Results of tracing of a light curve of the cataclysmic star KR Aur by various methods, signed in the picture. In the cases 2 and 4 the regression curve over all points is build. In the other 3 cases different smoothing methods with window size 71 pix (points) are used; b. Residual light curve with respect to the MMP smoothing by the MLTS (2.3) in a. (thick curve)

4. Tracing a large scale trend in stellar light curves with flickering

The flickering of symbiotic and cataclysmic stars produces complicated light curves where both large scale trends and short scale variations are of astrophysical interest.

Figure 5*a* shows a light curve (LC) of the cataclysmic variable KR Aur with 148 points (60 cm telescope of the Belogradchik AO, [13]). A significant sink of a 2-fold light decrease and duration of about 15 min dominates in the LC. The general behavior of the LC is fitted by a 2^{nd} degree polynomial (2.2) through the MLS and through the MLTS. Note that the MLS polynomial is affected by the sink of the LC and it has a concave curve, but the MLTS polynomial ignores the sink and shows a more realistic convex curve. The LC is also smoothed by the MMP with a window size 71

pix (points) through the MLS (2.1b), as well as through the MLTS (2.2) and MLTS (2.3). (For visualizing of the jaggedness of the MLTS result additional smoothing by the MLS has not been applied.)

The last mentioned smoothing may be considered the best and useful for deriving the "energy" of the sink: Figure 5*b* shows the residual LC with respect to the MMP, made though the MLTS (2.3).

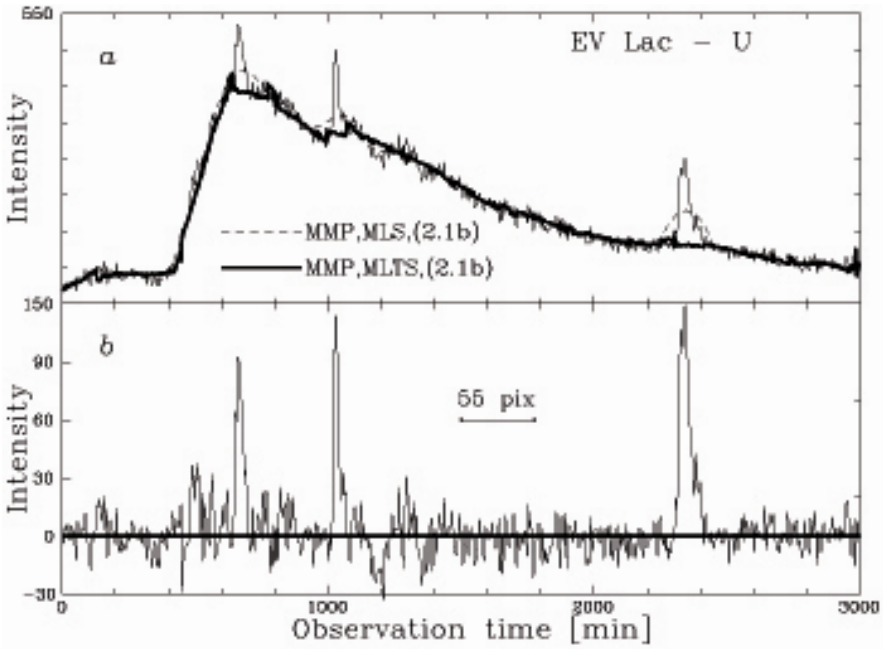

Fig. 6 a: Results of smoothing of the LC of a strong and continuous outburst of EV Lac with a window of 55 pix (points) by 2 methods, signed in the picture.
b. Residual LC with respect to the MMP smoothing by the MLTS (2.3)

Figure 6*a* represents the LC with 600 points of a remarkable power outburst of the active red dwarf star EV Lac (60 cm telescope of the Rozhen NAO, [14]). The general photometric behavior of the outburst is presented by smoothing with a window size of 55 points (275 min) by use of the MLS and the MLTS of type (2.1*b*). Note that the MLS smoothens and spreads the local short outbursts, while the MLTS ignores them.

Figure 6*b* shows that the residual LC with respect to MLTS is smoothened. (Additional LMS smoothing of the MLTS smooth is not applied). The residual LC elucidates clearly at least three well pronounced short time outbursts with a duration of 100 – 200 min. The applied MLTS

method gives possibility for deriving the energy of the main outburst as well as the energy of the flickering outbursts.

5. Tracing the spectral continuum among many spectral lines

The deriving of the continuum of a stellar spectrum containing numerous spectral lines is an important and difficult task. The MMP based on the MLTS gives a reasonable solution.

Figure 7*a* presents a part of the spectrum of the AM star HD 033254 through 900 data points with a step of 0.1 Å (2 m telescope of the *Rozhen* NAO [15]). The continuum seems to be linear and the regression line, build by the MLTS (2.1*b*), confirms clearly this impression. The respective LTS regression line is deviated down by the absorption spectral lines and it is useless. Furthermore, the smoothening by a window size of 71 pix (points) through the polynomial (2.1b) is applied by MLS and MLTS. The MLS curve twists accounting for the intensities of the lines, but the MLTS curve follows confidently the line of the continuum.

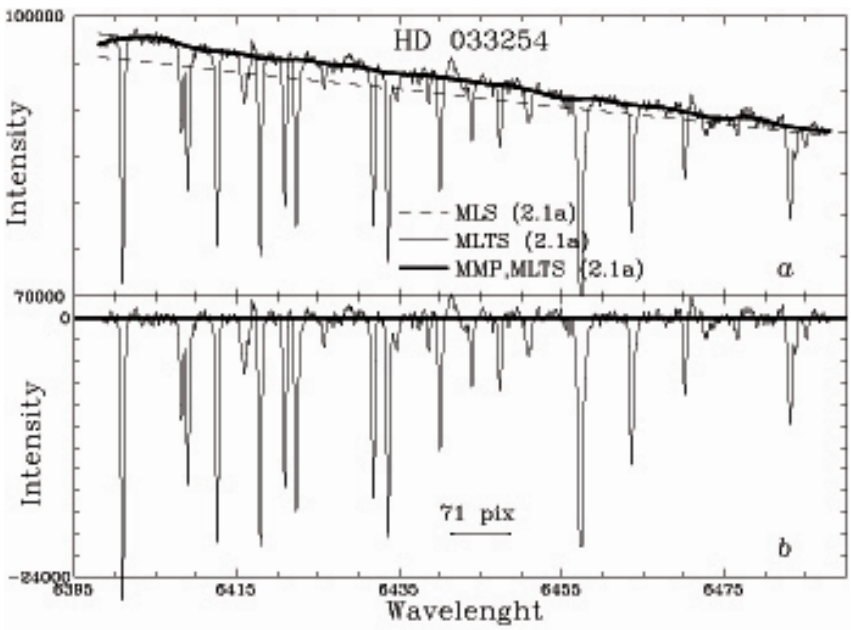

Fig. 7. a: Results of processing of a spectrum of the Am star HD 033254 by various methods, signed in the picture. In 2 cases, signed by "full", a regression line over all points is built. In the rest cases smoothing window size of 71 pix (7.1 A) is applied; b. Residual light curve with respect to the MMP smoothing by the MLTS (2.1b)

Figure 7*b* shows the residual spectrum with respect to the MLTS smooth and the equivalent widths of the spectral lines may be easy derived.

Figure 8 shows the central part of the spectrum, given in Fig.7. Smoothing with 2 different window sizes is applied and the results are practically identical. These examples show that the window size is not too crucial.

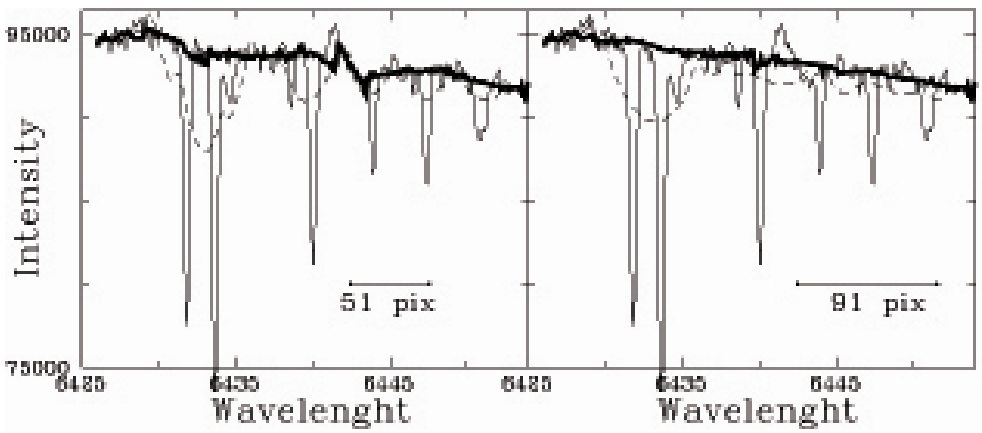

Fig. 8. Smoothing of the central part of the spectrum of HD 033254, given in Fig.7a by the MMP (2.1b) through MLS (dashed curve) or MLTS (thick curve) with a window of 51 pix or 91 pix. In both cases the MLTS (2.1b) follows the majority of the points, which are placed in the band of the continuum

Figure 9 shows an attempt for tracing the continuum in the complicated spectrum of the star HD 178449 with 900 points (2 m telescope of the *Rozhen* NAO, [15]). A MMP smoothing with a window size of 401 pix is applied through the MLS or MLTS. The MLS polynomial follows the middle part of the band of the data. On the contrary, the MLTS attempts to find and to follow the trend of some majority of points. This attempt is about to be successful up to 6000 A, but the right tail of the data is too short and the derived trend occurs broken. Essentially, this attempt for tracing some spectral continuum is not successful.

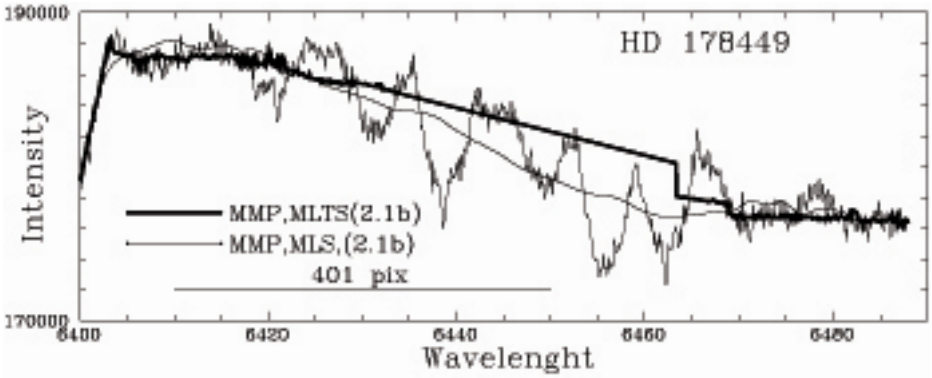

Fig. 9. Results of smoothing the spectrum of the star HD 178449 by application of a very large window of 401 pix (points). The right edge of the MLTS curve is broken because the spectrum is complicated and short

Figure 9 shows the significant difference between the results of MMP smoothing by the MLS or by the MLTS. It elucidates also the fact that the result of the MLTS smoothing cannot be easily predicted.

In the end, note that in the last example the computing time for the MLTS smoothening by means of the C-code (3.2) with the implementation of the algorithm (3.3) took about 10 min (roughly one second per data point), while when applying of the direct method (3.1) only the computing time should be about 100 times larger.

Conclusion

The main known advantage of the MLTS (Rousseeuw, 1984; Rousseeuw &, Leroy, 1987) in comparison with the MLS is its extremely high robustness with respect to outliers. Really the MLTS is able to ignore up to about 40% of the data, providing with a "mode" regression model. However, this method is not widely spread even when strong outliers are present because of its extremely high time consumption. This is understandable. When the data amount is not large, the user is able to reject the outliers that cannot be taken into consideration and to apply the ordinary MLS. However, in the case of many outliers or of many consecutive applications of a chosen regression model in the presence of outliers, the MLTS may be recommended. Apart from the examples given here, the MLTS may be useful in image processing and galaxy photometry.

We must note that we call the MLTS estimation to be “mode” but it is not just the mode, it must be slightly shifted toward the ignored large deviations. We consider this shift is very small.

Acknowledgements. The author is thankful to Dr. A. Antov and Dr. I. Kh. Iliev for the kindly allowed data, as well as for the valuable discussions about the text and the context of this paper.

References

1. A n d e r s o n, T. W., The statistical analysis of time series. 1971. New York: John Wiley & Sons
2. Br I l l I n g e r, D. R., Time series: Data analysis and theory. 1975. New York: Holt, Rinehart. & Winston.
3. O t n e s, R. K., E n o c h s o n, L., Applied time series analysis. 1978.,New York: John Wiley & Sons
4. S h u m w a y, R. H., Applied statistical time series analysis. 1988. Englewood Cliffs, NJ: Prentice Hall.
5. H e a s l e y, J. N., Publ. Astron. Soc. Pasific 96, 1984, 767
6. G e o r g i e v, Ts. B., Bull. Spec. Astrophys. Obs. 39, 1996, 131
7. T u k e y, J. W., Exploratory Data Analysis. 1977. Addison Wesley Publ. Co.
8. R o u s s e e u w, P. J., J. Am. Stat. Assoc. 79, 1984, 871
9. R o u s s e e u w, P. J., L e r o y A. M., Robust Regression and Oulier Detection. 1987. John Willy & Sons
10. G e o r g i e v, Ts. B., Bulg. Astron. J. 10, 2008. 93
11. D i m i t r o v, D., Private communication. 2007
12. T s v e t k o v a, S., B o e v a S., Bulg. Astron. J. 12, 2009, 4 3
13. A n t o v, A., Private communication. 2012
14. B o g d a n o v s k i, R., K o n s t a n t I n o v a – A n t o v a R., Private communication. 2013
15. B u d a j, J., I l i e v, I. Kh., Mon. Not. Roy. Astron. Soc. 346, 2003, 27-36

РОБАСТО ИЗГЛАЖДАНЕ НА РЕДОВЕ ОТ ДАННИ ЧРЕЗ ПРОСТ БЪРЗ АЛГОРИТЪМ. ПРЕКАРВАНЕ НА ТРЕНДА ПРИ ЗВЕЗДЕН ФЛИКЕРИНГ И КОНТИНУУМ ПРИ ЗВЕЗДЕН СПЕКТЪР

Цв. Георгиев

Резюме

Методът на движещия се прозорец от данни се използва широко при трасиране на едромащабното поведение на времеви редове, като оценката на централната точка на прозореца се базира на Метода на

най-малките квадрати (МНК). Обаче, обичайният МНК минимизира разсейването на всичките n квадрати на отклонения и затова е екстремално чувствителен към силно отклоняващи се данни. Една алтернатива е Методът на отбраните най-малки квадрати (МОНК) на Русю. Той минимизира само лявата част на квадратите на отклоненията, наредени по нарастване, включвайки поне $h = n/2+1$ данни. В дясната част на наредените квадрати на отклоненията може да присъстват произволно големи квадрати на отклонения, но МОНК ги игнорира. Така МОНК има асимптотична 50 % робастност спрямо силно отклоняващи се данни, докато робастността на МНК е определено 0 %. Обаче, докато коефициентите на обичайната МНК регресия се изчисляват чрез аналитично изведени формули, при МОНК това става чрез тестване на достъпни образци на възможни решения. Такива са: В едномерния случай – всяка данна; В двумерния случай – правата през всяка двойка точки; В тримерния случай – равнината през всяка тройка точки и т.н. Образецът, който има най-малко МОНК-разсейване се избира за решение. Главният недостатък на МОНК е, че в 2D, 3D и т.н. случаи той се нуждае от огромно изчислително време за да провери всички достъпни образци. Това може да отнеме милион-милиард пъти повече компютърно време отколкото времето за изчисляване на обичайна регресия. В гази работа е представен (i) прост бърз алгоритъм, който пропуска съседни комбинации с прогресивно увеличаваща се стъпка и може да редуцира изчислителното време хиляда – милион пъти. Представени са и (ii) възможностите на МОНК при изглаждане на редове от данни в два примера – за трасиране на кривата на блясъка на звезда в присъствието на избухвания и за прекарване на континуума на звезден спектър в присъствието на множество спектрални линии. Тук се имат предвид еквидистантни редове от данни но методът е приложим във всички случаи.