SMOOTHING SEQUENCES OF DATA BY EXTREME SELECTORS

Tertius de Wet and Willie Conradie
University of Stellenbosch, South Africa
tdewet@sun.ac.za

Non-linear smoothers based on the extreme selectors have been developed as a class with very powerful properties and ideally suited for application to data having impulsive noise, the type of data that often occur in the engineering and financial fields. Some of their properties make them ideally suited as a basis for teaching students about the art and science of data smoothing. These include inter alia their treatment of blockpulses of particular lengths as either signal or noise, its idempotency properties, which powerfully and visually demonstrate the mathematical concept of idempotence (which is often difficult for students to grasp) and the way that they systematically, measurably and monotonically "peel off" variation until one has a sufficiently smooth result. In this paper we define and discuss members of this class of smoothers and illustrate how their properties make them attractive aids in teaching aspects of nonparametric smoothing as well as aspects of Extreme Value Theory. A Standard and Poor 500 financial data set will be used for illustration purposes.

INTRODUCTION

Smoothing techniques are, broadly speaking, an approach to remove fluctuations from a time series with the purpose of uncovering patterns in the series, with a minimum of preconceptions and assumptions as to what those patterns should be. The goal is to separate the data into a smooth component (fitted curve, trend) and a rough component (residuals, noise), i.e.,

data = smooth + rough.

In the process of smoothing, the random error is reduced, thus making the variance of the smoother sequence small relative to the variance of the original sequence (Anderson, 1971). Smoothers fall into two basic categories: *linear* and *non-linear*. LULU smoothers is a class of non-linear smoothers introduced by Rohwer (1989), based on extreme selectors within moving windows. They have very powerful mathematical properties, are easy to understand and to compute and thus ideally suited as an aid to teaching students about smoothing. Furthermore, their distribution theory ties in naturally with Extreme Value Theory (EVT) and is thus a useful teaching application of the latter.

In this paper we will introduce the smoothers, discuss some of their mathematical and distributional properties and illustrate them using a financial data set.

DEFINITIONS

Let

$$x = \{..., x_{-3}, x_{-2}, x_{-1}, x_0, x_1, x_2, x_3, ...\}$$

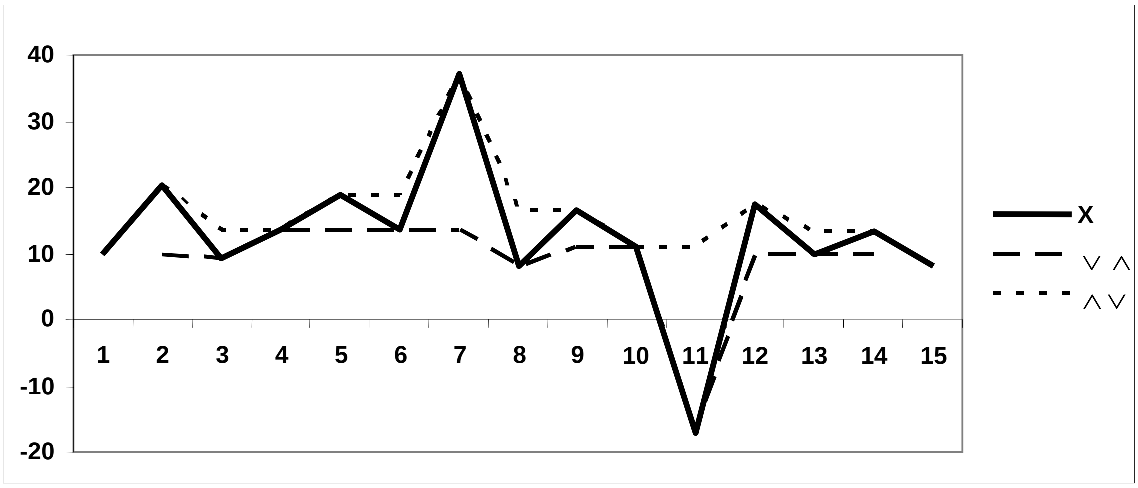
be a numerical sequence. LULU smoothers are compositions of the following two basic rank selectors operating on x. A forward operator $\vee^n$ is defined as:

$$\left(\vee^n x\right)_i = Max\left\{x_i, \ldots, x_{i+n}\right\}$$

and a backward operator $\wedge^n$ as

$$\left(\wedge^n x\right)_i = Min\left\{x_{i-n}, \ldots, x_i\right\}.$$

The operator $\vee^n$ is called *"forward"* since $\boldsymbol{x_i}$ is replaced by the *maximum* of x_i and the next *n* observations in the sequence and correspondingly $\wedge^n$ is called *"backward"* since x_i is replaced by the minimum of x_i and the *previous n* observations in the sequence. Clearly a single upward point will be removed by $\wedge x$ and a single downward point will be removed by $\vee x$. To remove both upward and downward isolated single points, the compositions $\wedge \vee$ and $\vee \wedge$ are needed. Similarly, sequences of consecutive upward and downward impulses of length n will be removed by $\wedge^n \vee^n$ and $\vee^n \wedge^n$. A simple illustration of this for *n*=1 is given below.



Denote these half smoothers as

$(U_n x)_i \equiv (\wedge^n \vee^n \; x)_i = Min\{\max(x_{i-n}, \; ..., x_i), \; ..., \max(x_i, \; ..., x_{i+n})\}$ *and*

$(L_n x)_i \equiv (\vee^n \wedge^n \; x)_i = Max\{\min(x_{i-n}, \; ..., x_i), \; ..., \min(x_i, \; ..., x_{i+n})\}$.

LULU smoothers are formed from compositions of L_n and U_n as building blocks. In particular, we can combine them as $U_n L_n$ and $L_n U_n$.

Now, an important aspect of many data analyses is to separate signal and noise in a data series. "Signal," however is a vague term and needs to be better defined in order to work with it in a sensible fashion. One approach to defining signal is in terms of so called blockpulses. Rohwer (2002) defines an *n*-blockpulse as a sequence

$$x = \{...0, b_1, b_2, ..., b_n, 0, ...\}$$

with $b_1 = b_2 = ... = b_n = b$ and infinitely many zeros on both sides. It is called upward if *b* is positive and downward if *b* is negative. In particular, blockpulses exceeding a certain length would then be considered signal and shorter blockpulses as noise. Since it follows quite easily that L_nU_n and U_nL_n remove blockpulses of increasing length as n increases, a natural smoother can be formed by applying these iteratively, for $n \geq 1$. Thus in order to smooth from above we form a flooring smoother

$$F_n \equiv U_n L_n U_{n-1} L_{n-1} .... U_2 L_2 U_1 L_1 = \left(U_n L_n\right) F_{n-1}$$

and to smooth from below, we form a ceiling smoother

$$C_n \equiv L_n U_n L_{n-1} U_{n-1} .... L_2 U_2 L_1 U_1 = \left(L_n U_n\right) C_{n-1}.$$

It can be proved that for each *n*: $U_n L_n \leq F_n \leq C_n \leq L_n U_n$, in other words the interval $\left[F_n, C_n\right]$ narrows the ambiguity of $\left[U_n L_n, L_n U_n\right]$ (see e.g., Rohwer, 2005). These smoothers provide a smoothed band that is useful in itself or may be used to construct a final smoothed sequence.

SOME MATHEMATICAL PROPERTIES

The LULU smoothers have a number of extremely attractive mathematical properties (see e.g., Rohwer, 1989; Rohwer and Toerien, 1991; Rohwer, 2005) – we mention only the following two.

Idempotence as well as *co-idempotence* hold for the LULU smoothers, i.e., for example for $L_n U_n$, we have:

$$\left(L_n U_n\right)^2 = L_n U_n \text{ and } \left(I - L_n U_n\right)^2 = I - L_n U_n.$$

Note that idempotence means that there is no "noise" left in the smoothed data and co-idempotence means that there is no "signal" left in the residual.

A second property we mention is with respect to the *handling of variation*. Since the purpose of smoothing is to reduce the variation in the data, it is extremely useful if a smoother does this in an ordered and measurable fashion. LULU smoothers have this property with respect to *total variation*. For a sequence $x \in l_1$, the latter is defined as

$$T(x) = \sum_{i=-\infty}^{\infty} | x_{i+1} - x_i |.$$

A LULU smoother P then has the property

$$T(x) = T(Px) + T(x - Px).$$

We can iterate on this, viz for P_1 and P_2 both LULU smoothers, it follows that

$$\begin{aligned} T(x) &= T(P_1 x) + T(x - P_1 x) \\ &= T(P_2 P_1 x) + T(P_1 x - P_2 P_1 x) + T(x - P_1 x). \end{aligned}$$

Consider for example the LULU ceiling smoother C_n for which

$$C_n = (L_n U_n) C_{n-1}.$$

Applying the above therefore gives

$$\begin{aligned} T(x) &= T(C_{n-1} x) + T(x - C_{n-1} x) \\ &= T(C_n x) + T(C_{n-1} x - C_n x) + T(x - C_{n-1} x). \end{aligned}$$

This variation reduction property is clearly extremely useful for applications. It gives a measure of the amount of variation "peeled off" during each successive application of a smoothing operation and is a diagnostic tool to decide when one could terminate the smoothing process.

DISTRIBUTIONAL PROPERTIES

It is not too difficult to obtain the exact distribution of LULU smoothers. For example, consider $L_n U_n$. Let $..., X_{-2}, X_{-1}, X_0, X_1, X_2, ...$ be a sequence of i.i.d. random variables with distribution function F_X. Denote by $F_{L_n U_n}$ the distribution function of $L_n U_n$ based on this sequence. Then, for $n = 1, 2, ...,$

$$\begin{aligned} F_{L_n U_n}(x) = F_X^{n+1}(x) + n(1 - F_X(x)) F_X^{n+1}(x) + (1 - F_X(x)).F_X^{2(n+1)}(x) \\ + \frac{1}{2}(n-1)(n+2)(1 - F_X(x))^2 F_X{}^{2(n+1)}(x), \end{aligned}$$

thus a polynomial in F_X. Also, for U_n we have

$$F_{U_n}(x) = F_X^{n+1}(x) + n(1 - F_X(x)) F_X^{n+1}(x).$$

Similar results hold for $U_n L_n$ and L_n.

Using these results and EVT, we can find the limiting distributions of $L_n U_n$ (respectively $U_n L_n$) and U_n (respectively L_n), as $n \to \infty$. These results show that U_n has the same limiting distribution as the second largest order statistic and that $L_n U_n$ lies asymptotically between the second and third largest order statistics. See Conradie, de Wet and Jankowitz (2006) for these results.

ILLUSTRATION

We illustrate the variation reduction property of LULU smoothers by applying it to Standard and Poor 500 data for the period 4 January 1999 to 3 October 2000.

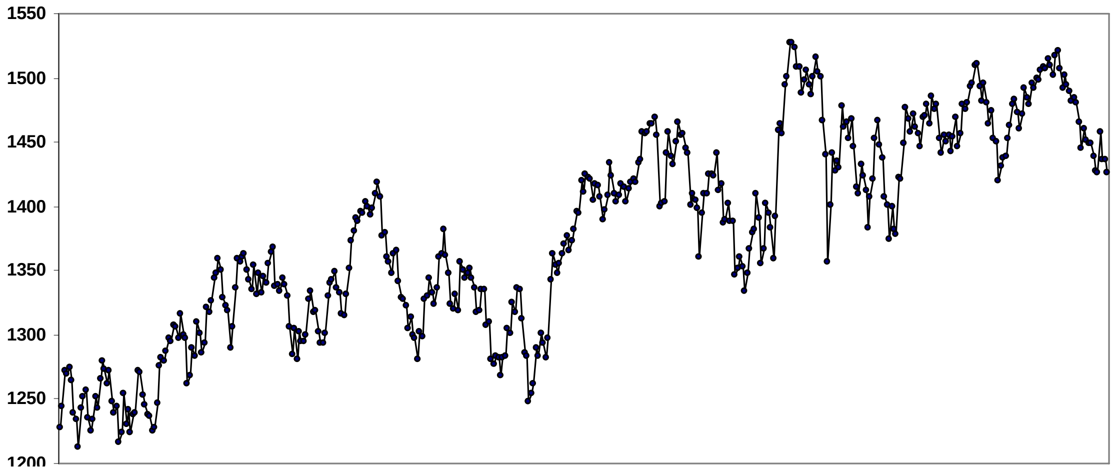


Figure 1: Standard and Poor 500 Data for the period 4 January 1999 to 3 October 2000

The total variation for this data set is:

$$T(x) = \sum_{i=-\infty}^{\infty} | x_{i+1} - x_i | = 5085.45$$

Smoothing by $L_1U_1 = C_1$ and $U_1L_1 = F_1$, using the decomposition $T(x) = T(C_1) + T(x - C_1)$, respectively $T(x) = T(F_1) + T(x - F_1)$, gives 5085.45 = 2835.22 + 2250.23, respectively 5085.45 = 2875.14 + 2210.31, or, as percentages, the decompositions give 100% = 55.75% + 44.25%, respectively 100% = 56.54% + 43.46%. The smoothed sequences are given in Figure 2 below.

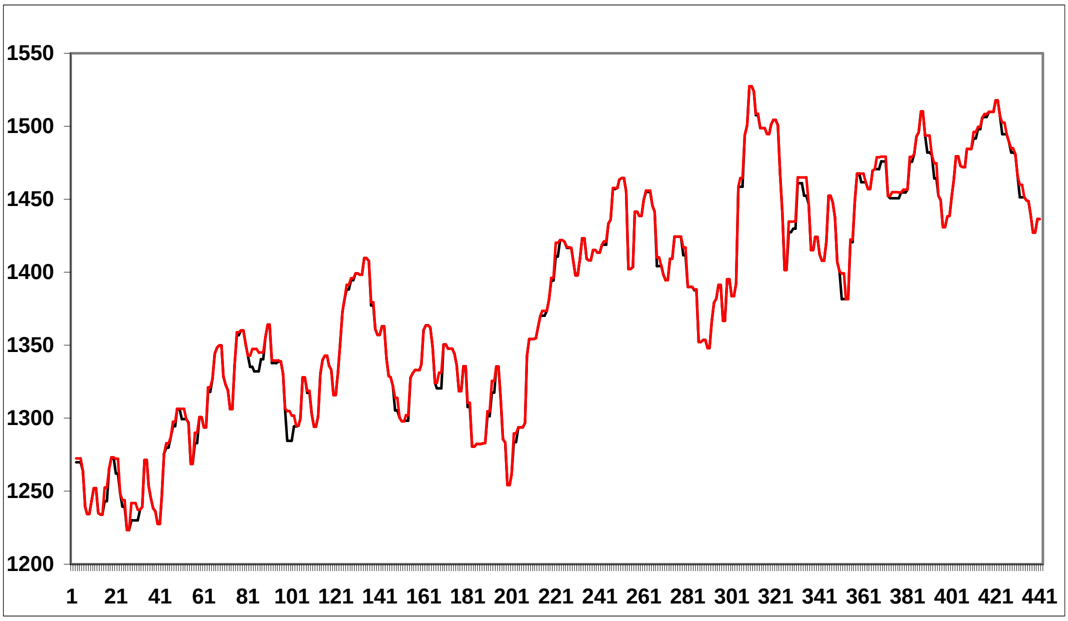


Figure 2: Data smoothed by $L_1U_1 = C_1$ and $U_1L_1 = F_1$. Where there is only one line, they coincide, otherwise C_1 denotes the upper line and F_1 the lower line

Now, if we apply the smoother L_2U_2 to C_1 (obtaining C_2), we obtain the further decomposition $T(C_1) = T(C_2) + T(C_1 - C_2)$, resulting in 5085.45 = 2079.8 + 755.42 + 2250.23. In percentages this is 100% = 40.90% + 14.85% + 44.25%. This means that a further 14.85% has been removed from the smoothed sequence obtained in the first iteration. Further smoothing (by applying L_3U_3 to C_2) does not produce any significant further reduction in the "smooth," indicating that further smoothing is not needed. The results for F_2 are very similar. The smoothed sequences C_2 and F_2 are given in Figure 3 below.

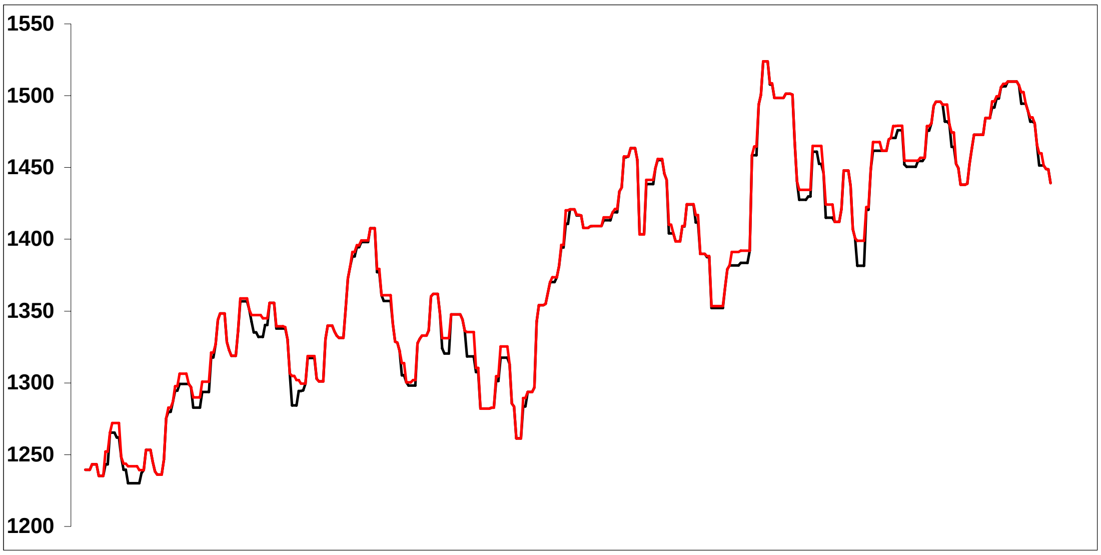


Figure 3: Data smoothed by C_2 and F_2. Where there is only one line, they coincide, otherwise C_2 denotes the upper line and F_2 the lower line.

CONCLUSION

This paper defined and discussed a class of non-linear smoothers based on extreme selectors with some very attractive properties. These properties can be used to illustrate some mathematical concepts in a straightforward fashion to students. Furthermore, the smoothers are easy to compute, even with standard spreadsheet software and students can thus easily implement them.

REFERENCES

Anderson, T. W. (1971). *The Statistical Analysis of Time Series*. New York: John Wiley Inc.

Conradie, W. J., de Wet, T. and Jankowitz, M. (2006). Exact and asymptotic distributions of LULU smoothers. *Journal of Computational and Applied Mathematics,* 186, 253-267.

Rohwer, C. H. (1989). Idempotent one-sided approximation of median smoothers. *Journal of Approximation Theory,* 58, 151-163.

Rohwer, C. H. and Toerien, L. M. (1991). Locally monotone robust approximation of sequences. *Journal of Computational and Applied Mathematics,* 36, 399-408.

Rohwer, C. H. (2002). Natural alternatives for one dimensional median filtering. *Quaestiones Mathematicae*, 25, 135-162.

Rohwer, C. H. (2005). *Nonlinear Smoothing and Multiresolution Analysis*. Basel: Birkhauser.