High-Fidelity Filter Based on Medians for Chemograms

Huajian Miao
SATech Software Engineering, Room 1001B/C, Zhongdian Mansion, No.1029, West Laoshan Road, Pudong, Shanghai, 200122, People's Republic of China

Abstract

A filtering method using the median in a moving data window and an algorithm avoiding level distortion on the signal peak top and valley bottom are proposed. The distinct feature of a median filter is its opposite effect on peaks with different widths. Peaks with small widths are removed thoroughly, whereas peaks with large widths are kept intact; thus the median filter can eliminate high-frequency noises effectively and keep the low-frequency signals nearly undistorted. Compared with other moving data window filtering methods (in particular, the widely used ones based on weighted sum), the median filter can remove spike noises more effectively because the median is not a compromise between noise and signal, and it is not affected by abnormal data or tendentious noises. The median filter can be connected in series, and its only parameter is easily set for a definite purpose. Keeping high fidelity to the width, height, and position of signal peaks, this method is especially feasible for use as a high-frequency noise eliminator in chemograms sampling data processing.

Introduction

In real-time processing of sampling data, a frequently used filtering method is based on a partial least-squares principle. It sets a window that contains n sequential sampling data and outputs the weighted sum as the smoothed one. The window is kept continuously moving so that the foremost datum in the window is moved out, and a new sampling datum is moved in as the last. In the course of the moving, a queue of smoothed data is generated by the window. Though having some smoothing effect, this kind of method may distort the signal. As usual, the more the signal-to-noise ratio increases, the more the signal is distorted (1). Moreover, the abnormal data that deviate greatly from the multitude cannot be rejected and are shared by the smoothed data. Consequently, the smoothing is severely affected. In order to overcome these limitations, the median, instead of the weighted sum, is adopted as the filtered output of the data window.

Algorithm Design

Suppose a peak whose width is k data points:
$...d_0, d_0, d_0, d_1, d_2...d_i...d_{k-1}, d_k, d_0, d_0, d_0...$

where d_0 is the baseline datum on both sides of the peak, and d_1, $d_2...d_i...d_{k-1}$ and d_k are peak data that ascend from the baseline to the peak top and then descend to the baseline.

The data window width n is set as $2m + 1$. When the data in the window are ranked by value, there are m data on both sides of the median that takes up the center point. When $k \leq m$, wherever the data window moved, the d_0 always takes up the center after the data in the window are ranked, so the k peak data will be filtered thoroughly.

When $k > m$, the m data on top of the peak are called peak top data, which are greater than the other peak data. The datum just below the m peak top data is assumed as d_z, which cannot be d_0 because the peak width is at least $m + 1$. When the window with a width of $2m + 1$ moves through the peak data, it contains all the uninterrupted m peak top data for $m + 1$ times; thus the medians output by the window are the same d_z for successive $m + 1$ times. Consequently, $m + 1$ data points on top of the peak are leveled out; this is called a level distortion in filtered data.

Through the above analysis, it can be concluded that if the median filter width is $2m + 1$, peaks for whose width $k \leq m$ are filtered thoroughly, and peaks for whose width $k > m$ are leveled on their tops. Similarly, this conclusion is suitable for negative peaks (i.e., valleys) as well.

For example, if it is assumed that peaks whose width are not greater than 2 are noise, m can be set at 2, and the data window width $n = 2 \times 2 + 1 = 5$. Then the noise taking up 2 data points in array 1 is filtered out, as is shown in array 2.

Original (array 1): 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 0, 0, 0, 0, 0, 0
Median (array 2): 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0

When the window with the same width is applied to array 3, in which a peak with five data points exists, three data points are leveled out on top of the peak, as is shown in array 4.

Original (array 3):	0,	0,	0,	0,	1,	2,	3,	2,	1,	0,	0,	0, 0
Median (array 4):		0,	0,	1,	2,	2,	2,	1,	0,	0		
p (array 5):		1,	2,	1,	2,	4,	2,	–1,	–2,	–1		
r (array 6):		2,	1,	2,	1,	0,	1,	2,	1,	2		
Modified median (array 7):		0,	0,	1,	2,	3,	2,	1,	0,	0		
Average (array 8):		0.2,	0.6,	1.2,	1.6,	1.8,	1.6,	1.2,	0.6,	0.2		

The level on the peak top or valley bottom makes signals distorted after filtering, so it is necessary to renovate the median filter. First, whether the window center datum is on the peak top or the valley bottom is evaluated according to the calculation of parameter p as follows, where d_i is a datum in the data window:

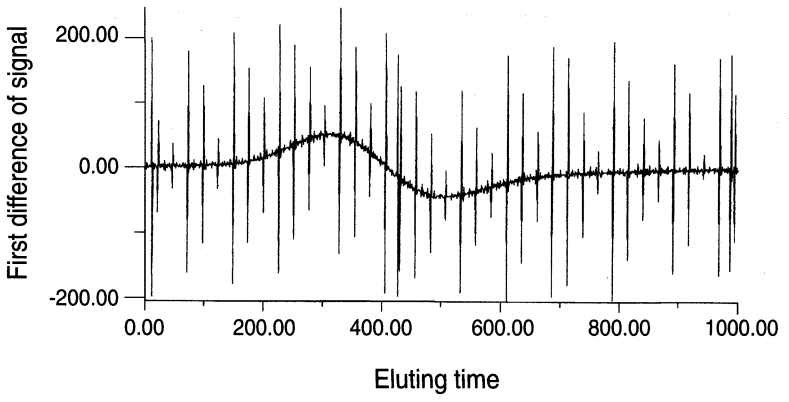

Figure 1. First difference curve of a chromatogram.

1. p is initialized to 0.
2. Index i is increased from 0 to $m - 1$. If $d_i < d_{i+1}$, then $p = p + 1$; if $d_i > d_{i+1}$, then $p = p - 1$.
3. Index i is increased from m to $2m - 1$. If $d_i > d_{i+1}$, then $p = p + 1$; if $d_i < d_{i+1}$, then $p = p - 1$.

According to this algorithm, if the center datum in the window is just on the peak top or the valley bottom, $p = 2m$ or $-2m$. If the data in the window strictly ascend, strictly descend, are completely equal, or are regularly rugged, $p = 0$. Between the above two extreme cases, the absolute value of p is greater than 0 and less than $2m$.

Second, in the data window, the data participating in filtering is revised as r data on both sides of the center datum and the center datum itself, and the variable r is called the filtering radius in the data window. After the evaluation of p, the filtering radius r can be calculated with the following formula:

$$r = m - |p|/2$$

Therefore, r is a variable changing with p. Thus, the number of data participating in filtering in the data window is no longer constantly $2m + 1$ but changeable as $2r + 1$. Under ideal circumstances, the above formula guarantees that the nearer the data window is to the peak top or the valley bottom, the smaller the filtering radius r in the data window gets. When the center data in the window are just on the peak top or the valley bottom, $r = 0$, and the data participating in filtering in the data window are only those in the center, so the output of the window is just the datum of the peak top or the valley bottom. Thus, the peak top or valley bottom is kept intact, and the level distortion on the peak top or valley bottom is avoided. On the contrary, if the center data in the window are not on the peak top or valley bottom, $r > 0$, and the further the center of the window departs from the peak top or valley bottom, the bigger r gets; it reaches its maximum m when the data in the window strictly ascend, strictly descend, are completely equal, or are regularly rugged. Therefore, when the window moves on either side of the peak or on the baseline, in ideal circumstances, the data in the window take part in filtering altogether, so the window still works as a normal median filter.

To illustrate the revised principle of median filter, the data in array 3 are filtered again. The values of the two parameters p and r in the course of filtering are shown in arrays 5 and 6 respectively, and the filtering result is shown in array 7. From array 7, no more level distortion can be seen generated on the filtered peak top, and the filtered data coincide with their originals completely.

However, the revised principle changes the width limits of the thoroughly removed peak and the completely reserved peak. It can be proved that, if the width of the data window is set to be $2m + 1$, peaks whose widths are not greater than $2m/3$ will be removed thoroughly, and peaks whose widths are not less than $2m - 1$ will be reserved completely. Peaks whose widths are between the two above limits will only have reduced height and width (i.e., be partially filtered).

Important Features

The median filter belongs to the kind of filter that uses the frequency difference of noise and signal. Its distinct feature has the opposite effect on peaks with different widths. As is proven above, it can thoroughly remove peaks with narrow widths and completely reserve peaks with large widths. It can therefore filter high-frequency noises while keeping low-frequency signals intact; it can filter just the spikes on a signal peak while hardly affecting the signal peak itself. Of course, keeping the peak width and height from distortion is especially important in the quantitative analysis of chemograms data processing. Other conventional filtering methods based on intensity smoothing, such as the one based on weighted sum, do not have this feature. Even if no noise exists, they still distort the real signal. For example, when the 5-datum window that outputs an equally weighted sum (i.e., average) is applied to array 3, the filtering result is shown in array 8. Compared with the original data, the filtered peak is distorted not only in height but also in width.

On the other hand, the median filter keeps fidelity not only to the signal peak width and height but also to its position. Evidently, when the data values in the window are strictly ascending or strictly descending, the output of the window is just the center datum, so there is no shift between the filtered data and the original ones. Of course, this lack of position shift is very important in many instances, such as the qualitative analysis in chemograms data processing.

Taking advantage of the high fidelity, the median filter can filter the filtered data over and over again. Those only partially filtered noise peaks will shrink both in width and height every time the median filter is used and will eventually be thoroughly removed. Finally, no changes take place in the filtered data, and the noises whose widths are less than some definite limit are thoroughly cleared. In this way, filtered data in which high-frequency noises scarcely exist are obtained. On the contrary, the more the method based on weighted sum is repeatedly used, the more the original signal is distorted. To

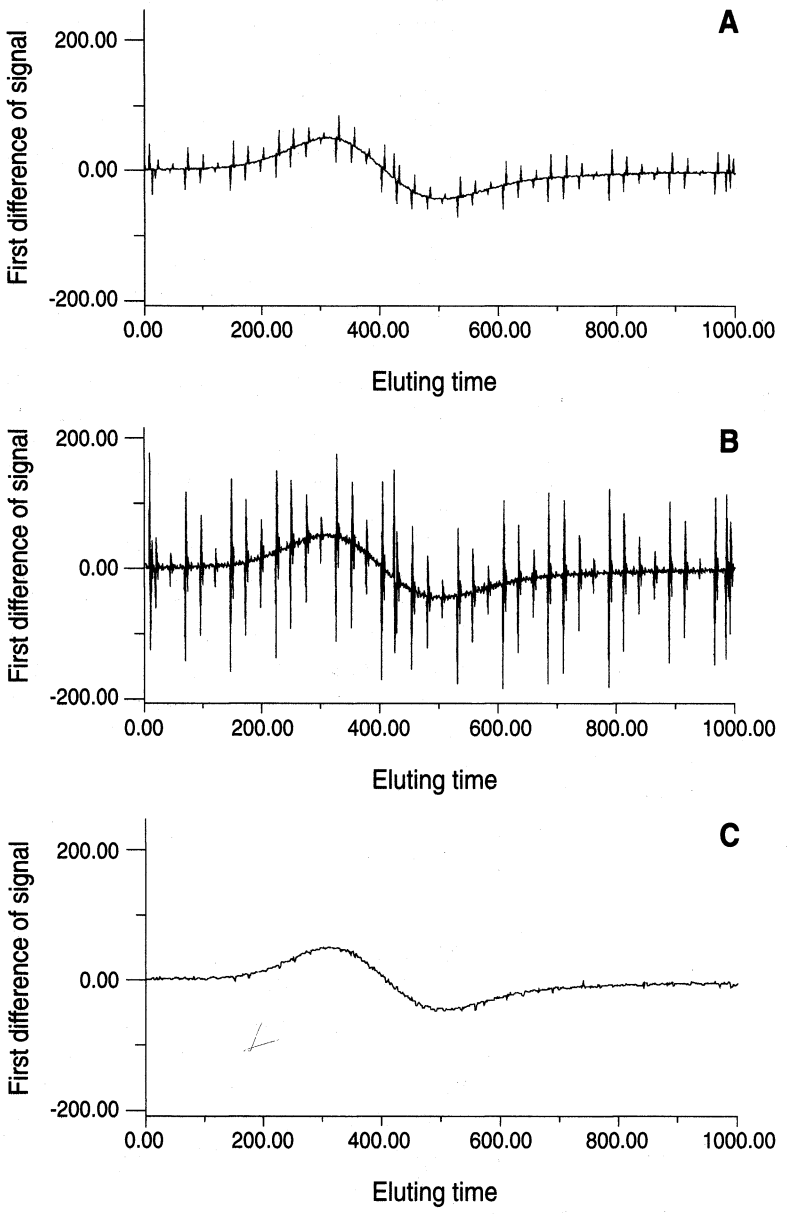

Figure 2. Effect of using different kinds of filtering methods on the data shown in Figure 1. (A) Average-based method, (B) second order curve fitting, and (C) median filter.

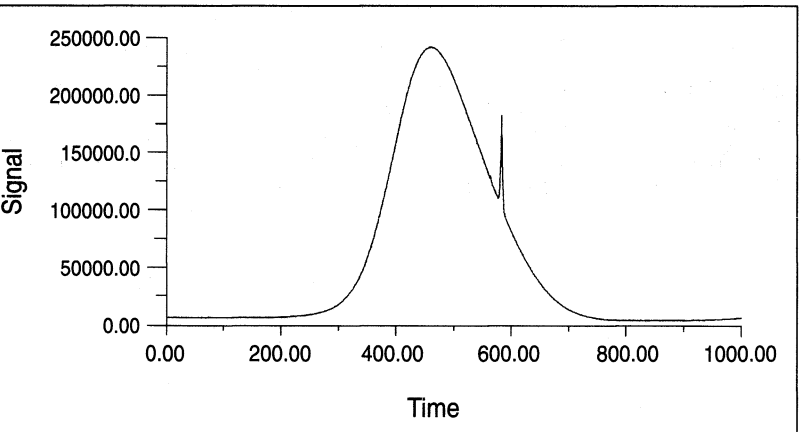

Figure 3. Two overlapping chromatographic peaks with very different sizes.

create the repeat filtering effect in only one scanning of the original data, the median filter can be connected in series (i.e., the output of the former filter can be used as the input of the latter one).

Apparently, the median filter has the inherent property of rejecting abnormal data and tendentious noises. These kinds of noises can be thoroughly removed in the course of filtering and do not exist in the filtered data. As for the filtering method based on weighted sum, the filtered data are only a compromise between signal and noise and, accordingly, are still influenced by the noises to some degree. Furthermore, the filtering method based on weighted sum makes the spike noise round and resemble a peak while making the real small peak even flatter. The rounded false peak and flattened real small peak will perplex the peak detection a great deal in chemograms data processing. On the contrary, the median filter will remove spike noise thoroughly and keep the real small peak intact (which is much wider than spike noise). That will favor the succeeding peak-detecting procedure, which is often disturbed by the spike noise, and further promote the small peak detection limit.

There is only one parameter, the width of the data window, whose effect is definite and should be set by the user. Like other conventional filtering methods, the wider the data window is, the smoother the filtered curve will be. Taking advantage of the property of high fidelity, the window width can be set as large as possible as long as the narrowest signal peak is not filtered, and the signal distortion is not a concern. Usually in the sampling data of chemograms, the noises are mostly electrical noises with high frequency, and the widths of the noise peaks are much different than the widths of the signal peaks, so the parameter of the data window is quite easy to set.

The algorithm of the median filter is not complicated. Though searching the median in the filtering radius range in the data window may take some time, it is no problem yet for the median filter to be applied in real-time processing. Furthermore, the median filter can be greatly sped up if some programming skills are used.

Application

The curve in Figure 1 consists of 1000 data that represent the first differences of chromatography sampling signals. From the figure, one can see many spikes with great amplitude

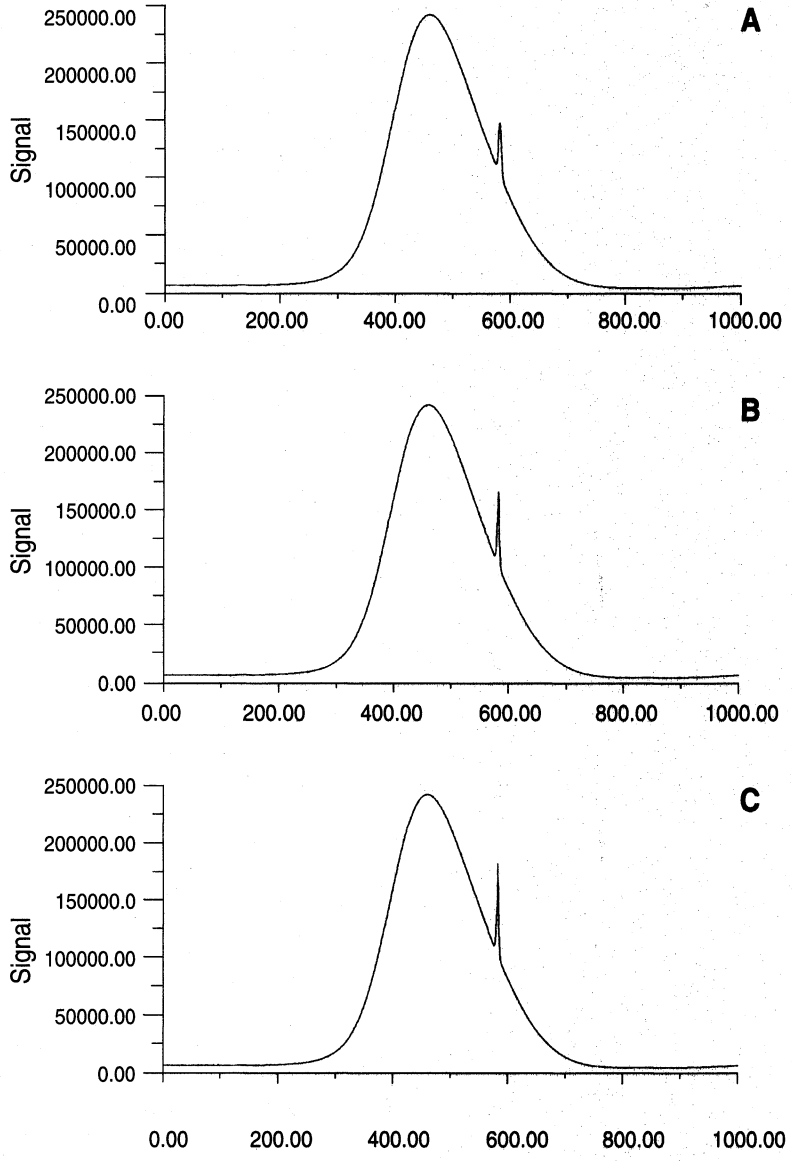

Figure 4. Effects of using different kinds of filtering methods on the data shown in Figure 3. (A) Average-based method, (B) second-order curve fitting, and (C) median filter.

on the curve. The data windows based on the average, second-order curve fitting, and median were used to smooth the data in Figure 1 on the condition that the window was set 5 data points wide. From the filtering effect shown in Figure 2, one can see that the median filter was the best.

Figure 3 shows the curve formed by two overlapping chromatographic peaks with very different sizes; the small peak takes up about 20 data points. The ideal filter should smooth the noise on the peak without affecting the small peak. The same three filters were used to smooth the data shown in Figure 3, and from the filtered curve shown in Figure 4, one can see that only the median filter met the requirement, whereas the other two filters distorted the small peak.

When the median filter is used as the on-line filter in chromatographic sampling data processing, the sensitivity of the peak-detecting method using parasecond-order derivatives (2) is further promoted. When the sampling data are severely interfered with by noise, even severely overlapping shoulder peaks are normally detected. Also, because the height and position of the inflection points on the first derivative curve of chromatography are hardly distorted by the median filter, the overlapping peaks deconvolving method aided by artificial neural networks (H. Miao and S. Hu. Chromatographic overlapping peak resolution by artificial neural network. *Chinese J. of Chem.*, in press.) can be put into practice.

Conclusion

The output of the median filter proposed in this paper is not a compromise between noise and signal, and thus is hardly affected by abnormal data or tendentious noises. Spike noise peaks are removed thoroughly, whereas signal peaks with large widths are kept intact. Ensuring peak size and position, which are two major parameters, to be the least changed, this high-fidelity filter is especially feasible for filtering chemograms from noises.

References

1. S.E. Bialkowski. Generalized digital smoothing filters made easy by matrix calculation. *Anal. Chem.* **61:** 1308–10 (1989).
2. H. Miao and S. Hu. Chromatographic peak detecting algorithm combining the first and second derivatives. *Chinese J. of Anal. Chem.* **22(3):** 247–50 (1994).

Manuscript accepted April 29, 1997.