Noise filtering:

the ultimate solution?

*Yuri Kalambet**
Sergey Maltsev

Ampersand Ltd., Kurchatov sq. 2, Moscow123182, Russian Federation
kalambet@ampersand.ru

Briefs

A noise filtering method that provides the lowest possible confidence interval for every data point.

Abstract

A method of noise filtering based on confidence interval evaluation is described. In the case of the approximation of a function, measured with error by a polynomial or other functions that allow estimation of the confidence interval, a minimal confidence interval is used as a criterion for the selection of the proper parameters of the approximating function. In the case of the polynomial approximation optimized parameters include the degree of the polynomial, the number of points (window) used for the approximation, and the position of the window center with respect to the approximated point. The Method is demonstrated using generated and measured chromatograms. The special considerations on confidence interval evaluation and quality of polynomial fit using noise properties of the

data array are discussed. The Method provides the lowest possible confidence interval for every data point.

Introduction

Any measurement contains a signal portion and random error caused by the electronics utilized, variation of ambient conditions, radio interferences, etc. This error should be diminished as much as possible to achieve the best estimate of the measured signal. There are many methods of noise reduction, both linear (moving average, Gaussian, Savitzky-Golay [1], Fourier transform-based) and nonlinear (median filtering)[2]. However, most of these methods change the shape of the object, e.g., a peak in chromatography or capillary electrophoresis may change its shape after noise filtering, and the better the noise reduction is at the baseline, the more significant change of the peak shape is observed.

Novel linear methods have emerged, such as noise reduction, based on wavelet transform [2]. They do not provide a final solution either. The main problem in all of the methods is a lack of clear-cut quantitative criterion of the filtering quality.

On the other hand, we can approximate our data set with a moving polynomial (similar to the method of Savitzky and Golay [1]) and calculate the confidence intervals for every approximation. Theory of confidence intervals is a well-established technique, widely used in the calibration of systems of different nature. For single-dimensional data, the confidence interval can be estimated with [3]:

$$C_Y = t_{n-p}^{½\alpha} \cdot S \cdot \sqrt{u_*} \qquad (1)$$

Here

$$S^2 = \frac{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})' \cdot (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n - p};$$

$$u_* = \mathbf{x}'_* (\mathbf{X}' \cdot \mathbf{X})^{-1} \mathbf{x}_*;$$

n - number of data points used for polynomial approximation (gap of the filter);

p - number of parameters of the polynomial (power +1);

$$\mathbf{X} = \left\{ \begin{matrix} 1 & x_1 & x_1^2 & \dots & x_1^{p-1} \\ 1 & x_2 & x_2^2 & \dots & x_2^{p-1} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_n & x_n^2 & \dots & x_n^{p-1} \end{matrix} \right\}$$ - matrix of values on independent axis (usually a time or position axis);

$\mathbf{Y} = \{y_1, y_2 ..., y_n\}$ - vector of detector response values;

$\mathbf{x}'_* = \{1, x_*, ..., x_*^{p-1`}\}$;

x_* - position at which smoothed (approximated) value is estimated;

$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \cdot \mathbf{X})^{-1} \mathbf{X}' \cdot \mathbf{Y}$ - Polynomial coefficients for regression;

t_m^{δ} - Student's coefficient for confidence probability (1-*δ*) and *m* degrees of freedom.

We applied confidence interval calculation principles to noise reduction task.

Algorithms

Filtering Algorithm for fixed window and degree of the polynomial

To simplify our task we will consider chromatographic data as measured with a constant data rate and we will not consider the case of re-sampling. Input array is just an array of raw data, and output array consists of the same number of data points and an estimate of confidence interval for every data point. An algorithm of noise filtering using confidence intervals works as follows:

1. Evaluate points and confidence intervals for all points within a selected window.

2. For all points within the window compare new confidence interval with that in the output array. If the new interval is smaller than stored in the output array or the point was not evaluated, replace stored point and its confidence interval.

3. Shift evaluation window and go to step 1.

So, every point of the chromatogram is approximated *n* times and an estimate with the best confidence interval remains as filtered value. Computational complexity of this simple Confidence filter is comparable with that of convolution, (e.g. Savitzky-Golay) and linearly depends on the product
(window width)·(degree of the polynomial).

Already this simplest implementation provides some benefits over the traditional Savitzky-Golay filter in several important cases: baseline step between two peaks (Figure 1); triangular peak originating from capillary electrophoresis; outlier point. The main benefit of the

Confidence filter in this case is that the points close to an abrupt change of signal level (step) are not disturbed by this step.

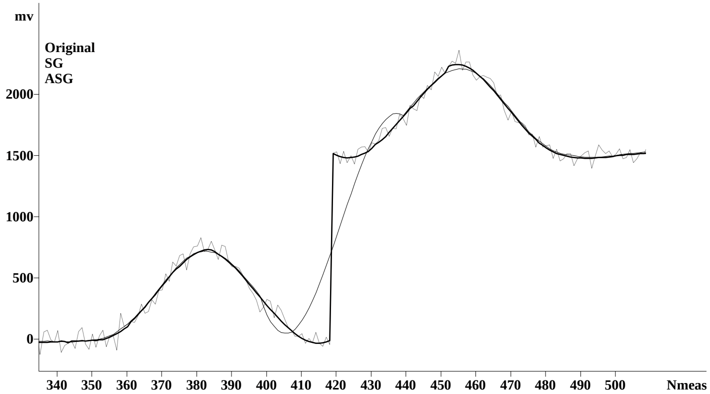

Figure 1. *Filtering with conventional Savitzky-Golay (SG, thin line) and Confidence adaptive non-central approximation filter (ASG, thick line). Original data are drawn with dotted line. The filter gap is the same in both cases and equals 41 points.*

Filtering Algorithm for variable window and degree of the polynomial

Obvious improvement of the simplest filter is changing the window and/or degree of the polynomial. Smaller windows are expected to give better estimates of steep slopes and bigger windows – better noise reduction for long baseline regions. However, small and large windows may lead to errors in filtering for different reasons. For small window we have a rather high probability of an accidental good fit of the polynomial, where confidence interval estimate using Formula 1 will give a too optimistic estimate. This error is caused by the fact that Formula 1 gets an estimate of experimental error from the small subset of the data array. Another problem exists in the case of large windows: decrease of confidence interval due to a large number of degrees of freedom may provide a formally very good confidence interval for rather poor approximation polynomial.

The solution for both problems can be easily provided, if we assume that we know parameters of the noise in our data array. That is, we assume that the noise is white, noise density probability is constant throughout the array and does not depend on measurement number or value, and noise standard deviation equals σ. As S^2 from formula 1 is an unbiased estimate of σ^2 [3], we can assume that all cases when S from Formula 1 is below σ are accidental and we should use σ for the estimate of confidence interval instead of S:

$$S_i = \{S,\ S > \sigma; \quad \sigma, S \le \sigma\} \tag{2}$$

Another criterion, based on known noise level, relies on the fact, that distribution of S has its own width, which quickly decreases with increasing size of the window [3]:

$$\mathrm{Var}(S^2) \sim \frac{1}{(n-p)}$$

Note, that the value

$$(n-p)\mathrm{Var}(S^2)$$

is a constant for all window widths and polynomial degrees properly fitting our data.

So, we select such a rejection coefficient *k* that the polynomial is treated as improper evaluation of the data array due to the wrong approximation model (window width and degree of the polynomial) if it satisfies condition:

$$S^2 \geq S_R^2 = \sigma^2 + k \cdot \sqrt{\mathrm{Var}(S)_{\mathrm{exp}ected}} = \sigma^2 + k \cdot \sqrt{\frac{\mathrm{Var}\left(S_n^2\right)\left(n_n - p_n\right)}{(n-p)}} \qquad (3)$$

where subscript n corresponds to the values originating from noise definition window.

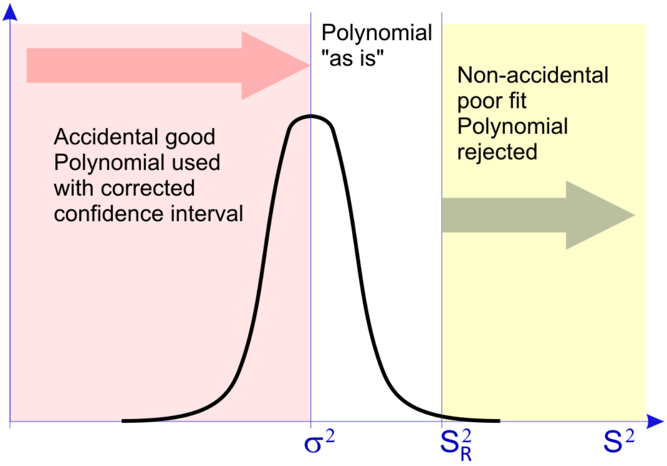
Polynomial "as is"
Non-accidental poor fit Polynomial rejected
Accidental good Polynomial used with corrected confidence interval
σ^2
S_R^2
S^2

Figure 2. *Distribution of dispersion for valid approximations and behavior of approximation procedure depending on S.*

Evaluating noise level

Now we have to find a way to estimate σ using our data array. The robust way of noise estimation was selected: user has to define width of noise window and a degree of the polynomial to approximate data using this window. Requirements to the window and degree are that 1) most of signal peculiarities, which are treated by the user as noise in his data are effectively suppressed if the array is filtered by Savitzky-Golay algorithm with this width and 2) most of the data array should be properly described by the polynomials of this window/power. Data array is approximated $3 \cdot L/n$ times, where L is the size of the data array and n – size of the noise definition window; every time window is shifted to higher indices W/3 points. S^2 value from each approximation is stored in the new array E, which is used for estimation of σ^2 and Var(S^2) in several look-through passes. During the first pass we calculate average and variance of values in E, on the second pass we accept only values, that pass the condition

$$E_i \leq \sigma^2 + 3 \cdot \sqrt{\mathrm{Var}\left(S^2\right)}$$

and re-calculate new σ2 and variance. This outlier rejection procedure is repeated until σ2 and variance stop changing, but not more than 5 times; it effectively rejects all outliers, originating from the regions with poor approximation of the data, such as the baseline steps, jumps, sharp peaks caused by sample injection.

Outline of the Confidence filtering algorithm:

1. Evaluate noise level using noise definition window width n and degree of the polynomial p-1, get σ2 and Var(S2) estimates;
2. Define a list of window widths and degrees of polynomial to be applied to filtering;
3. Fill output array with input data and confidence intervals with

   $t_1^{½\alpha} \cdot \sigma$

4. Select the first element of the list;
5. For all possible positions of the polynomial within input array: the approximate data within the window; evaluate S2; if S2 is too big and fits condition (3), skip position; if S2 is below σ2 replace S2 with σ2 (condition 2). For all points within a window compare new confidence interval (calculated with corrected S) with output value. If the new interval is smaller than stored in the output array, replace stored point and its confidence interval;
6. Select the next element of the list; if the list is complete stop filtering.

We selected to start implementation of the described noise filtering procedure with changing window at fixed (cubic) degree, δ=0.05 (corresponds to 95% confidence level) and k=2. To improve calculation speed, logarithmic steps were used, increasing or decreasing window √2 times with every step. In addition to filtering with the noise definition window, three steps were performed upwards and three downwards, increasing overall window width range to 8. Even this quite simple implementation provided excellent results, which are shown below.

Results

White noise

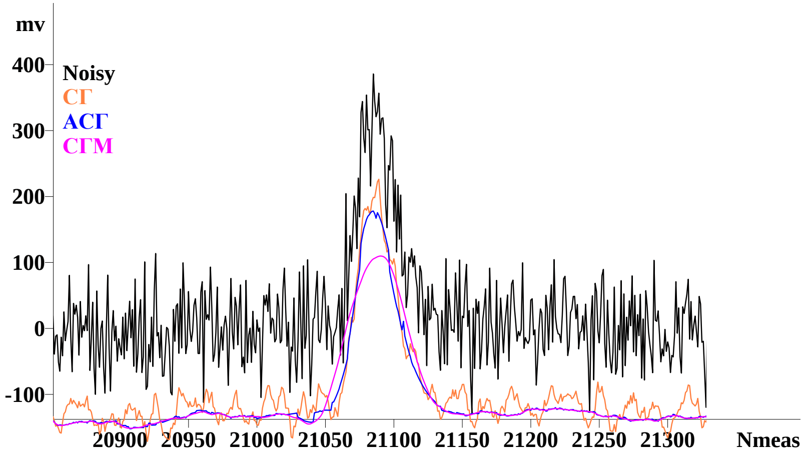

Figure 3. *Filtering artificial chromatogram of EMG [4] peak with white noise applied; black line – original data; blue line – Confidence filter with noise definition width of 31; magenta – Savitzky-Golay filter with width of 85 (corresponding to the maximum window width allowed for Confidence filter); light brown – Savitzky-Golay filter with width of 13 (corresponding to the minimum window width allowed for Confidence filter). Quality of baseline filtering corresponds to the widest window and peak shape does not change.*

Pump pulsations

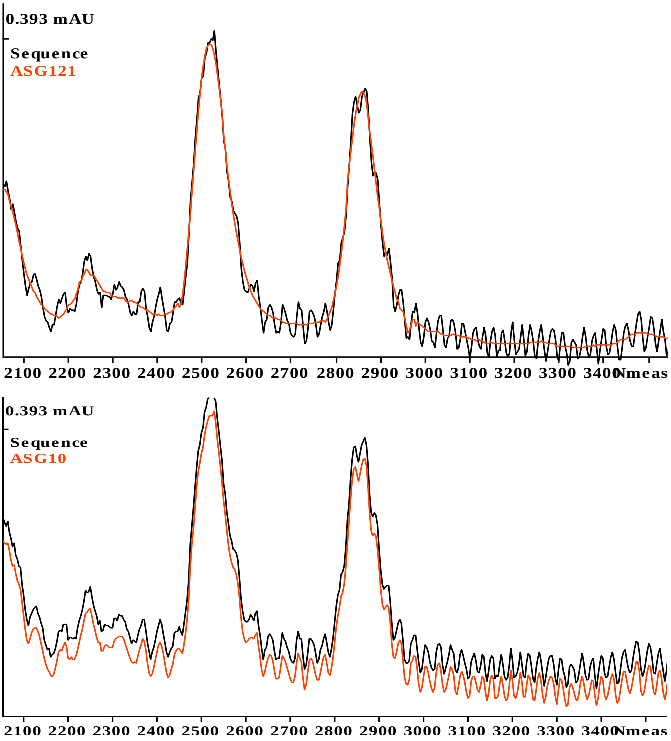

Figure 4. *a) Pump pulsations are effectively suppressed by Confidence filter using noise definition window width of 121 (light brown line). Black line – original data.*

b) When narrow noise definition window 11 (corresponding to half cycle of pump pulsation) is used, pump pulsations are not suppressed, just smoothed. Curves are shifted along Y axis to avoid overlapping.

Capillary electrophoresis

One of the most interesting fields for application of Confidence filter, as some CE peaks are very narrow and other triangular.

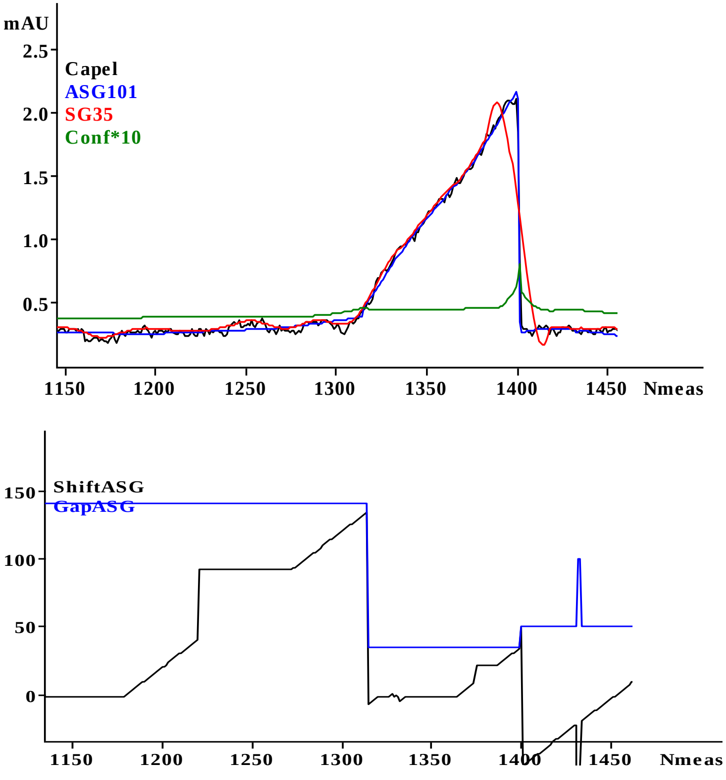

Figure 5. *Approximation of Capillary Electrophoresis peak (black line) with Confidence filter, width=101(blue line), and Savitzky-Golay filter (red line), width = 35 (corresponds to the smallest window of Confidence filter). In this case Confidence filter has undoubted advantage, as eliminates noise much better and makes much less*

disturbance near the steep end of the peak. Green line shows 10 times amplified confidence interval of the approximation by Confidence filter.

b) Behavior of the polynomial width (blue line) and of the distance of the point used for approximation from the center of the polynomial (black line), positive – to the right, negative – to the left. Note, that peak top is approximated by non-central approximation.

Discussion

Our implementation of the Confidence Filter does not reject outliers; rather it avoids them, leaving them as they are. It's rather easy to imagine a modified procedure, based on the robust regression methods [5], and outliers will be eliminated. However, there is a danger behind such approach, as robust regression may obscure model errors, so we currently prefer to deal with separate object-dependent outlier elimination procedures.

The confidence interval is a very natural criterion of approximation quality and it perfectly fits the case of noise filtering. In the case of variable window width and/or degree of the polynomial additional criteria based on noise estimate have to be applied to avoid effects of an accidental good fit for small approximation windows and peak suppression in wide windows.

The algorithm of the Confidence Filter very effectively suppresses baseline noise and significantly improves detection and quantification limits. Even non-white noise, such as pump pulsations or chemical noise can be suppressed; in addition the peak shape does not suffer. Peak metrology gets a chance to become a science, definitions of LOD and LOQ have to be re-considered using the confidence interval information.

References

1. Savitzky, A.; Golay, M.J.E. (1964). "Smoothing and Differentiation of Data by Simplified Least Squares Procedures". *Analytical Chemistry* **36** (8): 1627–1639.

2. Felinger A.; Data analysis and signal processing in chromatography / Data Handling in Science and Technology – v.21 ELSEVIER, 1998.

3. Linear Regression Analysis (Wiley Series in Probability and Statistics) by George A. F. Seber and Alan J. Lee (Feb. 5, 2003).

4. McWilliam, I. G.; Bolton, H. C., Instrumental Peak Distortion. I. Relaxation Time Effects, Anal. Chem. 1969, 41, 1755-1762.

5. Ricardo Maronna, Doug Martin and Victor Yohai, *Robust Statistics - Theory and Methods*, Wiley, 2006