Noise filtering: the ultimate solution

Yuri Kalambet, Yuri Kozmin*, Sergey Maltsev

Ampersand Ltd., Moscow, Russia

*Institute of Bioorganic Chemistry RAS, Moscow, Russia

Summary

The ultimate solution of the problem "Which smoothing method is better" is offered. 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 allows estimation of the confidence interval, the 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 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.

The Method is demonstrated using generated and measured data. Improvement of noise reduction compared to competing methods can vary depending on the input data, but always exists. Excellent noise reduction properties are combined with conserved object shape (e.g. chromatographic peak, photographic object) without artifacts. The method requires extra computations, which can be easily paralleled.

Theory

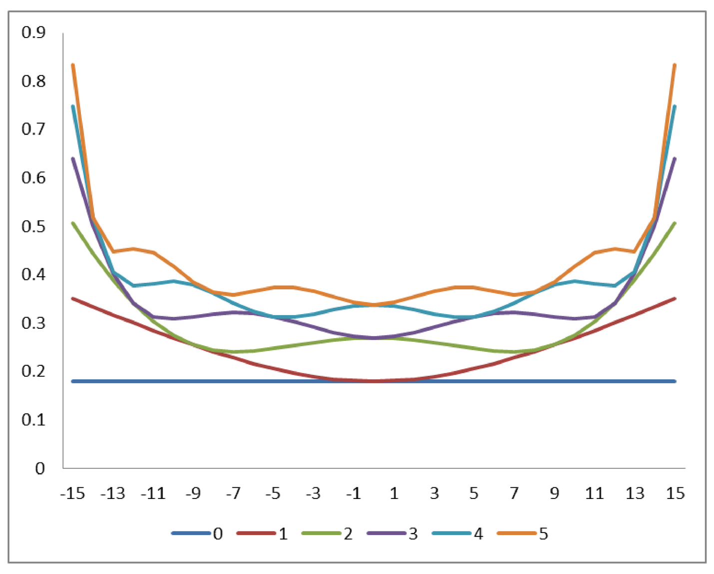
$$C_Y = t_{n-p}^{(1/2)\alpha} \cdot S \cdot \sqrt{u_*}$$

where $S^2 = \frac{(Y - X\hat{\beta})'(Y - X\hat{\beta})}{n-p}$ $\quad u_* = x'_*(X' \cdot X)^{-1} x_*$

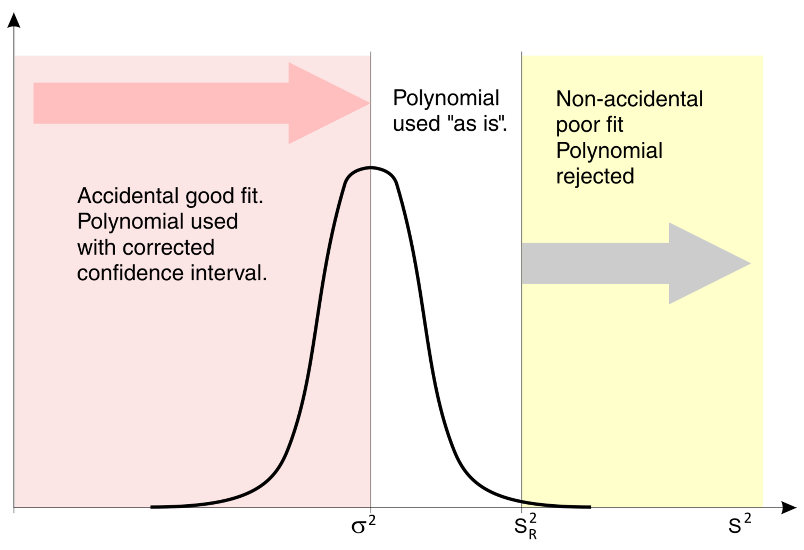
n - number of data points used for polynomial approximation (gap of the filter);
p - power of the polynomial;
$\mathbf{X}$ - matrix of x power values on independent axis (time);
$\mathbf{Y}$ - vector of detector response values;

$$\hat{\beta} = (X' \cdot X)^{-1} X' \cdot Y \qquad x'_* = \{1, x_*, \ldots, x_*^p\}$$

t_m^{δ} - Student's coefficient for confidence probability (1-δ) and m degrees of freedom
x_* - position at which smoothed (approximated) value is estimated.

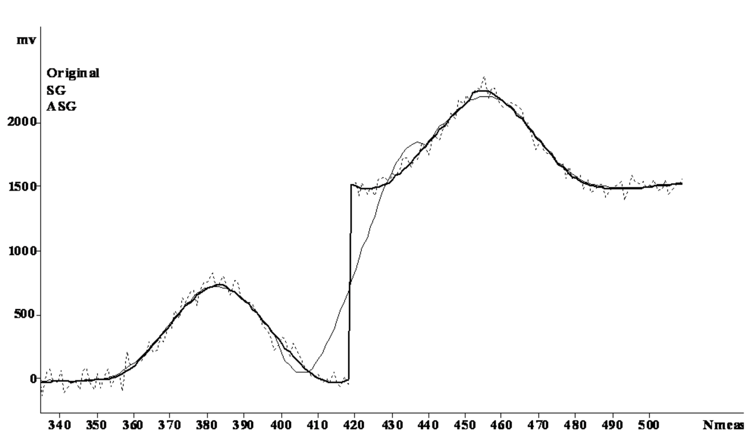


Profile of √(u*) term for the gap 31 and polynomial degrees from 0 to 5.



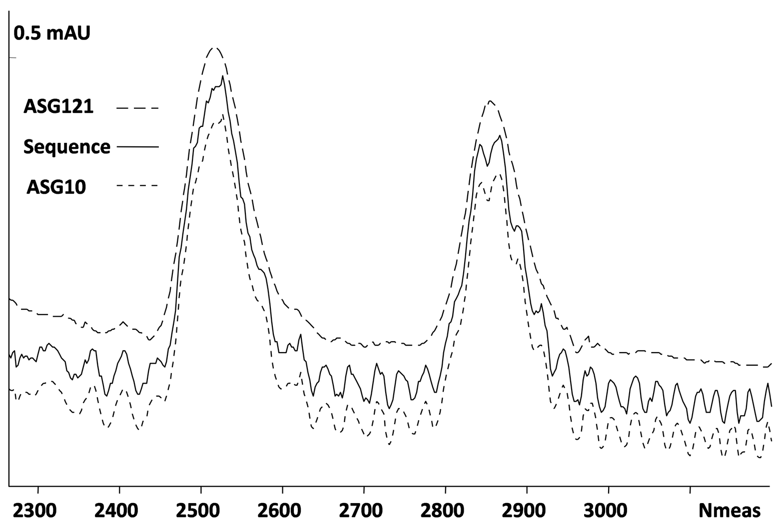
Modification of confidence interval estimation algorithm for particular polynomial depending on S^2 of this poynomial

Handling baseline steps



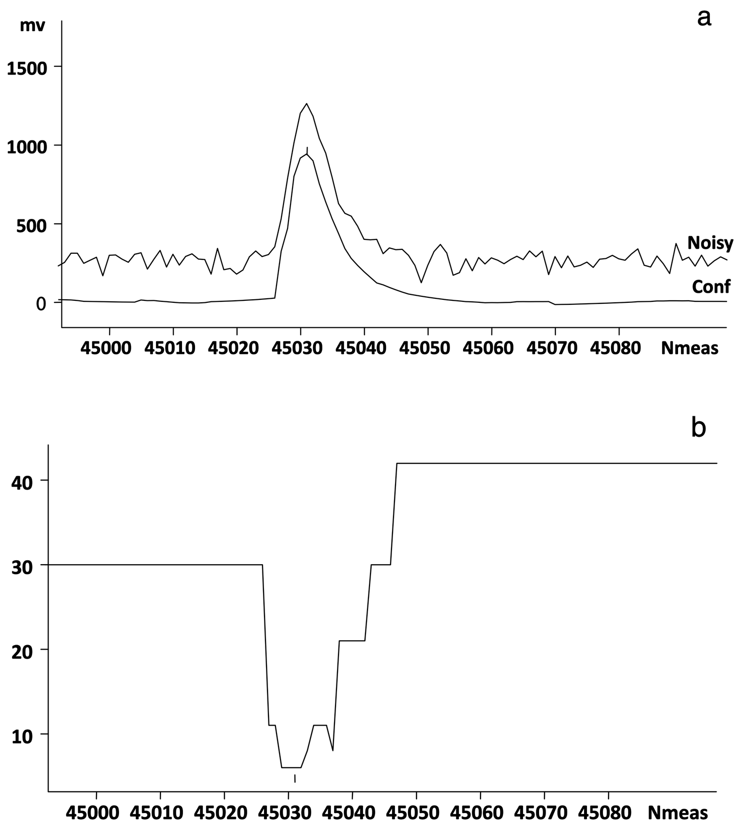
dotted – raw data; thick line – Confidence Filter; thin line – Savitzky-Golay filter

Non-white noise



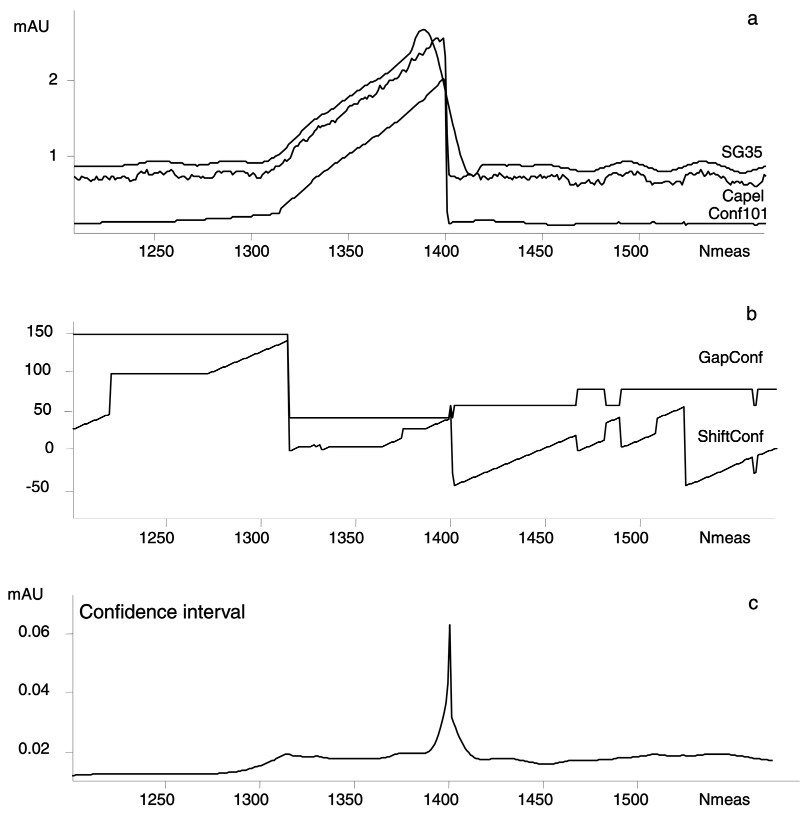
Pump pulsations are effectively suppressed by Confidence filter using noise definition window width of 121 (dashed line). Solid line – original data. When narrow noise definition window 11 (corresponding to half cycle of pump stroke) is used (dotted line), pump pulsations are not suppressed, just smoothed. Curves are shifted along Y axis to avoid overlapping.

Filtering white noise



a) Filtering artificial chromatogram of EMG [4] peak with white noise applied; solid line – original data (Noisy); dashed line – Confidence filter with noise definition width of 31; b) Approximation window size depending on window position.

Triangular peak



a) Approximation of CE peak (solid line) with Confidence filter (dotted line), and Savitzky-Golay filter (dashed line). b) Polynomial width (dashed line) and of the distance of the point used for approximation from the center of the polynomial (solid line). c) confidence interval of the approximation by the Confidence filter.

Signal to noise ratio of the peak

- **Area** $\quad C^2_{Area} = \sum_{Peak} C_i^2 + w \cdot \left(C^2_{Baseline}\right)$
- **Height** $\quad C_h^2 = C^2_{Y(x_h)} + C^2_{Baseline}$

Discussion

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 fit quality criteria based on noise estimate have to be applied to avoid effects of accidental good fit for small approximation windows and peak suppression in wide windows.

The algorithm of Confidence Filter very effectively suppresses baseline noise and significantly improves detection and quantification limits. Even non-white noise can be suppressed, such as pump pulsations, chemical noise; in addition peak shape does not suffer. No doubt, methods that utilize additional information about signal and noise, such as reconstruction of pump pulsation profile and subtraction of this profile from the signal before noise filtering, may give better results, but they will be much more expensive, as will require building a model of the process.

The math used to compute confidence interval (Formula 2) is based on assumption of homoscedastic noise. The case of heterocsedastic noise is much more complicated and in some cases can be avoided by scaling of the raw data. As an example of such scaling one can imagine taking square root of the signal before smoothing and then using square for final answer – in the case of any counting detector such a procedure has good theoretical base.

The Confidence algorithm can be extended to data of higher dimensions, e.g. 2-D, by approximation of smoothed data in different dimension. The heteroscedastic noise model in this case is a big benefit, as for the second dimension, regression may gain additional accuracy by accounting for different confidence intervals of data points.

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**