Estimation of Gaussian noise in spectra by the selective polynomial fit.

S. Charonov, Horiba France SAS, 455 Avenue Eugène Avinée, 59120 Loos, France

E-mail : serguei.charonov@horiba.com

Abstract.

This article describes an algorithm for estimation the variance of Gaussian noise. The data is smoothed using the Savitsky-Golay polynomial filter. Absolute differences between original and smoothed data are sorted in ascending order. The initial part of this sequence is selected for analysis. The result of calculation mean value of differences can be used to estimate the variance of the noise. By selecting points for analysis, the impact of cosmic ray noise and other artifacts can be reduced. The use of the proposed method for artificial and real spectra shows the ability to effectively estimate the noise variance. The algorithm contains no user-defined parameters.

Keywords.

Gaussian noise, variance estimation.

Introduction.

One of the important steps in data preprocessing is noise reduction. But often not only noise, but also signal features can be distorted or removed. A noise estimation is required to distinguish between noise and features, and to adjust the parameters of noise reduction algorithms. Many algorithms for this task have been invented (1-6). The proposed method is very simple, it can be applied directly without any analysis of the features of the data, and it is robust to a very wide data type. It can operate in the presence of cosmic ray noise, which is a contaminate factor for CCD spectral detectors. The method can also be used unchanged for 2D images and other multidimensional data.

Algorithm.

The algorithm includes the following steps.

1. The raw data is smoothed by a Savitsky-Golay polynomial filter (8) with a small window size and a small order. This work uses values of 5 for window and 2 for order.

2. The absolute difference values between filtered and raw data are sorted in ascending order. The mean is calculated for the initial part of this sequence. The variance of Gaussian noise is proportional to the mean.

It is known that spectra of a wide range of types are well approximated by the sum of the Gaussian and Lorentzian functions. These functions can be fitted by a polynomial with a small relative error.

$$Raw = GL + Noise \quad (1)$$

$$\text{Filter} = \text{SG(GL)} + \text{SG(Noise)} \tag{2}$$

$$\text{Diff} = |\text{Filter-Raw}| = |\text{SG(GL)} - \text{GL} + \text{SG(Noise)} - \text{Noise}| \tag{3}$$

where SG is the Savitsky-Golay filter. For all real situations, we can assume that

$$|\text{SG (GL)} - \text{GL}| << |\text{SG (Noise)} - \text{Noise}| \tag{4}$$

Figure 1 shows the logarithm of the mean value of the absolute difference for the Gaussian and Lorentz peaks.

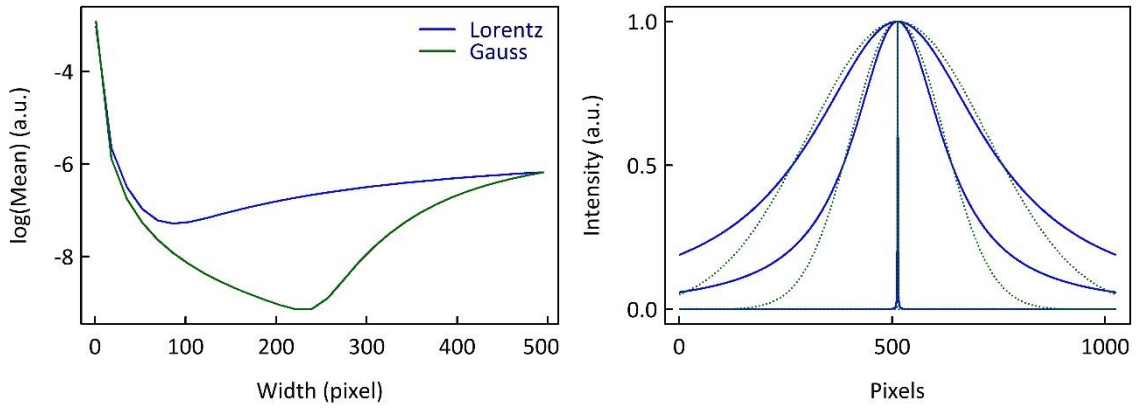

Fig. 1. Logarithm of the mean of the absolute difference as a function of the width of the peak on the left. Examples of peaks for widths = 1, 18 and 36 pixels on the right.

Noise variance after applying Savitsky-Golay filter and subtracting raw data

$$\text{Variance}_{\text{Filter}} = \frac{\sqrt{18}}{\sqrt{35}} \times \text{Variance}_{\text{Raw}} \tag{5}$$

The variance value with using of all data points is calculated as

$$\text{Variance} = \frac{\sqrt{\pi}}{2} \times \text{Mean(Diff)} \tag{6}$$

When only part of data points is used the value of variance will be

$$\text{Variance} = \frac{\sqrt{\pi}}{2} \times \text{Mean}(\text{Diff}_{\text{Thr}}) \times \frac{\text{Thr}}{1-\text{z}} \tag{7}$$

$$\text{z} = \text{erf}^{-1}(2 \times (1 - \text{Thr})) \tag{8}$$

Where 0 <Thr≤1 is the threshold that determines how many points are used, erf^{-1} is the inverse error function and $\mathrm{Diff}_{\mathrm{Thr}}$ is initial part of the sequence.

Results.

The proposed method is applied to real and artificial spectra. The variance of the noise is 1.0 and the threshold is 0.5 for all examples. Figure 2 shows the result of applying the method to pure artificial noise spectra. The result is compared with the direct calculation of the variance.

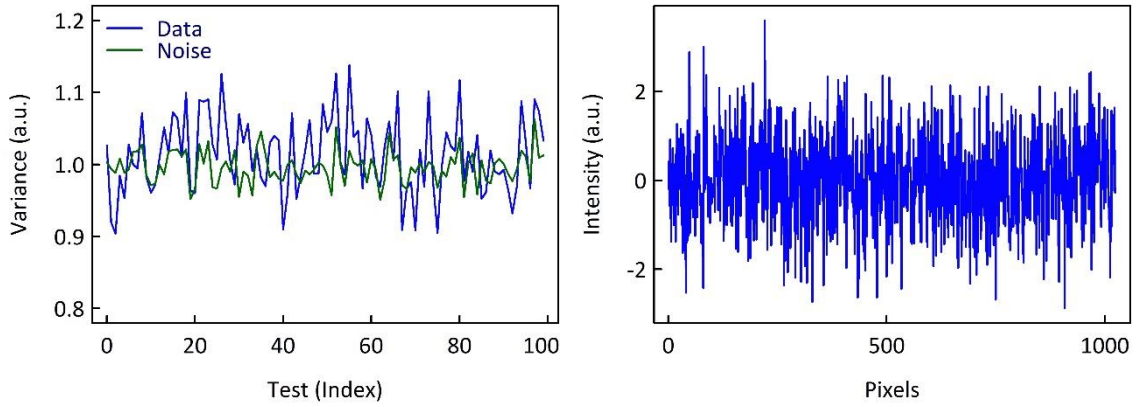

Fig. 2. Estimating the variance of the noise in the data versus direct variance calculation on the left. Example of noise data on the right.

Figures 3 and 4 show the results of modeling a spectrum with different numbers of Gaussian and Lorentzian peaks with widths from 1 to 10 pixels. The baseline is modeled as the sum of the wide Gaussian and Lorentz peaks.

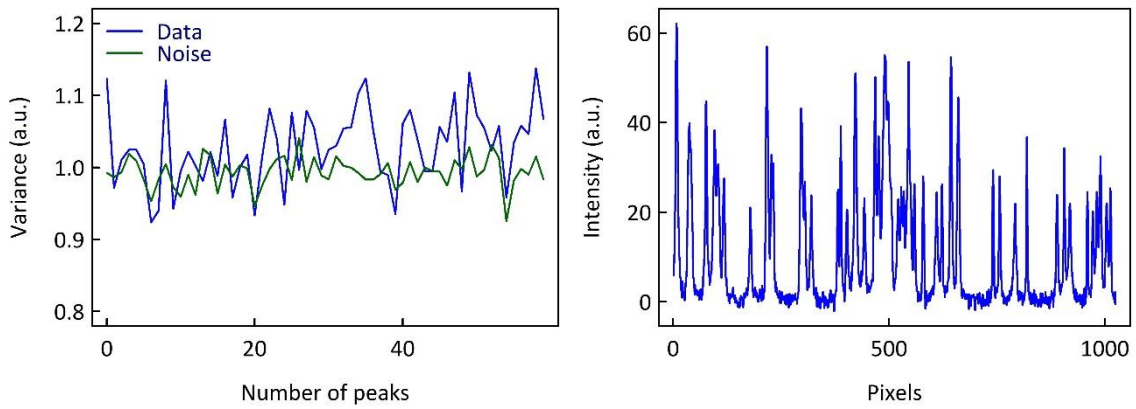

Fig 3. Estimation of variance for different number of peaks on the left. An example of a spectrum with 60 peaks on the right.

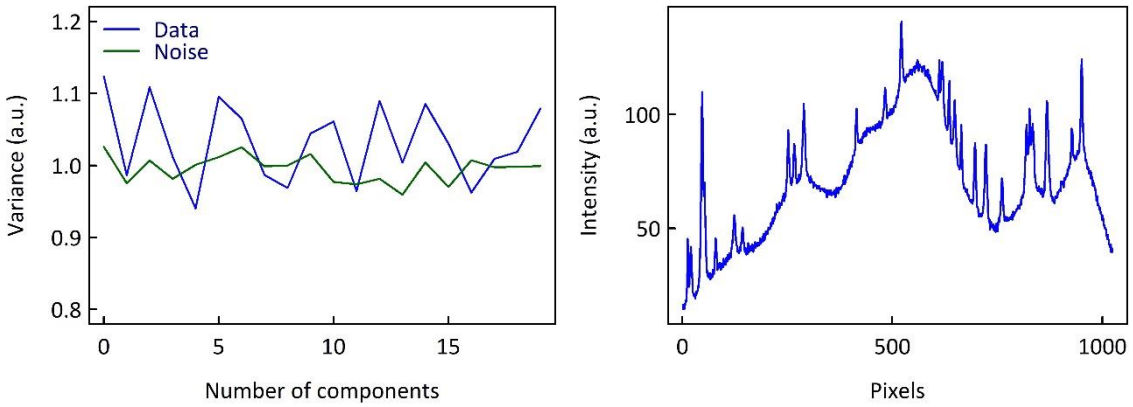

Fig. 3. Estimation of variance for different numbers of baseline components on the left. An example of a spectrum with 20 components on the right. The number of peaks is 30.

Figure 5 shows the result for an artificial spectrum with 30 peaks and 10 baseline components contaminated with spike noise with a pixel spike probability from 0% to 5%. The thresholds used are 0.5, 0.7, and 1.0. The result demonstrates the importance of using the threshold to get a good estimation.

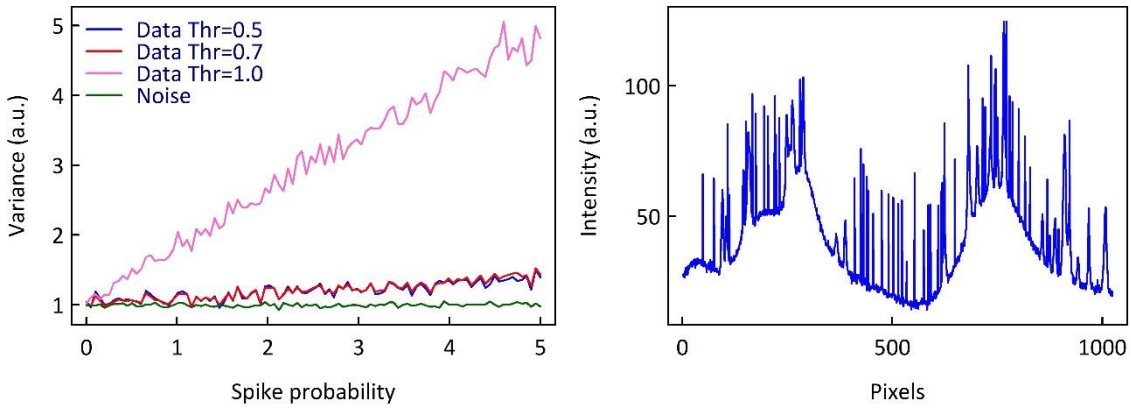

Fig. 5. Estimation of the noise variance for data contaminated by spike noise for different threshold levels on the left. Example of spectra with 5% noise on the right.

Figure 6 shows the estimated variance for different noise levels. The number of peaks is 30, the baseline components are 10, and the peak noise is 0.5%.

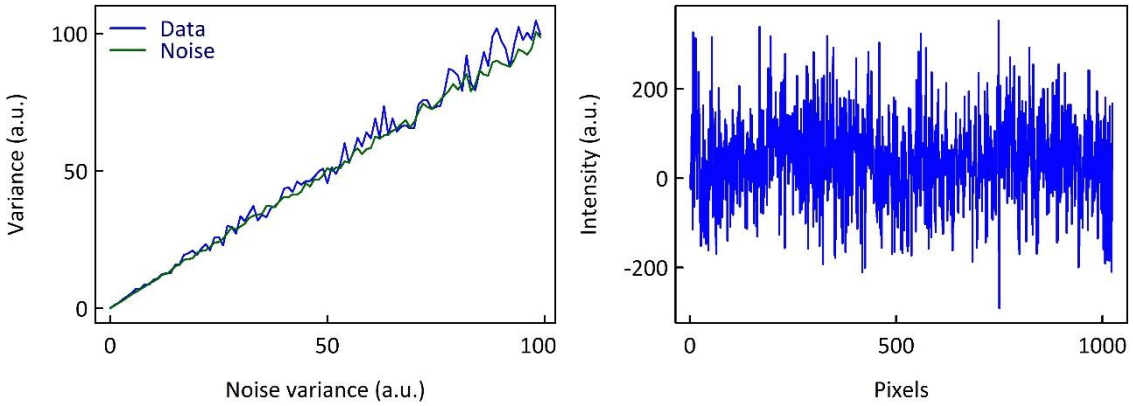

Fig. 6. Calculated variance as a function of noise level on the left. An example of spectra with a noise variance of 100.

The method was tested on one of the standard signal functions used to test data processing algorithms and real quasi-clear Raman spectra of minerals (8). Figures 7,8,9 shows the result of applying the method. The gray graph corresponds to noise variance = 10.

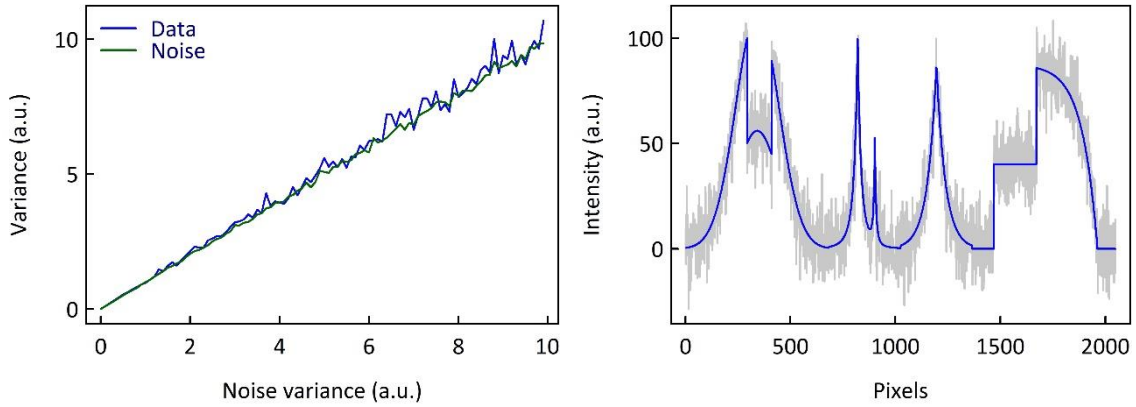

Fig. 7. Calculated variance as a function of noise level on the left. The original function of the artificial signal and the noisy one on the right.

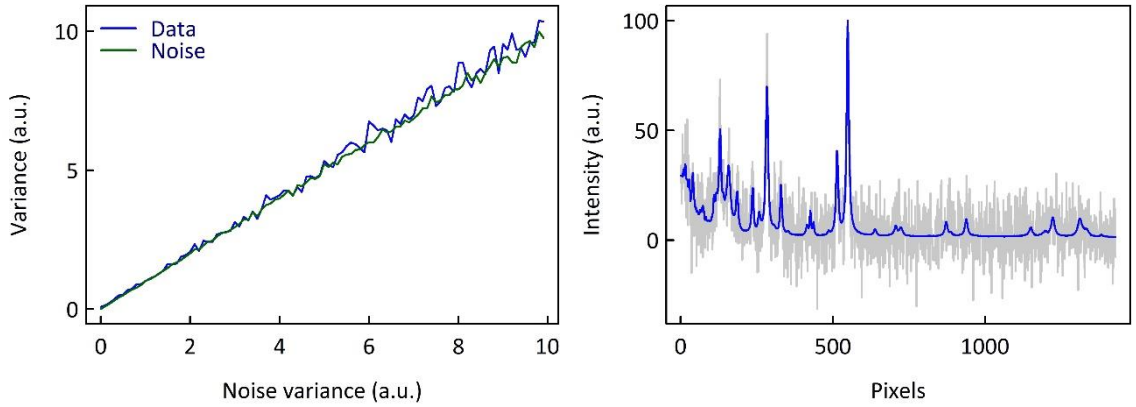

Fig. 8. Calculated variance as a function of noise level on the left. The original Raman spectrum of Albite and the noisy one on the right.

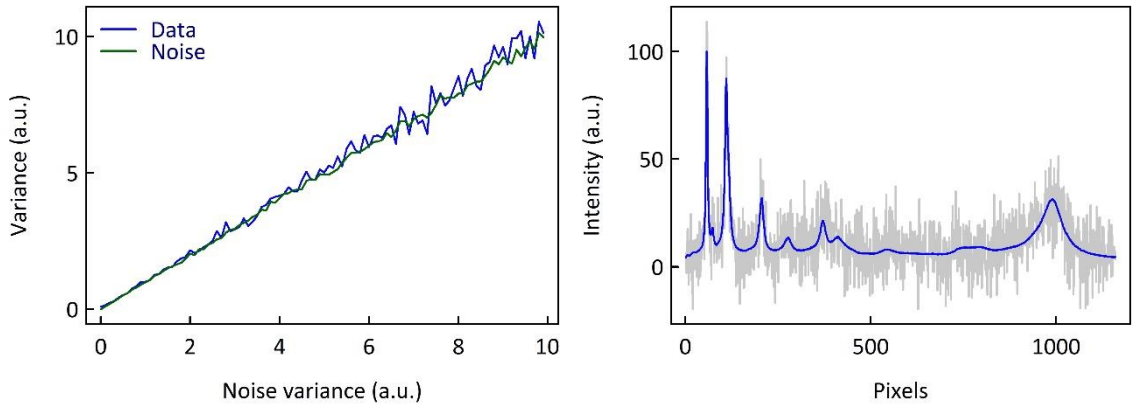

Fig. 9. Calculated variance as a function of noise level on the left. The original Raman spectrum of Hematite and the noisy one on the right.

The results obtained allow to conclude that the difference between the estimation and the real value is less than 15%. The method is robust and does not require user input parameters. It can be used without modification for 2D images. Figure 11 shows the result for astronomical artificial images. The images contain 500 stars and 100 galaxies.

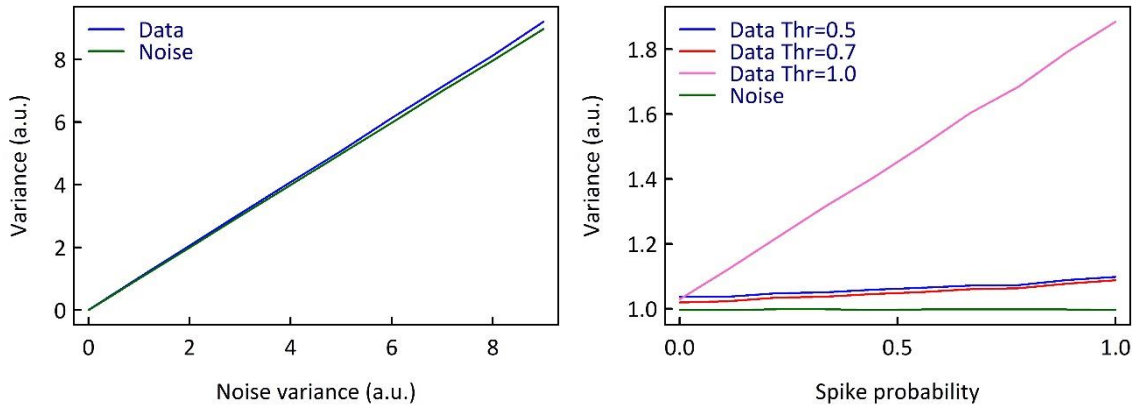

Fig 10. Calculated variance as a function of noise level on the left. Estimation of the noise variance for data contaminated by spike noise for different threshold levels on the right.

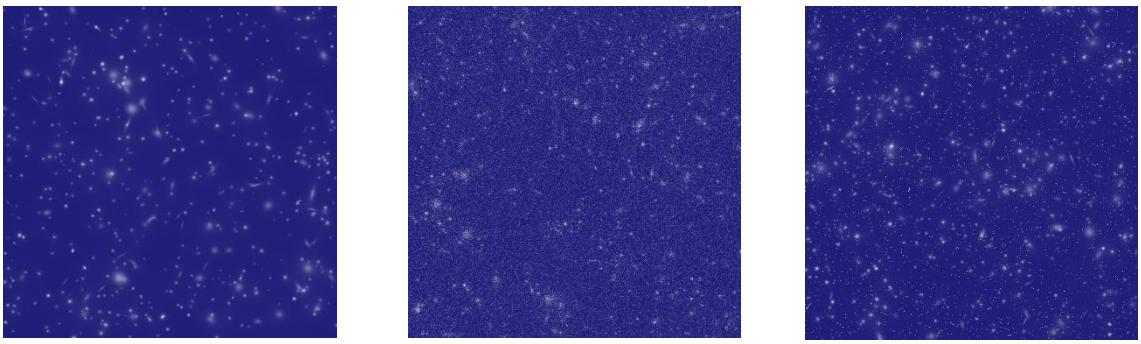

Fig 11. Artificial image of the sky. Image without noise on the left, with a noise variance of 10 in the center and with 1% spike noise on the right.

Conclusions.

The proposed method shows good performance for spectroscopic data. It is simple to implement and does not require user input. The method can also be applied to 2D images and other multidimensional data.

Supplemental Material

Datasets can be downloaded from (9). The online version of the method implementation is available at (10).

References.

[1] D. Makovoz, "Noise Variance Estimation In Signal Processing," 2006 IEEE International Symposium on Signal Processing and Information Technology, 2006, pp. 364-369, doi: 10.1109/ISSPIT.2006.270827.

[2] St. Pyatykh, L. Zheng, and J. Hesser "Fast noise variance estimation by principal component analysis", Proc. SPIE 8655, Image Processing: Algorithms and Systems XI, 86550K. 2013. doi: 10.1117/12.2000276

[3] G. A. Einicke, G. Falco, M. T. Dunn and D. C. Reid, "Iterative Smoother-Based Variance Estimation," in IEEE Signal Processing Letters, vol. 19, no. 5, pp. 275-278, May 2012, doi: 10.1109/LSP.2012.2190278.

[4] S. Sari, H. Roslan and T. Shimamura, "Noise Estimation by Utilizing Mean Deviation of Smooth Region in Noisy Image," 2012 Fourth International Conference on Computational Intelligence, Modelling and Simulation, 2012, pp. 232-236, doi: 10.1109/CIMSim.2012.89.

[5] C. Liu, R. Szeliski, S. Bing Kang, C. L. Zitnick and W. T. Freeman, "Automatic Estimation and Removal of Noise from a Single Image," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 2, pp. 299-314, Feb. 2008, doi: 10.1109/TPAMI.2007.1176.

[6] D. Guo, Y. Wu, S. S. Shitz and S. Verdú, "Estimation in Gaussian Noise: Properties of the Minimum Mean-Square Error," in IEEE Transactions on Information Theory, vol. 57, no. 4, pp. 2371-2385, April 2011, doi: 10.1109/TIT.2011.2111010.

[7]. A. Savitzky, M.J.E. Golay. "Smoothing and differentiation of data with simplified least squares procedures". Anal. Chem. 1964. 36: 1627–1639, doi: 10.1021/ac60214a047

[8]. Handbook of Raman Spectra for geology. http://www.geologie-lyon.fr/Raman/

[9]. Data sets. https://doi.org/10.6084/m9.figshare.14991567

[10]. The online noise calculator.
https://spectralmultiplatform.blogspot.com/p/mathsmooth.html