Chemometrics and Intelligent Laboratory Systems 161 (2017) 8–14

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

A new automatic threshold selecting criteria for spectroscopy data processing

Yuanjie Liu[a,*], Yude Yu[a], Xiaoguang Zhou[a], Chong Wang[b]

[a] State Key Laboratory of Integrated Optoelectronics, Institute of Semiconductors, Chinese Academy of Sciences, P.O. Box 912, Beijing 100083, PR China
[b] School of Foreign Languages, Renmin University of China, Beijing 100872, PR China

ARTICLE INFO

Keywords:
Automatic thresholding
Spectroscopy processing
Noise elimination

ABSTRACT

A nonparametric and unsupervised method of automatic threshold selection to eliminate noise for spectroscopy data processing is described in this paper. A detecting scheme, named bi-trapezoid criteria, is devised where the threshold is selected according to the turning corner present on the uprising intensity trace. This selecting procedure is very simple, utilizing only basic summation on the sorted intensity sequence to optimize threshold for distinguishing between noises and signals. This approach is effective in selecting appropriate value to filter noise when the distribution of noises is not preconditioned or the gap between signals and noises is not obvious. Testing on both artificial and authentic data under specific quantitative evaluation condition shows that this new method performs better than previous ones.

1. Introduction

Signals and noises are critical elements in treating spectroscopy data. Thresholding is almost ubiquitous when carrying out the process to distinguish signals from noises. Therefore, the selection of threshold is the key to success. When using threshold to eliminate noises and to keep the signal peaks in data processing of various analytical data, including Raman spectra, X-ray diffraction, fluorescence, etc., the determination of the threshold value is usually manual (the most straightforward and often the most effective way). However, with an increasing number of data generated by an analytical instrument these days automatic processing is often necessary. A robust, efficient threshold selection method plays an extremely important role in this regard. Currently, there are several popular choices of automatic threshold determination criteria. Classical schemas include Otsu's method [1] derived from histogram strategy, using principle of Stein's Unbiased Risk Estimate [2,3] method, minimax thresholding[4,5] method, segmented regression [11] and the triangle method [12]. In this work, we describe a new method, called bi-trapezoid criteria, as an alternative. Compared with the above-mentioned methods, we show it to be more robust and effective.

At the core, the bi-trapezoid criteria we developed is an improved version of histogram method. Previously, a variety of techniques has been proposed for determining thresholds in 2-D image processing in order to extract objects from their background by slicing. For one dimensional data, the same idea could be applied as well: the signal is equivalent of object or foreground emerged from background noises. If the intensity ranges occupied by the objects and the background are sufficiently large (i.e., the high signal on a low background), there will be a valley in the intensity histogram between two peaks corresponding to these ranges, and one can set a threshold located at the bottom of this valley [6]. Other methods based on the combination of these ideas have been proposed (e.g. [7],) including method attempting to define thresholds that vary from one part to another [8].

In an ideal situation, the histogram may have a deep valley between two peaks representing the objects and the background (or the signal and the noise for one dimensional spectroscopy data). But in reality, limitation of "valley" methods arises when objects occupy only a small portion with resulting peaks of very different sizes, making it difficult to uniquely locate the valley. One most renowned method to overcome this problem was developed by Otsu, in which discriminant analysis is performed to evaluate the "goodness" of and automatically select an optimal threshold that best separate classes of intensities. This method outperforms others in most cases and becomes the algorithm of choice in this category. However, it still fails in some cases when the gap between the signal and noise is not obvious.

Other non-histogram based strategies, including the one developed by Donoho, used statistical principles for more complex situations with the help of wavelet, often dubbed SURE principle , including RigrSURE and HeurSURE thresholding. Derived from the minimization of Stein's unbiased risk estimation, these algorithms avoid any prior hypotheses on the noise-free signal. The deficiency, however, is that they are only

* Corresponding author.
E-mail address: yjliu@semi.ac.cn (Y. Liu).

http://dx.doi.org/10.1016/j.chemolab.2016.12.006
Received 14 July 2016; Received in revised form 19 November 2016; Accepted 9 December 2016
Available online 10 December 2016

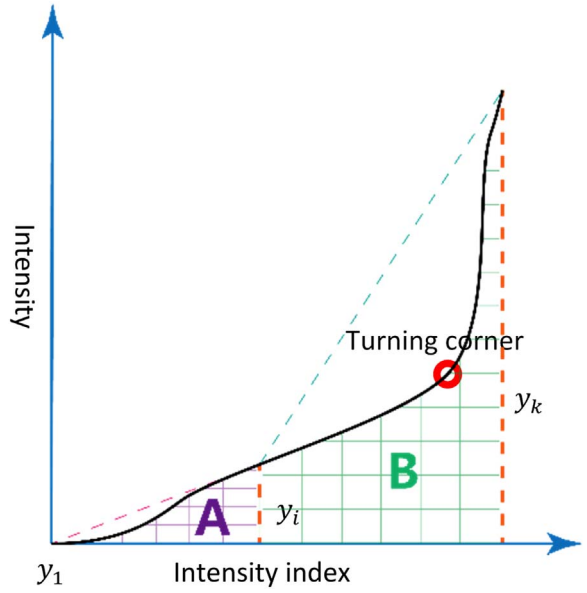

Fig. 1. The geometric view of the formulation (1), the optimization object is aiming to find a position bisecting the area underneath the data trace into two parts that look like trapezoids. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

effective when the noise follows standard normal distribution. Another approach represented by minimax thresholding [4,5] tries to minimize the maximum estimation error for all signals and performs admirably in estimating signals mixed with white Gaussian noise with thresholding estimators in orthogonal bases but fails if the noises do not follow Gaussian distribution. Compared with these precondition-required methods, the new algorithm proposed in this article shares more similarity with the segmented regression [11] and the triangle method [12]. The difference between them, as well as performances of each one, is to be analyzed in the following sections.

When exploring signal recognition and noise elimination, another group of methods, named penalized least squares [13], are also worth of noting. P-spline approaches [14] and PLS based smoother [15] proposed by Paul H. C. Eilers, which can treat both high frequency noises and low frequency fluctuations (known as baseline drift), are representative of this kind of algorithms. Since the discussion here is

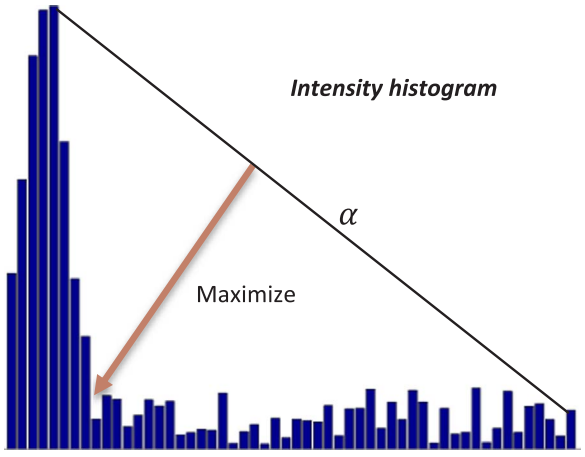

Fig. 3. Triangle method schematic.

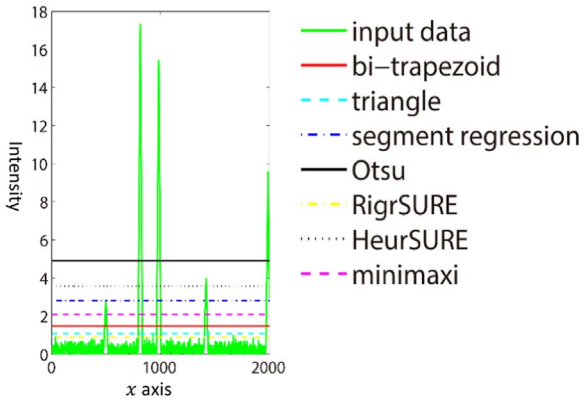

Fig. 4. Left: Testing on artificial data. The simulated data is synthesized with random peaks and standard normal distributed noises. Right: The legend of figures, affecting on Fig. 4, Fig. 5 and Fig. 6.

mainly on threshold selection, and data with baseline drift usually has already been corrected before subjecting to this process, comparison with the PLS methods will not be performed in this paper.

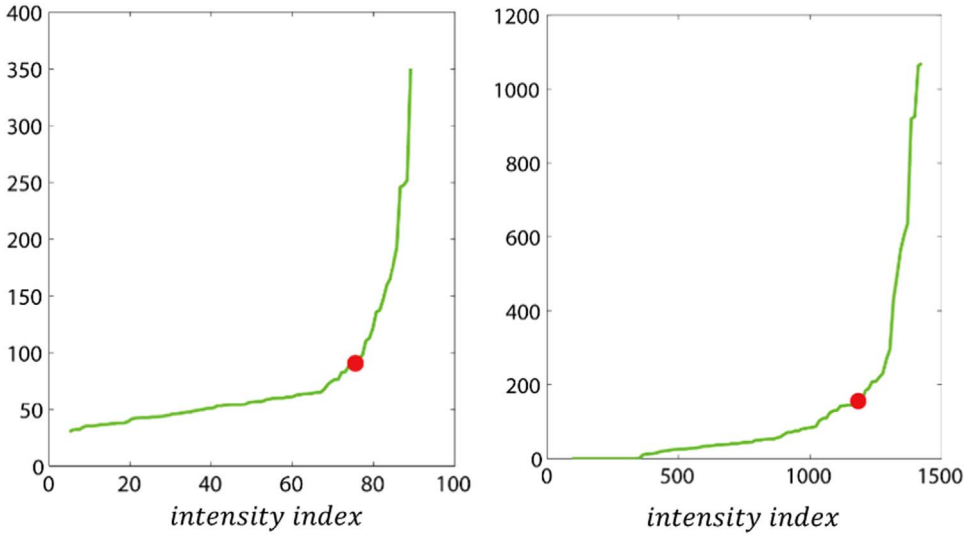

Fig. 2. Locating corner of sorted intensity trace from practical experiments' data. The left is from XRD of Ferberite and the right is from the Raman spectroscopy of Variscite. The red points are corner points detected by Algorithm 1's locating step (step 3). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).