Baseline Correction by Improved Iterative Polynomial Fitting with Automatic Threshold

Article *in* Chemometrics and Intelligent Laboratory Systems · May 2006

DOI: 10.1016/j.chemolab.2005.08.009

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Chemometrics and Intelligent Laboratory Systems 82 (2006) 59 – 65

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

Baseline correction by improved iterative polynomial fitting with automatic threshold

Feng Gan [*], Guihua Ruan, Jinyuan Mo

School of Chemistry and Chemical Engineering, Sun Yat-Sen University, Guangzhou 510275, P.R. China

Received 2 August 2004; received in revised form 6 July 2005; accepted 16 August 2005
Available online 11 November 2005

Abstract

An iterative polynomial fitting method is proposed for the estimate of the baseline of chemical signal. The new method generates automatic thresholds by comparing the chemical signal with the calculated signal from polynomial fitting in the iterative processes. The signal peaks are cut out consecutively in the iterative processes so the polynomial fitting will finally give a good estimation of the baseline. Simulated data and real data from capillary electrophoresis experiment are used to demonstrate the feasibility of the proposed method.

Keywords: Data processing; Baseline correction; Iterative polynomial fitting; Automatic threshold

1. Introduction

Noise and baseline are two very common problems in analytical chemistry. Both of them will lead to the deterioration of accuracy and precision. Noise is a high-frequency signal while baseline variation is a low-frequency one. A lot of methods, such as mean filtering, exponential smoothing, Savitzky–Golay, Fourier transfer, have been used in coping with noise problems. Recently, wavelet [1–4] has got more interest in the denoising of signals. However, report on baseline correction is relatively few.

Baseline problem is a relatively complicated one. Conventional frequency analysis cannot give a theoretic description of the baseline information. It is well known that baseline is a low-frequency signal but it is difficult to make a theoretic distinguishing of baseline information from others. So it is difficult to develop a theoretically perfect method to cope with baseline problem. Approximate estimate of baseline has been the general method. Conventionally, a straight line is used to connect the two ends of a signal peak. The straight line is taken as the baseline and further calculation of peak area or peak height is based on it. If the straight line does not fit the real baseline, the calculation will lead to errors. Recently, new approaches [5–9] have been made to make a better estimate of the baseline. Golotvin and Williams [5] take two steps to cope with the baseline problem. The first step is the baseline recognition by setting a rule and the second is the baseline modeling. Ruckstuhl et al. [6] put forward a robust baseline estimate which uses robust local regression to estimate baseine. Some works [7–9] estimate baseline by using wavelet analysis. Shao et al. [7] make baseline correction by finding an approximation of the baseline in the decompose process. Ma and Zhang [8] obtain the baseline by discarding the elements attributed to the analyte signal. Wang and Mo [9] take a special way to cope with baseline problem. In their work, Mexican hat wavelet is used to the regression of signal. The part of a peak that is over the estimated one constructs a threshold and a new signal is generated by cutting out the part of the peak. An iterative process is implemented for the further regression based on the new signal until a stable baseline is obtained.

In this paper, we take an improved iterative polynomial fitting to estimate baseline based on the idea of automatic threshold [9]. As the analytical signal is usually complicated profile, polynomial fitting with lower power will not give a good estimate to the whole signal but a rather good approximate estimation of the baseline. We implement iterative

[*] Corresponding author.
E-mail address: cesgf@zsu.edu.cn (F. Gan).

0169-7439/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2005.08.009

processes to adjust the signal peaks step by step and finally give a best estimation of the real baseline. This method is quite simple and easy to implement. Both simulated and real signal are used to test the utility of this method and the results are rather good.

2. Methodology

A chemical signal can be expressed in the following form,

$$y(x_i) = b(x_i) + s(x_i) + \varepsilon_i \tag{1}$$

where, $y(x_i)$ is measured signal, $s(x_i)$ is true signal, $b(x_i)$ is baseline, and ε_i is measurement error.

Eq. (1) tells us that if $s(x_i)$ is eliminated, the baseline is obtained. We use iterative polynomial fitting to cope with this problem. As the chemical signal usually takes a complicated form, it can be predicted that the polynomial fitting with lower power will not give perfect fitting to the whole signal but an approximate estimate of the baseline. The unfitted parts of the signal peaks which are above the estimated baseline generate the automatic thresholds. By cutting out the parts, one can obtain new signal $y(x_i)$, which is then used in further consecutive polynomial fitting. An iterative process is used in both the polynomial fitting of the whole signal and the cutting out of the parts of the peaks. By this way, $s(x_i)$ will be finally eliminated and a good estimate of the baseline is obtained.

Polynomial fitting is a set of classic mathematic methods. One of the simple versions of the polynomial fitting is to describe a function by a set of variables with the form of x, x^2, x^3, ..., x^n. To a function $f(x)$, the equation of polynomial fitting can be written as following,

$$y(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + \cdots + a_nx^n \tag{2}$$

where, a_0, a_1, a_2, a_3,...,a_n are the coefficients to be determined; n is the power of the polynomial function.

To a series of variables $x_1, x_2,...,x_m$, Eq. (2) can be written in following matrix form,

$$\begin{bmatrix} y(x_1) \\ y(x_2) \\ y(x_3) \\ \cdots \\ y(x_m) \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ 1 & x_3 & x_3^2 & \cdots & x_3^n \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_m & x_m^2 & \cdots & x_m^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \cdots \\ a_n \end{bmatrix} \tag{3}$$

Eq. (3) can be written in the following concise matrix form,

$$\mathbf{y} = \mathbf{Xa}. \tag{4}$$

Then, the polynomial fitting result is as following,

$$\hat{\mathbf{b}} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{5}$$

where, $\mathbf{b}$ is an estimate of $\mathbf{y}$. Superscript T and -1 denote the matrix transpose and matrix inversion, respectively.

Based on the discussion above, we put forward an algorithm for the estimation of baseline as followings.

Step 0 For original signal $\mathbf{y}_0$, set the power n.

Step 1 Establish polynomial fitting equation as shown in Eq. (3).

Step 2 Calculate the polynomial fitting result $\mathbf{b}_k$ by Eq. (5) in kth iteration.

Step 3 Compare $\hat{\mathbf{b}}_k$ with $\mathbf{y}_{k-1}$ in the regions of signal peaks, if $y_{k-1}(i) > \hat{\mathbf{b}}_k(i)$, $y_k(i) = \hat{\mathbf{b}}_k(i)$, $i = 1, 2, \ldots$. Here, i is in the region where $y_{k-1}(i) > \hat{\mathbf{b}}_k(i)$.

Step 4 If following criterion is reached, stop; otherwise, take $\boldsymbol{y}_k$ as new signal and go to Step 1.

$$\rho = \frac{\|\mathbf{b}_k - \mathbf{b}_{k-1}\|}{\mathbf{b}_{k-1}} < 0.001 \tag{6}$$

where, $\hat{\mathbf{b}}_k$ and $\hat{\mathbf{b}}_{k-1}$ are polynomial fitting results at kth and $(k-1)$th iterations, respectively. At zero iteration, $\hat{\mathbf{b}}_0 = \boldsymbol{y}_0$. If the criterion (6) holds, $\hat{\mathbf{b}}_k$ will be the estimated baseline.

Step 5 If the estimated baseline matches the signal well, stop; otherwise, change power n and go to Step 0.

3. Experimental section

3.1. Simulated signals

A series of simulated signals are used to evaluate the method we put forward in this paper. The real capillary electrophoresis (CE) signal is the reference of these simulations so we use the same units as CE for the axes. Simulated signal peaks are generated by Gaussian function and random noise is added in the simulated signals.

3.2. Experimental signals

3.2.1. Data set 1

This data is generated from CE analysis of amino acids standard solution with reversed EOF using high-frequency conductivity detector. Separation voltage: −16.0 kV; injection time: 15.0 s, fused silica capillary: 150 μm (i.d.) × 43.0 cm; buffer: diethylamine: H_3BO_3: CTAB: carbonamide: β-CD (15.0: 0.1: 0.1: 50.0: 0.45 mmol/L); amino acids standard solution, 5.0×10^{-5} mol/L; There are nine components in the system, they are L-arginine, L-aspartic, L-glutamic acid, Cystine, L-glycine, L-serine, L-threonine, L-isoleucine and L-proline, according to their times of appearance. The instrument is homemade.

3.2.2. Data set 2

This data set is also from above experiment but the analytical conditions have a little change. Separation voltage is −12.0 kV and others keep the same.

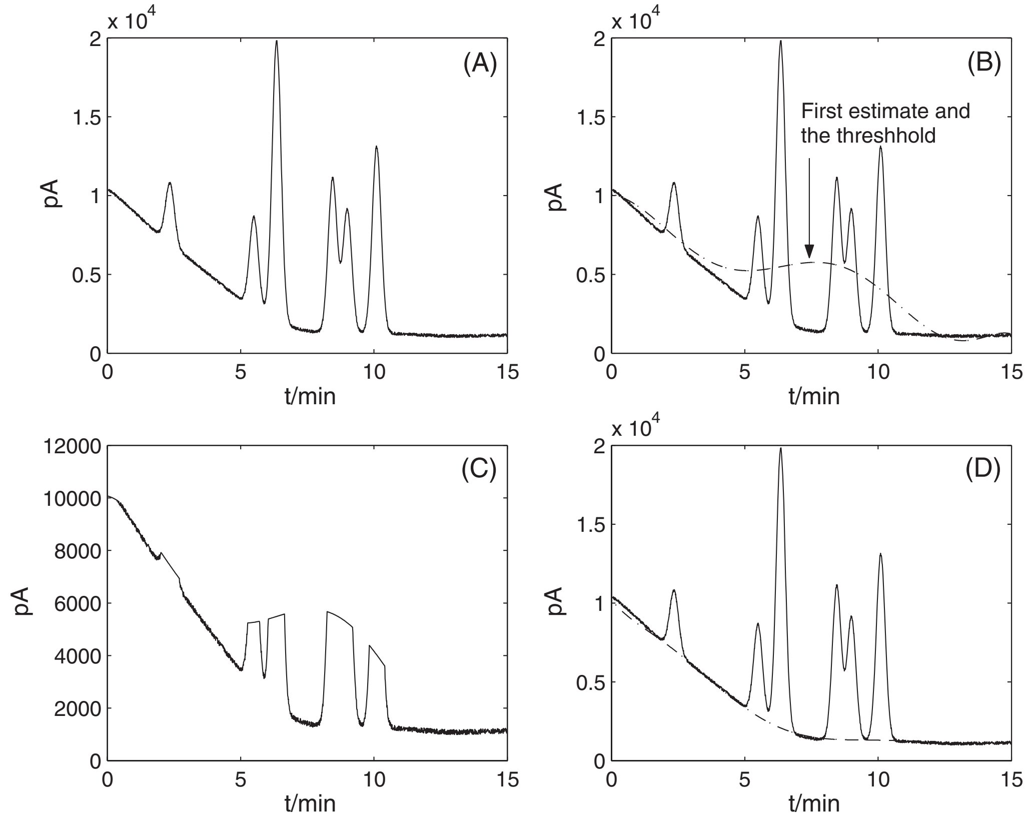

Fig. 1. Simulated signal 1 is shown in (A); (B) shows the first polynomial fitting result(dashed line). The dashed line will also be the automatic threshold. (C) shows the new signal by cutting out the parts of the peaks that is above the threshold. (D) shows the final estimated baseline (dashed line). The power of the polynomial is 7.

3.3. Programming

All programs are written using Matlab 6.5 and run under Windows 2000 on a personal computer (RAM 1G, CPU 2.6 GHz).

4. Results and discussion

Fig. 1 shows the basic process for the estimation of baseline. Fig. 1(A) is the original signal. The signal peaks are separated so there are enough baseline segments

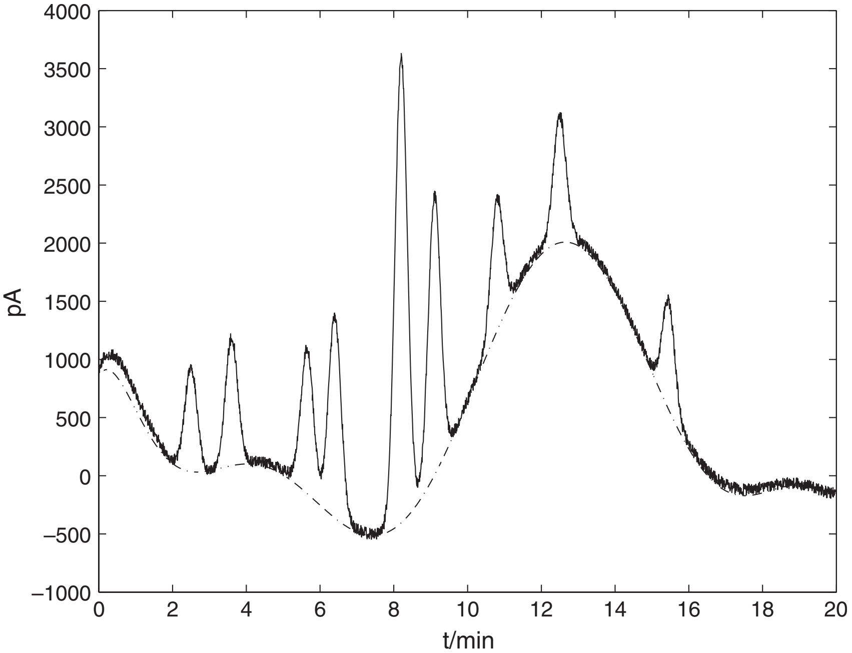

Fig. 2. Simulated signal 2 (solid line) and the estimated baseline (dashed line). The power for polynomial is 11.

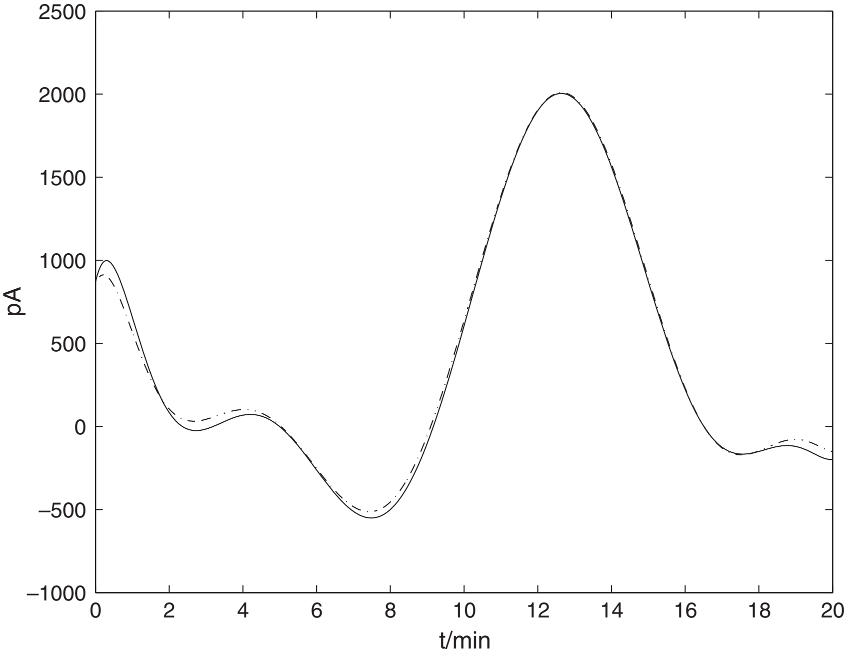

Fig. 3. Real baseline (the solid line) and the estimated baseline (dashed line) for simulated signal 2. The power for polynomial is 11.

between two peaks, which offers more information in the estimation of baseline. The dashed line in Fig. 1(B) is the first estimation of the original signal by our method. One can see that the fitting is not good. However, this bad fitting offers us the chance to adjust the original signal. By cutting out the parts of the peaks that are above the dashed line, an adjusted new signal is obtained which is shown in Fig. 1(C). The new signal replaces the original one and is used in the next iteration process. This adjustment is implemented in every iteration process and finally, a good estimation of the baseline is obtained which is shown in Fig. 1(D).

Fig. 2 shows relatively complicated signal. We increased the overlapping of signal peaks to study its affect on the estimation of baseline. The dashed line in both Figs. 2 and 3 shows that the estimated baseline matches both the signal and the real baseline well. The reason is that there is enough baseline information between the signal peaks although the

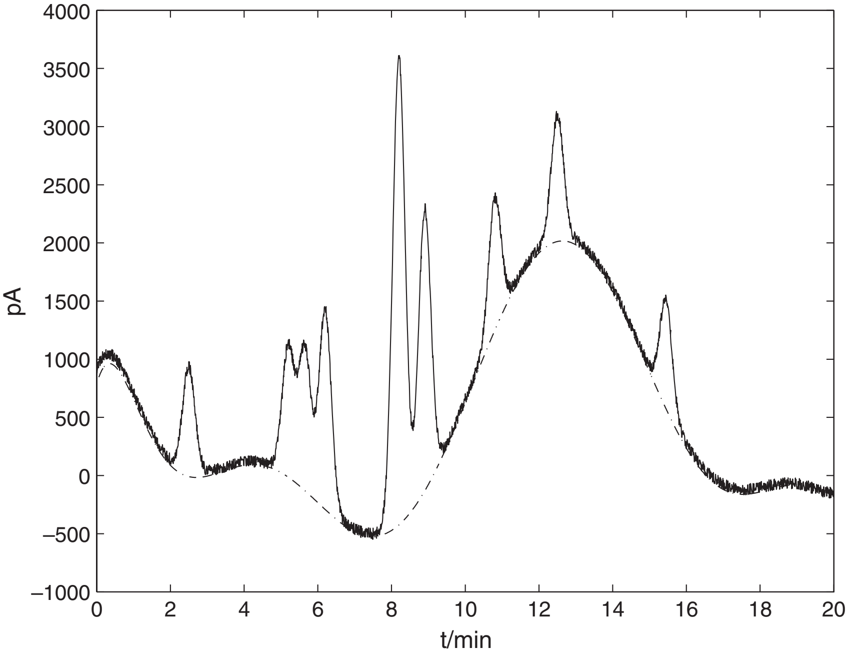

Fig. 4. Simulated signal 3 (solid line) and the estimated baseline (dashed line).

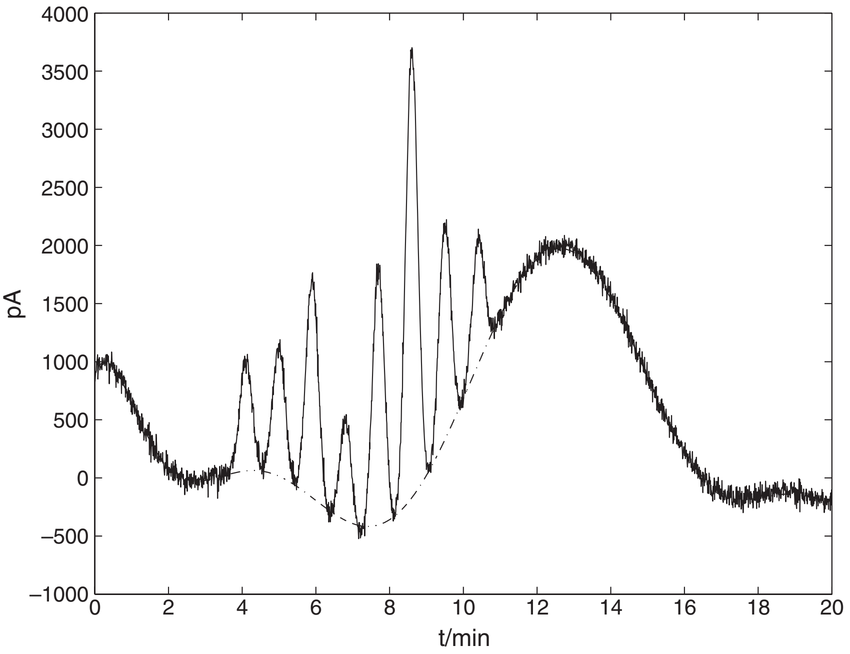

Fig. 5. Simulated signal 4 (solid line) and the estimated baseline (dashed line). The power of the polynomial is 11.

overlapping of the signal peaks increase. Fig. 4 shows an even more overlapping signal peaks but the estimated baseline still matches the signal well. The reason is just the same as above.

We studied another situation where the baseline information is not enough. In a simulated signal shown in Fig. 5, we moved the signal peaks closer so there is less baseline information between some peaks. Under this situation, it is difficult for us to determine whether there is real baseline information between some peaks or there is just the overlapping of signal peaks. Although the estimated baseline seems matching the signal well as shown in Fig. 5, it does not give a perfect fitting to the real baseline shown in Fig. 6.

Fig. 7 shows the real data set 1 and the estimated baseline. One can see from this figure that the baseline should be a large curve and the calculated baseline fits the signal rather good. Fig. 8 shows a relatively complicated situation. One can see that the estimated baseline also fits the data set well except for two local regions. What causes the sharp changes at these two regions is not clear but they will not represent the real trend of

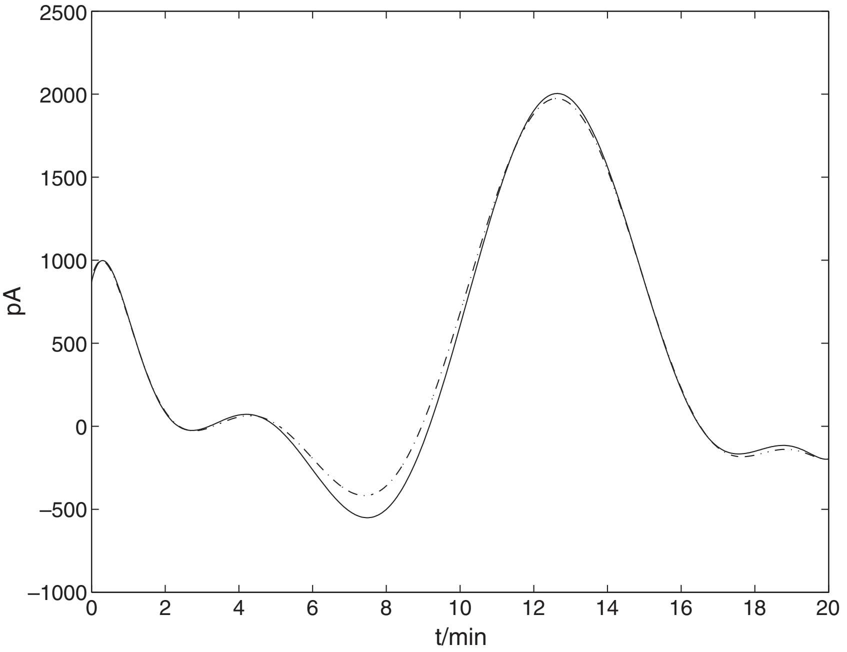

Fig. 6. Real baseline (the solid line) and the estimated baseline (dashed line) for simulated signal 4. The power of the polynomial is 11.

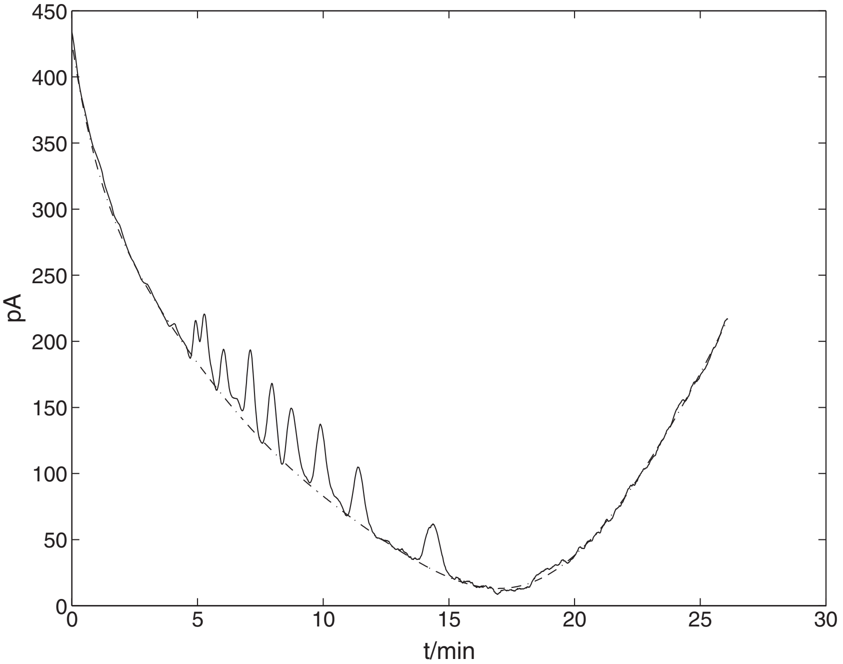

Fig. 7. Data set 1: electrophorogram of amino acids standard solution in CE with reversed EOF using high-frequency conductivity detector. The dashed line is the calculated baseline.

the baseline. Further calculations of peak area and peak height based on the calculated baseline are acceptable.

5. Conclusion

Using polynomial fitting with automatic threshold is an interesting approach to cope with baseline problem. The attractive part of this method is easy to understand and implement. Polynomial fitting is a good way to describe functions with lower power, which is just the characteristic of the baseline. On the other hand, calculating the baseline by a consecutive way offers the chance of approaching the baseline step by step. At each step, the information from signal peaks is reduced and the baseline information takes the control position, which finally reaches to the best estimation of the baseline.

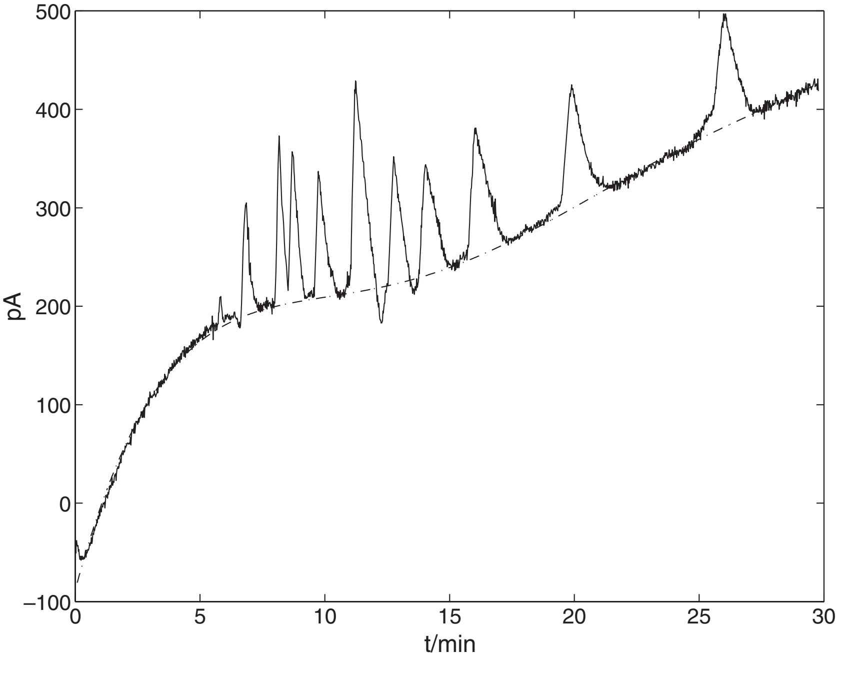

Fig. 8. Data set 2: electrophorogram of amino acids standard solution in CE with reversed EOF using high-frequency conductivity detector.The dashed line is the calculated baseline.

Acknowledgement

This work was supported by grants from the National Natural Science Foundation of China (No. 20475067). Thanks to Ph.D. student, Haiyun Zhai, for her offering the real data sets.

References

[1] C.R. Mittermayr, S.G. Nokolov, H. Hutter, M. Grasserbauer, Chemom. Intell. Lab. Syst. 34 (1996) 187–202.

[2] B. Walczak, D.L. Massart, Chemom. Intell. Lab. Syst. 36 (1997) 81–94.

[3] V.J. Barclay, R.F. Bonner, I.P. Hamilton, Anal. Chem. 60 (1997) 78–90.

[4] C. Perrin, B. Walczak, D.L. Massart, Anal. Chem. 73 (2001) 4903–4917.

[5] S. Golotvin, A. Williams, J. Magn. Reson. 146 (2000) 122–125.

[6] A.F. Ruckstuhl, M.P. Jacobson, R.W. Field, J.A. Dodd, J. Quant. Spectrosc. Radiat. Transfer 68 (2001) 179–193.

[7] X.G. Shao, W.S. Cai, Z.X. Pan, Chemom. Intell. Lab. Syst. 45 (1999) 249–256.

[8] X.G. Ma, Z.X. Zhang, Anal. Chim. Acta 485 (2003) 233–239.

[9] Y. Wang, J.Y. Mo, in: Chemical J. on Internet, vol. 5, 2003, pp. 16–19.