Contents lists available at ScienceDirect

International Journal of Mass Spectrometry

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

Interpolational and smoothing cubic spline for mass spectrometry data analysis

A.V. Chudinov[a,b], Wei Gao[a,*], Zhengxu Huang[a], Weiguang Cai[c], Zhen Zhou[a], V.V. Raznikov[b], V.I. Kozlovski[b], I.V. Sulimenkov[b]

[a] Institute of Atmosphere Environment Security and Pollution Control, Jinan University, Guangzhou 510632, China
[b] Institute of Energy Problems of Chemical Physics, Russian Academy of Sciences, Chernogolovka, Russia
[c] Guangzhou Hexin Analytical Instrument CO LTD, Guangzhou 510530, China

ARTICLE INFO

Article history:
Received 6 September 2015
Received in revised form 9 November 2015
Accepted 16 November 2015
Available online 11 December 2015

Keywords:
Peak approximation
Peak localization
Peak detection
Spline interpolation
Spline approximation
Spline for histogram
Mass accuracy
Mass calibration

ABSTRACT

A special cubic spline for histograms was examined as a possible tool for the processing of mass spectrometry data. Algorithms of interpolating and smoothing cubic splines are described shortly. It was shown that peak apex localizations using splines yield similar results as peak apex determinations using parabolas and centroids for high intensity peaks, while for low intensity peaks, the peak apex localizations using splines are superior. It also was shown that peak localization accuracies for real experimental time-of-flight mass spectral data are higher in case of peak localizations that use apex determinations as compared to peaks determined by centroid or Gaussian approximation of peaks. Some examples of peak detection algorithms using splines are given. A new method of peak intensity calculations is proposed taking into account the baseline. All approaches are directly applicable for the processing of TOF MS data, though specific changes may be desirable for other MS data.

1. Introduction

One of the most important parameters of mass spectrometers is the achievable mass accuracy. It was shown that centroid peak location [1,2], mass resolving power *R*, and mass accuracy $\sigma_{m/z}$ (measured in ppm) are connected to each other by the relation:

$$\sigma_{m/z} = \frac{10^6}{R\sqrt{N}}, \tag{1}$$

where *N* is the number of ions in the peak. In case 1/*R* is substituted by a mean square width of a peak, Eq. (1) is just a standard error of the mean [3]. However, two things should be taken into account about Eq. (1). First, the *m*/*z* scale is absolute, in other words, the error of the *m*/*z* scale calibration can be disregarded. Second, the peak localization error corresponds only to a "good" shape of the peak, such as a Gaussian with fast disappearing tails. Hence, Eq. (1) may be considered as an upper/lower limit for a mass accuracy/error.

The regular mass calibration law in time-of-flight mass spectrometry is [4,5]:

$$t_i = t_0 + k_1(m/z)_i^{1/2} + k_2(m/z)_i + \cdots + k_p(m/z)_i^{p/2}. \tag{2}$$

where all calibration coefficients $t_0, k_1, k_2, \ldots, k_p$ can be found using linear regression [3]. Actually, only the first two terms in Eq. (2) have a physical meaning and are usually sufficient for a good mass calibration. Some investigators reported an increase of mass accuracy when higher order terms were used [4,5] in Eq. (2). However, supposedly, higher terms just describe biases correlated with intensity ratios between peaks used for calibrations, though in special cases like MALDI-TOF MS with delayed extraction, higher terms in Eq. (2) may be obligatory for a correct mass calibration [6].

In the case of mass spectrometry, it is possible to use a more convenient representation of Eq. (2) by expanding *m*/*z* into a series of *t* [7]. Hence, the simplest and adequate equation for mass calibration in TOF MS mass spectra is

$$m/z = \left(\frac{t - t_0}{k_1}\right)^2. \tag{3}$$

As mentioned above [8], peak localizations using centroids are not ideal in case of peak tails. Another variant of a peak localization is to approximate a peak by a Gaussian function. However, a

[*] Corresponding author at: Institute of Atmosphere Environment Security and Pollution Control, Jinan University, Guangzhou, 510632, China.
Tel.: +86-020-82071910; fax: +86-020-82071902.
E-mail address: w.gao@hxmass.com (W. Gao).

http://dx.doi.org/10.1016/j.ijms.2015.11.008
1387-3806/© 2015 Elsevier B.V. All rights reserved.

Gaussian approximation may yield even more erroneous results because of peak asymmetries and tails [9]. It can be productive to use only points in a peak above a baseline or use only a small number of neighboring points around a peak top for the centroid estimation. In addition, polynomials of higher than second order can be used for the description of a quasi-Gaussian peak [9]. However, unfortunately, all these methods lead to an overcomplication of the data processing algorithms. Also, it is interesting to note, that centroids in time and m/z scales do not coincide exactly with each other. For example, if Eq. (3) was used, then

$$\overline{m/z} = \overline{\left(\frac{t-t_0}{k_1}\right)^2} = \left(\frac{\bar{t}-t_0}{k_1}\right)^2 + \frac{(\overline{t^2}-\bar{t}^2)}{k_1^2} = \frac{((\bar{t}-t_0)^2+\sigma_t^2)}{k_1^2}, \tag{4}$$

where over-lines denote centroid values (or averages).

In the current work, a more sophisticated way of peak apex localizations is chosen by using cubic spline functions [10,13]. Previously, classical spline approximations were used in the processing of mass spectrometric data to calibrate the mass scale of static magnetic mass spectrometers [2,11] as well as for baseline approximations [12]. The method of so-called quasispline approximation for peak apex localizations and mass resolution enhancements was described also [8,9,12]. As mentioned already, the spline interpolation produces more exact results, while using high-order interpolating polynomials is less effective due to the so-called Runge phenomenon [10,13]. By using spline functions, it is possible to not only easily locate peak maxima but also to estimate other peak parameters like peak widths and peak areas, which could be useful for some applications. In addition, since cubic spline coefficients contain approximations of peak shapes and their derivatives [10,13], cubic splines can be used as part of peak detection algorithms. Moreover, as will be shown below, a smoothing of raw mass spectral data is also possible [10,13].

The main objective of the work here is to examine a special cubic histogram spline as a possible tool for mass spectrometry data processing. Histogram splines of arbitrary orders were already developed earlier [14], which we will call here as histosplines, for short. Using histosplines, more exact estimations of peak parameters can be obtained, because time-of-flight peaks are also histograms of ion flow probability distributions. This assumption could not only be valid for ion counting detection systems, where time to digital converters (TDC) are used, but also for ion detection systems that use transient recorders with analog to digital converters (ADC), because any ADC has a finite window in a time scale.

2. Theory

A spline function is a continuously differentiable piecewise polynomial of desired order that interpolates raw data at an arbitrary time position between t_{i-1} and t_i time bins. In the work here, all splines are restricted to third-order polynomials:

$$S_i(t) = a_i + b_i(t-t_{i-1}) + c_i(t-t_{i-1})^2 + d_i(t-t_{i-1})^3. \tag{5}$$

For regular spline interpolations, the coefficients a_i, b_i, c_i, and d_i can be obtained from the conditions:

$$\begin{aligned} &S_i(t_i) = S_{i+1}(t_i) = I_i \\ &S_i{}'(t_i) = S_{i+1}{}'(t_i) = b_{i+1} \quad , \\ &S_i{}''(t_i) = S_{i+1}{}''(t_i) = 2c_{i+1} \end{aligned} \tag{6}$$

where I_i is an ion current intensity measured at an ith spectrum point. For $M+1$ of raw data points, all coefficients in Eq. (5) can be expressed via c_i [10,13]. Assigning further for equally spaced data:

$$t_i - t_{i-1} = h = 1 \tag{7}$$

where the c_i can be obtained [10,13] by:

$$c_i + 4c_{i+1} + c_{i+2} = 3I_i - 6I_{i+1} + 3I_{i+2} \tag{8}$$

In order to solve Eq. (8), additional assumptions are necessary. In a regular cubic spline [10,13], it can be assumed that

$$c_1 = c_{M+1} = 0 \tag{9}$$

In ref. [15], the condition of a minimum norm of third derivative's breaks was proposed to estimate boundary condition for splines. However, in this case, the obtained matrix for the system of equations is not a Toeplitz band [16]. Hence, relatively slow algorithms for solutions should be used. Admittedly, this could be dramatically slow for big data sets.

In case of a histospline, the first of Eq. (6) can be rewritten as:

$$\int_{t_{i-1}}^{t_i} S_i(t)\,\mathrm{d}t = I_i \tag{10}$$

$$S_i(t_i) = S_{i+1}(t_i) = a_{i+1} \tag{11}$$

So that Eq. (8) can be transformed to:

$$c_i + 11c_{i+1} + 11c_{i+2} + c_{i+3} = 12I_i - 24I_{i+1} + 12I_{i+2}. \tag{12}$$

Thus, one extra boundary parameter can be set as compared to a regular spline. It is possible, for example, to set:

$$c_1 = c_M = c_{M+1} = 0. \tag{13}$$

Necessity of this additional parameter follows from the fact that one I_i value corresponds to two time bins, t_{i-1} and t_i. Please, also, note that third-order histospline corresponds to a fourth-order spline accordingly to,

$$\begin{aligned} \int_0^t S_i(t)\,\mathrm{d}t = &\sum_{k=0}^{k=i-1} I_k + a_i(t-t_{i-1}) + \frac{b_i}{2}(t-t_{i-1})^2 + \frac{c_i}{3}(t-t_{i-1})^3 \\ &+ \frac{d_i}{4}(t-t_{i-1})^4, \end{aligned} \tag{14}$$

where t is between t_{i-1} and t_i, and $I_0 = 0$.

In order to remove experimental noise from the raw data, smoothing can be applied. The data smoothing is using splines eq.

$$(1-p)\sum_{i=1}^{M} \frac{(I_i - S(t_i))^2}{\sigma_i{}^2} + p\int_0^{t_M} (S^{(m)}(t_i))^2\,\mathrm{d}t \to \min \tag{15}$$

needs to be minimized [13], where, p is a smooth factor, σ_i are the weightings and m is an order of a spline derivative. Eq. (15) is a classical Lagrange eq. for a conditional extremum problem [17]. Alternatively, it is possible to use a smoothing algorithm [8,9,18] first and to build then an interpolating spline (without smoothing).

Please note that cubic splines can be expressed as linear combinations of B-splines [10,13,19]

$$S(t) = \sum_i \alpha_i B_{i,n}(t). \tag{16}$$

At case of cubic splines, order of B-splines is $n = 3$ in Eq. (16). However, for cubic histosplines, fourth-order B-splines need to be used. Splines of any order can be easy obtained using B-splines. However, in this work, regular spline Eq. (5) was used, as it is more computationally simple especially for cubic splines [10,13,19].