Available online at www.sciencedirect.com
SCIENCE DIRECT®
Applied Mathematics Letters 18 (2005) 1412–1417
Applied Mathematics Letters
www.elsevier.com/locate/aml

A numerical method for mass spectral data analysis☆

Anthony J. Kearsley[a,*], William E. Wallace[b], Javier Bernal[a], Charles M. Guttman[b]

[a]*Mathematical and Computational Sciences Division, National Institute of Standards and Technology, Gaithersburg, MD 20899-8910, United States*
[b]*Polymers Division, National Institute of Standards and Technology, Gaithersburg, MD 20899-8541, United States*

Received 18 January 2005; accepted 24 February 2005

Abstract

The new generation of mass spectrometers produces an astonishing amount of high-quality data in a brief period of time, leading to inevitable data analysis bottlenecks. Automated data analysis algorithms are required for rapid and repeatable processing of mass spectra containing hundreds of peaks, the part of the spectra containing information. New data processing algorithms must work with minimal user input, both to save operator time and to eliminate inevitable operator bias. Toward this end an accurate mathematical algorithm is presented that automatically locates and calculates the area beneath peaks. The promising numerical performance of this algorithm applied to raw data is presented.
Published by Elsevier Ltd

1. Introduction

Modern mass spectrometers are capable of producing large, high-quality data sets in brief periods of time [13]. It is not uncommon for a synthetic polymer to produce a spectrum with hundreds of peaks. This motivates the design of automated data analysis algorithms capable of rapid and repeatable processing of raw mass spectrometer data. While many algorithms for the analysis of raw mass spectrometer output already exist, they all require significant operator input. In some cases smoothing parameters must be selected, in other cases one must identify peaks from noise or vice versa, and many algorithms assume

☆ Contribution of the National Institute of Standards and Technology and not subject to copyright in the United States.
* Corresponding author.
E-mail address: ajk@cam.nist.gov (A.J. Kearsley).

0893-9659/$ - see front matter Published by Elsevier Ltd
doi:10.1016/j.aml.2005.02.033

the functional form of data close to peaks or troughs. Once the data has been processed, for example peaks or troughs have been selected and the area underneath portions of the data have been calculated, there is still no standard or point of comparison [7,8].

The goal of this work is to present an algorithm with the potential of automatically identifying peak structure from raw mass spectrometer output without the use of smoothing, parameter specific filtering, or manual data analysis. This method requires no knowledge of peak shape and no pre- or post-processing of the data. Experience to date on *matrix-assisted laser desorption/ionization–time of flight mass spectrometry* (MALDI-TOF-MS) shows that the power spectrum of the noise cannot be predicted solely from the experimental conditions; therefore, blind application of smoothing and/or filtering algorithms may unintentionally remove information from the data. The new method does not have this failing. It does not require equal spacing of data points. It does require one single sensitivity parameter that can be accurately estimated. The sensitivity parameter's size can be bounded from below by knowledge of the ultimate resolution of the instrument and can be well approximated automatically using statistical properties of the raw data.

At present there is no single algorithm that will always and accurately identify peak structure in raw mass spectroscopy data without operator input. However an algorithm that produces output independent of *any* operator parameter selection or signal to noise estimation would be of tremendous benefit for the purpose of comparison (e.g. [11]).

2. Algorithm

In this section a two-phase algorithm is outlined. Described is a method for identifying what will be called *strategic points*, by solving a sequence of maximum orthogonal (Euclidean) distance problems [6]. Once these strategic points have been obtained, a nonlinear programming problem (NLP) is solved to find the optimal line segments which will constitute our solution.

Consider the collection of N *raw* data pairs, $D \in \Re^{N\times 2}$. Without loss of generality assume that the raw data, $D = [d_{ij}]$, with $i = 1 \ldots N$ and $j = 1, 2$, is strictly monotone in the first coordinate, $d_{11} < d_{21} < \cdots < d_{N1}$. In the case where raw data is not monotone it can be re-ordered or one can apply a simple isotonic regression (see [10,9]). Given any two pairs in the data set, say $(d_{k1}, d_{k2}) = d_k$ and $(d_{l1}, d_{l2}) = d_l$, one can define the line segment connecting them to be $s(d_k, d_l)$ and the set of points between d_k and d_l to be $\mathcal{I}(d_k, d_l) = \{d_j : k1 < j1 < l1\}$. Given a collection of data points D and a line segment, say $s(d_k, d_l)$, one can rapidly locate the data point(s) in $\mathcal{I}(d_k, d_l)$ that maximize(s) the orthogonal distance (see for example [5,2]) from $s(d_k, d_l)$ to the point(s). For simplicity, assume that there is only one point, say $\hat{d}_k$. Here $\hat{d}_k$ would solve

$$\max_{\hat{d}_k \in \mathcal{I}(d_k, d_l)} \operatorname{dist}(\hat{d}_k, s(d_k, d_l)) \tag{1}$$

and have optimal value, say $f(\hat{d}_k) \geq 0$. Our goal is to construct a piecewise linear approximation to the data that is accurate to within a tolerance, say τ. If $f(\hat{d}_k) \geq \tau$, then the point, $\hat{d}_k$, can become a new endpoint to two new line segments, $s(d_k, \hat{d}_k)$ and $s(\hat{d}_k, d_l)$, and the process can be continued [6] until $f(\hat{d}_k) \leq \tau$ for all data points. The tolerance τ can be estimated statistically for any given data set (see [12]). The collection of all points that solve problems (1) will constitute our set of *strategic points.*

Next, given a collection of, say, M strategic points $\hat{d}_m$, one can find the 'optimal' piecewise linear fit by solving an equality constrained nonlinear optimization problem as follows. Consider two adjacent strategic points, say $\hat{d}_p$ and $\hat{d}_{(p+1)}$, and assume that there are Q data points in $\mathcal{I}(\hat{d}_p, \hat{d}_{p+1})$, i.e., there

are Q non-strategic points between $\hat{d}_p$ and $\hat{d}_{p+1}$). The solution of the minimization problem

$$\min_{\hat{d}_{p2},\hat{d}_{(p+1)2}} \sum_{i=1}^{Q} \frac{1}{2}(d_{i2} - s(\hat{d}_{p2}, \hat{d}_{(p+1)2}))^2$$

finds the optimal height (or second coordinate) for the strategic points $\hat{d}_p$ and $\hat{d}_{p+1}$. Because a *continuous* piecewise linear function is sought, the constraints imposing continuity between solutions must be included. Given M strategic points one arrives at a nonlinear programming problem with M variables and $M - 1$ linear equality constraints enforcing that endpoints of adjacent line segments must be equal. The solution of this problem provides the optimal height, in the least squares sense, with respect to data between adjacent strategic points. Again, the problem is coupled through the continuity constraints that ensure a continuous piecewise linear function.

The algorithm can be stated as follows:

0 given D,
1 do while maximum $f(\hat{d}_k) < \tau$,
  – solve orthogonal distance problem (1) resulting in M strategic points $\hat{D}$,
2 solve nonlinear programming problem (with M variables and $M - 1$ constraints) adjusting second coordinate of the strategic points.

In theory, the problem of identifying the data point with maximum orthogonal distance may not yield a unique solution, but we have yet to observe this in numerical experimentation.

Upon completion of the algorithm one is left with a continuous piecewise linear approximation to raw data from which maxima and minima can more easily be extracted [1]. Once a peak and two adjacent troughs have been identified, the area underneath that peak can be approximated through a quadrature rule or by calculating the area of the polytope of strategic points between the two adjacent troughs.

3. Numerical results

In this section the numerical behavior of the algorithm is described. As a numerical example for this short work, we selected *polyethylene glycol* (PEG) for demonstrating the performance of the algorithm. This data set contains 19 772 pairs of data and was selected because it has essentially no baseline to contend with and therefore makes an excellent problem for demonstrating the 'peak-picking' aspect of the algorithm presented here. The algorithm has been applied to numerous other mass spectrometry data sets and a more comprehensive description of the numerical behavior can be found in [12].

The maximum orthogonal distance problem in the first step of the algorithm can be solved rapidly by sweeping through the data from left to right. The second step of the algorithm requires the solution of a nonlinear programming problem. Currently a sequential quadratic programming algorithm described in [4,3] is employed, although any large scale NLP algorithm would suffice.

The algorithm was coded in Fortran95 and is installed on a 450 MHz SPARCstation Ultra 80 using *IEEE floating point arithmetic (64 bit)*. When applied to the raw PEG data, the value of τ was estimated to be $\tau = 0.47234$. In addition, the algorithm was applied with four different selected values of τ. Obviously, different numbers of strategic points will result from various selections of τ, as shown in Table 1. However the numbers of peaks (and associated area approximations) were virtually identical for values of τ between 0.25 and 1.0.

Table 1

Values of τ	0.25	0.5	0.75	1.0	0.47234
Number of strategic pts.	8031	7856	6999	6251	7855
Number of peaks found	831	831	830	825	831
Elapsed CPU time (s)	18.84	16.12	15.03	14.67	16.66

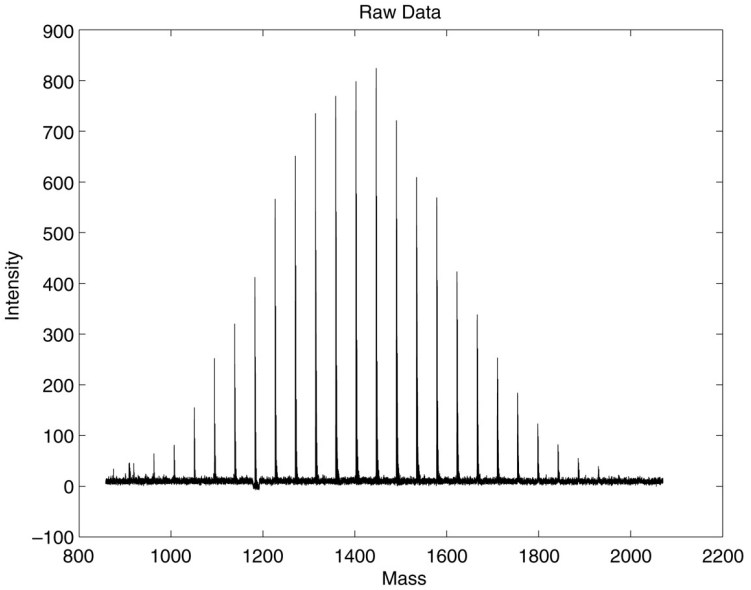

Fig. 1. Raw (PEG) data.

When plotted in their entirety, the raw data (Fig. 1) and the processed data (Fig. 2) appear to be identical. This illustrates, not surprisingly, that the algorithm results in a piecewise linear approximation to the data. Close examination shows that the processed data more clearly exhibits peaks and troughs. In Fig. 3 the solution closely follows the raw data; however, in Fig. 4 between mass values of 1844.5 u and 1845.75 u the solution identifies as a single peak what appears from inspection of the raw data to be three separate peaks.

Where this algorithm chooses a single parameter (which can be estimated statistically [12]), most other algorithms require far more parameter selections. The algorithm presented here is robust with respect to changes in the data and is completely reproducible. Solutions produced from this algorithm form an excellent tool for comparison.

4. Conclusions

We have presented an automated two-stage algorithm for rapid, robust and reproducible identification of peaks (and troughs) in raw mass spectrometry data. The algorithm does not rely on smoothing or parameter-driven filtering techniques; instead it requires only one parameter (which can be estimated directly from the data).

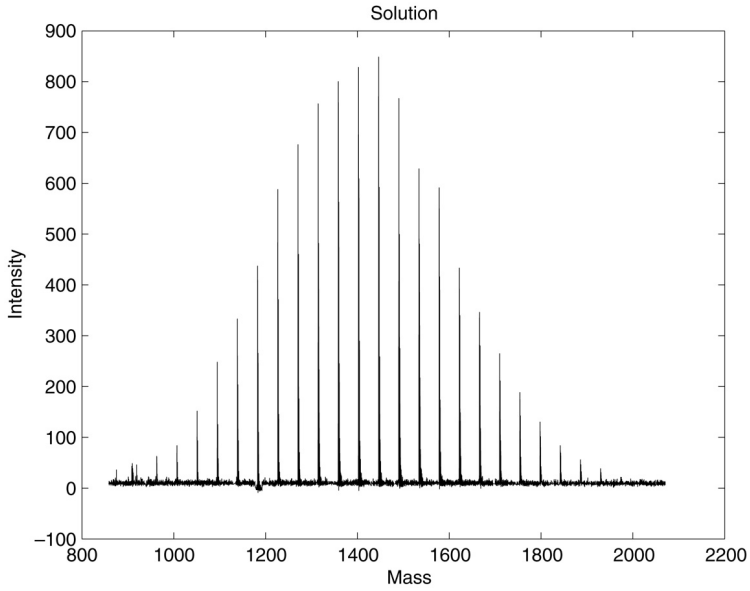

Fig. 2. Processed (PEG) data.

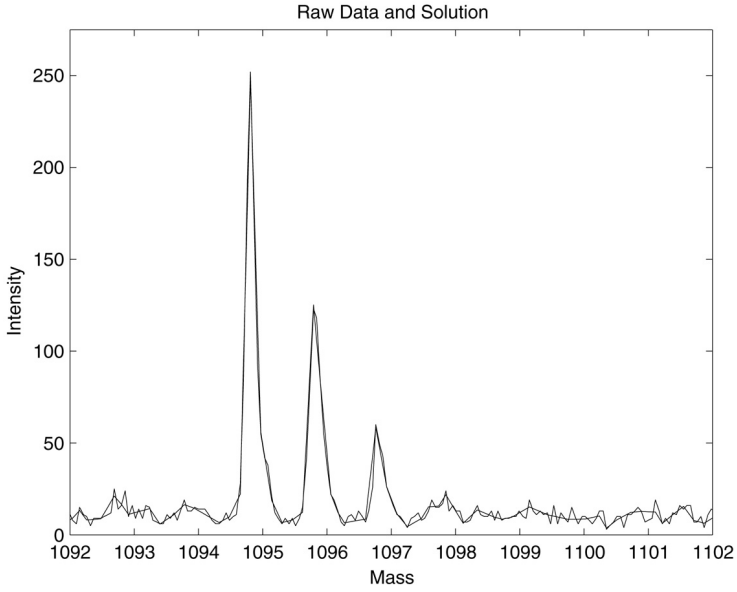

Fig. 3. Raw and processed (PEG) data.

The algorithm is very fast and produces reasonable results for wide ranges of the single parameter τ. For smaller values of τ, clearly the algorithm may incorrectly identify peaks on an order of magnitude less than or equal to the order of magnitude of the error or noise in the data. If τ is too large, very small peak structure may not be properly identified. However, the robustness and the reproducibility of this algorithm makes it a natural first choice for processing raw mass spectrometry data.

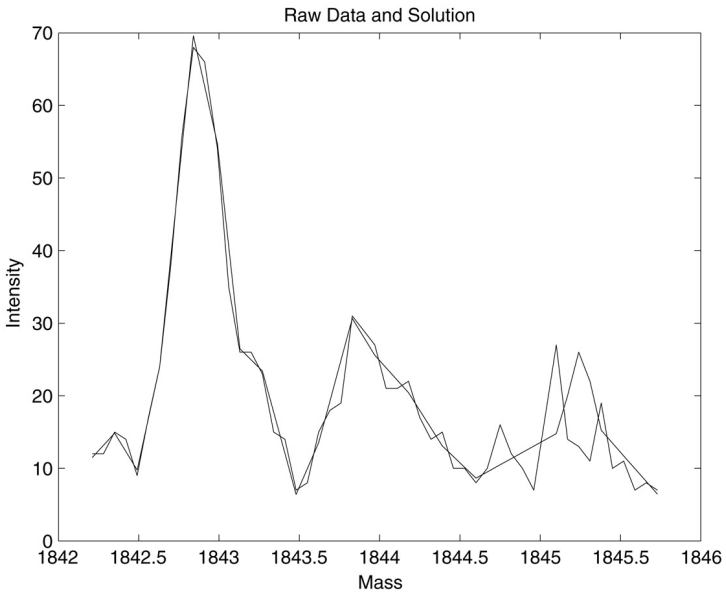

Fig. 4. Raw and processed (PEG) data.

References

[1] I. Barrondale, F.D.K. Roberts, An improved algorithm for discrete L1 approximation, SIAM J. Numer. Anal. 10 (4) (1993) 839–848.
[2] P.T. Boggs, R.H. Byrd, R.B. Schnabel, A stable and efficient algorithm for nonlinear orthogonal distance regression, SIAM J. Sci. Stat. Comput. 8 (1987) 1052–1078.
[3] P.T. Boggs, A.J. Kearsley, J.W. Tolle, A global convergence analysis of an algorithm for large scale nonlinearly constrained optimization problem, SIAM J. Optim. 9 (4) (1999) 833–862.
[4] P.T. Boggs, A.J. Kearsley, J.W. Tolle, A practical algorithm for general large scale nonlinear optimization problems, SIAM J. Optim. 9 (3) (1999) 755–778.
[5] P.T. Boggs, J.E. Rogers, Orthogonal distance regression, in: W. Fuller, P. Brown (Eds.), Contemporary Mathematics, vol. 112, American Mathematical Society, Providence, RI, 1990, pp. 183–194.
[6] D.H. Douglas, T.K. Peucker, Algorithms for the reduction of the number of points required to represent a digitized line or its caricature, Canad. Cartographer 10 (2) (1973) 112–122.
[7] C.M. Guttman, S.J. Wetzel, W.R. Blair, B.M. Fanconi, J.E. Girard, R.J. Goldschmidt, W.E. Wallace, D.L. Vanderhart, NIST-sponsored interlaboratory comparison of polystyrene molecular mass distribution obtained by matrix assisted laser desorption/ionization time of flight mass spectometry: Statistical analysis, Anal. Chem. 73 (2001) 1252–1262.
[8] S.D. Hanton, Mass spectrometry of polymers and polymer surfaces, Chem. Rev. 101 (2) (2001) 527–569.
[9] A.J. Kearsley, Projections onto order simplexes and isotonic regression, J. Res. Natl. Inst. Stand Technol. (in press).
[10] A.J. Kearsley, R.A. Tapia, M.J. Trosset, On the solution of the isotonic regression problem on parallel computers, in: S. Schaffler, H. Fischer, B. Riedmuller (Eds.), Applied Mathematics and Parallel Computing; Festschrift fur Professor Dr. Klaus Ritter, Physica-Verlag, Heidelberg, Germany, 1996, pp. 141–147.
[11] W.E. Wallace, C.M. Guttman, Data analysis methods for synthetic polymer mass spectrometry: Autocorrelation, J. Res. Natl. Inst. Stand Technol. 107 (2002) 1–17.
[12] W.E. Wallace, A.J. Kearsley, C.M. Guttman, An operator independent approach to mass spectrometric peak identification and integration, Anal. Chem. 76 (9) (2004) 2446–2452.
[13] J.T. Watson, Introduction to Mass Spectrometry, Lippincott Williams & Wilkins, USA, 1997.