

# A Numerical Method for Mass Spectral Data Analysis <sup>\*</sup>

Anthony J. Kearsley <sup>†</sup>    William E. Wallace <sup>‡</sup>    Charles M. Guttman<sup>‡</sup>  
Javier Bernal<sup>†</sup>

March 17, 2004

## 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 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. Promising numerical performance of this algorithm on raw data is presented.

## 1 Introduction

Modern mass spectrometers are capable of producing large, high-quality data sets in brief periods of time ([10]). It is not uncommon for a synthetic polymer to produce a spectra 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 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 the functional form of data close to peaks or troughs. Once the data has been processed, for example peaks or troughs have

---

<sup>\*</sup>Contribution of the National Institute of Standards and Technology and not subject to copyright in the United States.

<sup>†</sup>Mathematical and Computational Sciences Division, National Institute of Standards and Technology, Gaithersburg, MD, 20899-8910

<sup>‡</sup>Polymers Division, National Institute of Standards and Technology, Gaithersburg, MD, 20899-8541

been selected and the area underneath portions of the data have been calculated, there is still no standard or point of comparison ([5, 6]).

The goal of this paper is to present an algorithm with the potential to automatically identify 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/ionisation-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. However, it does require one single sensitivity parameter. This parameter's size can be bounded from below by knowledge of the ultimate resolution of the instrument and can be well approximated automatically by 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. [8]).

## 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 ([4]). 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,  $D \in \mathfrak{R}^{2 \times N}$ . Without loss of generality assume that the raw data,  $D = [d_{ij}]$ , is strictly monotone in the first coordinate,  $d_{11} < d_{21} < \dots < d_{N1}$ . In the case that raw data is not monotone it can be re-ordered or one can apply a simple isotonic regression [7]. 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  $s(d_k, d_l)$ . For any such line segment and any data point one can rapidly locate the point that maximizes orthogonal distance from  $s(d_k, d_l)$ , say  $\hat{d}_k$ . Here  $\hat{d}_k$  would solve,

$$\max_{d_{k1} \leq \hat{d}_{k1} \leq d_{l1}} \text{dist} \left( \hat{d}_m, s(d_k, d_l) \right) \quad (2.1)$$

and have optimal value, say  $f(\hat{d}_m)$ . The point,  $\hat{d}_k$ , can now 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 until the distance of the data point with greatest orthogonal distance from associated line segment falls beneath some prescribed tolerance, denoted here by  $\tau$ . The tolerance  $\tau$  can be estimated statistically for any given data set (see [9]). The collection of points that solve problems (2.1) will constitute our set of strategic points.

Secondly, given a collection of, say  $M$ , strategic points  $\hat{d}_m$ , one can find the ‘optimal’ piecewise linear fit by solving an equality constrained NLP. Consider a two adjacent strategic points, say  $\hat{d}_p$  and  $\hat{d}_{(p+1)}$  and further assume that there are  $Q$  data points above and beneath the line segment connecting  $\hat{d}_p$  and  $\hat{d}_{(p+1)}$  (i.e. there are  $Q$  non-strategic points between  $\hat{d}_p$  and  $\hat{d}_{(p+1)}$ ). The solution of the minimization problem,

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

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

The algorithm can be stated as follows:

0 Given  $D$  and  $\tau$

1 Do while maximum  $f(\hat{d}_m) < \tau$ ,

– Solve orthogonal distance problem (2.1) 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 this does not pose any difficulty in practice as it has yet to be observed in numerical experimentation.

Upon completion of the algorithm one is left with a continuous piece-wise 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.

Values of $\tau$	0.25	0.5	0.75	1.0
Number of strategic pts.	8031	7856	6999	6251
Number of peaks found	831	831	830	825
Elapsed CPU Time (secs)	18.84	16.12	15.03	14.67

### 3 Numerical Results

In this section the numerical behavior of the algorithm is described. As a numerical example for this short paper, we selected *polyethylene glycol* (PEG) to demonstrate the performance of the algorithm. This data set contains 19772 pairs of data and was selected because it has essentially no baseline to contend with and therefore makes an excellent problem to demonstrate 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 is currently in preparation ([9]).

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 [3, 2] 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 algorithm identified different numbers of strategic points for various selections of  $\tau$ , as shown in Table 3. However the number of peaks (and associated area approximations) were virtually identical for values of  $\tau$  between 0.25 and 1.0.

When plotted in entirety, the raw data (Figure 1) and the processed data (Figure 2) appear to be identical. Closer examination shows that the processed data more clearly exhibits peaks and troughs. In Figure 3 the solution correctly identifies all peak structure with little ambiguity; however Figure 4 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 [9]) 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.

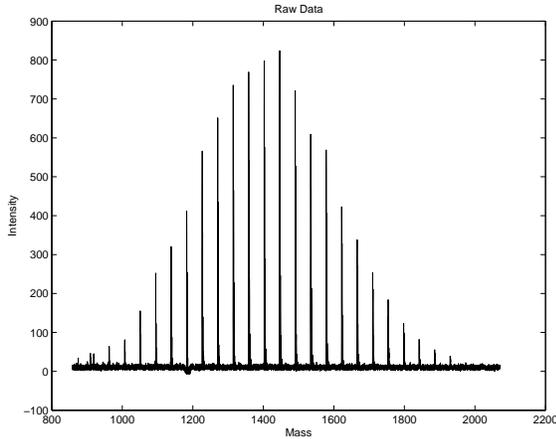


Figure 1: Raw (PEG) data

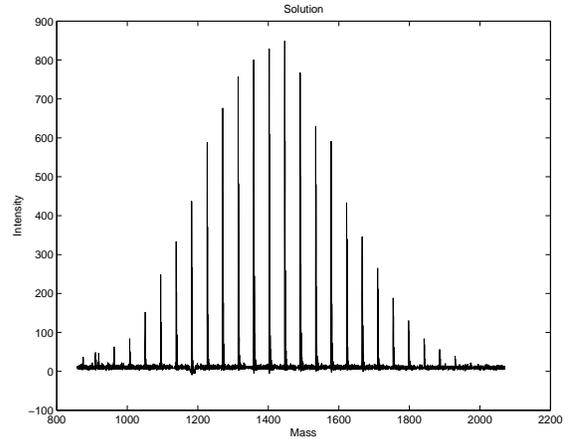


Figure 2: Processed (PEG) data

## 4 Conclusions

We have presented an automated two-stage automatic 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).

The algorithm is very fast and produces reasonable results for wide ranges of the single parameter  $\tau$ . For smaller values of  $\tau$ , clearly the algorithm may incorrectly identify peaks on the order of magnitude less than or equal to the order of magnitude of error or noise in the data. If  $\tau$  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.

## References

- [1] I. Barrondale and F. D. K. Roberts. An improved algorithm for discrete L1 approximation. *SIAM J. Numer. Anal.*, 10(4):839–848, 1993.
- [2] P. T. Boggs, A. J. Kearsley, and J. W. Tolle. A global convergence analysis of an algorithm for large scale nonlinearly constrained optimization problem. *SIAM J. Optim.*, 9(4):833–862, 1999.

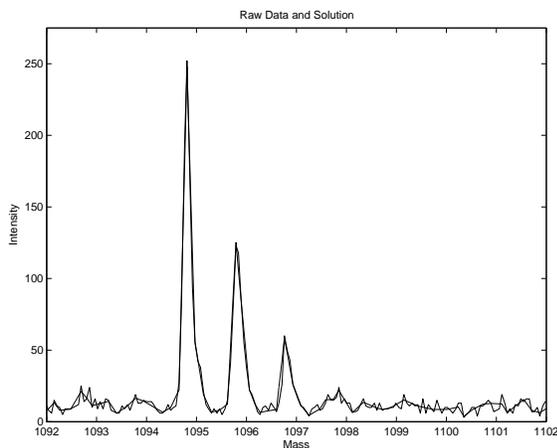


Figure 3: Raw and processed (PEG) data

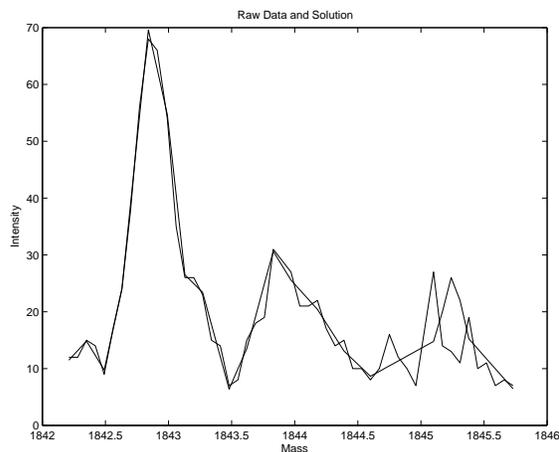


Figure 4: Raw and processed (PEG) data

- [3] P. T. Boggs, A. J. Kearsley, and J. W. Tolle. A practical algorithm for general large scale nonlinear optimization problems. *SIAM J. Optim.*, 9(3):755–778, 1999.
- [4] D. H. Douglas and T. K. Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *The Canadian Cartographer*, 10(2):112–122, 1973.
- [5] C. M. Guttman, S. J. Wetzel, W. R. Blair, B. M. Fanconi, J. E. Girard, R. J. Goldschmidt, W. E. Wallace, and D. L. VanderHardt. NIST- sponsored interlaboratory comparison of polystyrene molecular mass distribution obtained by matrix assisted laser desorption/ionization time of flight mass spectrometry: Statistical analysis. *Anal. Chem.*, 73:1252–1262, 2001.
- [6] S. D. Hanton. Mass spectrometry of polymers and polymer surfaces. *Chemical Reviews*, 101(2):527–569, 2001.
- [7] A. J. Kearsley, R. A. Tapia, and M. J. Trosset. On the solution of the isotonic regression problem on parallel computers. In S. Schaffler H. Fischer, B. Riedmuller, editor, *Applied Mathematics and Parallel Computing; Festschrift fur Professor Dr. Klaus Ritter*, page TBA, Heidelberg, Germany, 1996. Physica-Verlag.
- [8] W. E. Wallace and C. M. Guttman. Data analysis methods for synthetic polymer mass spectrometry: Autocorrelation. *NIST Journal of Research*, 107:1–17, 2002.

- [9] W. E. Wallace, A. J. Kearsley, and C. M. Guttman. An operator independent approach to mass spectrometric peak identification and integration. *Anal. Chem.*, accepted.
- [10] J. T. Watson. *Introduction to Mass Spectrometry*. Lippincott Williams & Wilkins, USA, 1997.