Advanced Numerical Methods for Polymer Mass Spectral Data Analysis

William E. Wallace*, Charles M. Guttman*, Anthony J. Kearsley+, Javier Bernal+

*Polymers Division and +Mathematical & Computational Sciences Division
National Institute of Standards and Technology, Gaithersburg, MD 20899-8541 USA

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. These new algorithms must work with minimal user input, both to save operator time and to eliminate operator bias. This second criterion is crucial for NIST's goal of creating a synthetic polymer Standard Reference Material for absolute molecular mass distribution measurement by mass spectrometry. Toward this end a mathematical algorithm is presented that accurately locates and calculates the area beneath peaks using only reproducible mathematical operations and a single user-selected sensitivity parameter.

A non-linear programming algorithm using an L_2 approximation to an L_1 fit was employed [1]. Given a dataset of N points we find a collection of strategic points and find the unique optimal piece-wise linear function passing through the x coordinate of each strategic point. This defines a set of function maxima and minima corresponding to the peak maxima and the peak limits. The original data is then integrated under each peak maximum between the two adjacent minima.

The method reported on in this presentation is a two-step algorithm. The first portion of the algorithm requires the selection of *critical* or *strategic* points. These points are selected based on an iterative procedure that identifies points whose orthogonal distance from the end-point connecting line segment is greatest. Once a point with greatest orthogonal distance from the mean has been identified, it joins the collection of strategic points and, in turn, becomes an end-point for two new line segments from which a point with greatest orthogonal distance. This numerical scheme is performed until the greatest orthogonal distance to any end-point connecting line segment drops beneath a prescribed threshold value. This threshold value is the only algorithmic parameter. Clearly the selection of these points does not require equally spaced data. The second phase of the algorithm requires the solution of an optimization problem, specifically, locating strategic point heights (or adjusting strategic y-axis values associated strategic x-axis values) that minimize the sum of orthogonal distance from raw data. This problem is a nonlinear (and non-quadratic) optimization problem that can be accomplished quickly using a modern nonlinear programming algorithm (e.g. [2]).

This method requires no knowledge peak shape and no preprocessing of the data, i.e. smoothing. Our experience 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. Method does not require equal spacing of data points. However, it does require the operator to choose the sensitivity parameter. This parameter's size can be bounded from below by knowledge of the ultimate resolution of the instrument; future work will focus on methods to choose this parameter. A publicly accessible, secure Web application for on-line, real-time application of the algorithm is planned, as well as an automated baseline compensation algorithm.

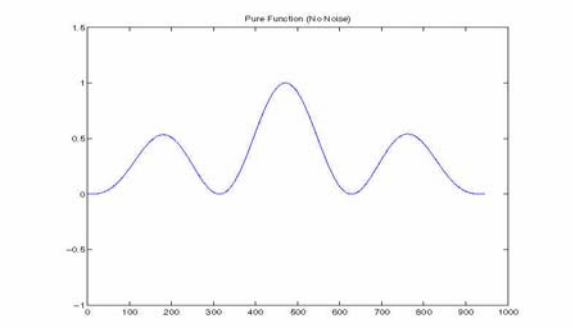
[1] I. Barrondale and F. D. K. Roberts, ``An Improved Algorithm for Discrete L1 Approximation'', SIAM J. Numer. Anal. 10 (1993) 839.

[2] P. T. Boggs, A. J. Kearsley and J. W. Tolle, ``A practical algorithm for general large scale nonlinear optimization problems'', SIAM J. Opt. 9(3) (1999) 755.

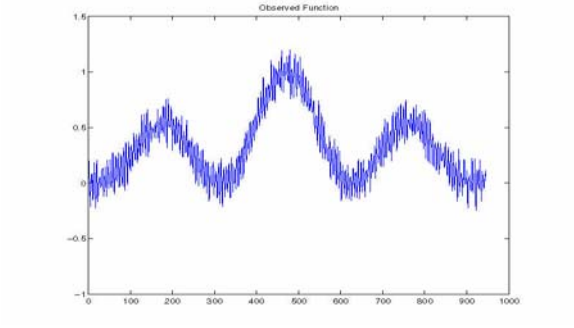
Algorithm

Consider a fictitious three-peak scenario with added noise.

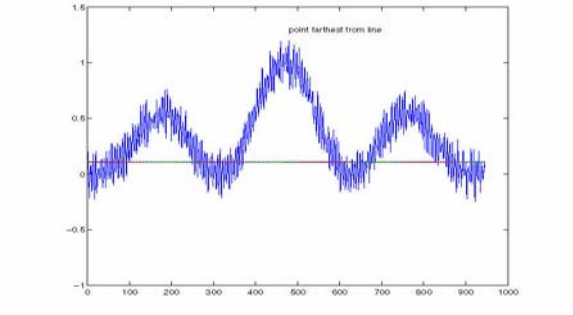
A) Original data



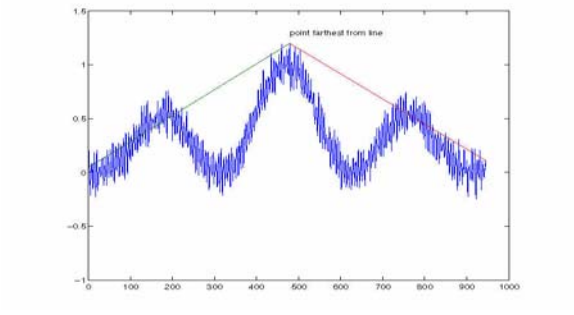
B) With added noise



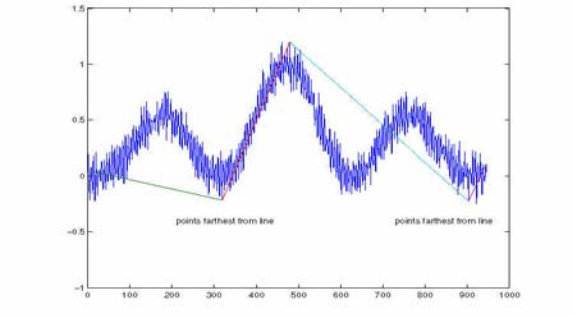
C) Take 1st & last points



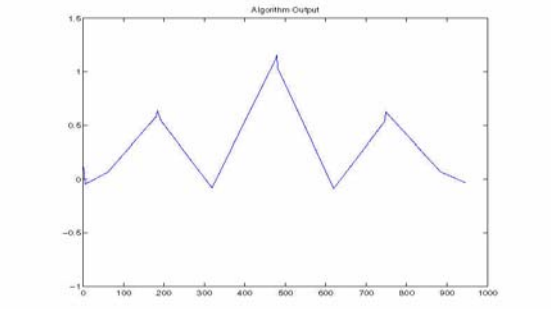
D) Find point greatest distance normal to line



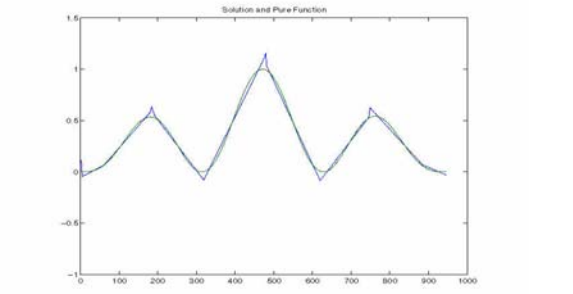
E) Iterate to sensitivity limit



F) Output peak positions and widths



F) Comparison to original data



Results:

Analytical Areas (exact values)

100.	**(left peak)**
175.	**(middle peak)**
100.	**(right peak)**

Algorithm Calculated Areas

101.609	**(area +1.6%)**
178.591	**(area +2.0%)**
102.873	**(area +2.9%)**