2020 American Control Conference
Denver, CO, USA, July 1-3, 2020

Improved Peak Detection for Mass Spectrometry via Augmented Dominant Peak Removal

Daniel Y. Abramovitch*

Abstract— **In several common measurement modes of mass spectrometry systems, the measurements produced are an ordered pair of abundance (an amplitude) versus mass to charge ratio (m/z). This mass spectrum can be viewed as delta functions comprising actual abundance of the ions with that m/z value convolved with a a smearing function due to the measurement process. Peak detection refers to the method of extracting estimates of these precise delta functions (mass locations) and amplitudes from this smeared response. Current peak detection and centroiding in mass spectrometry is particularly susceptible to errors when there is significant overlap between peaks. This paper explains the issues with current methods and presents a set of algorithms inspired by curve fitting and system ID methods in control [1] that dramatically reduce these issues. The algorithms are computationally simple, suitable for implementation in the embedded system of an analytical instrument, and produce dramatically improved results in the peak center estimates, particularly when there is significant peak overlap in the measured peak spectrum.**

I. Introduction

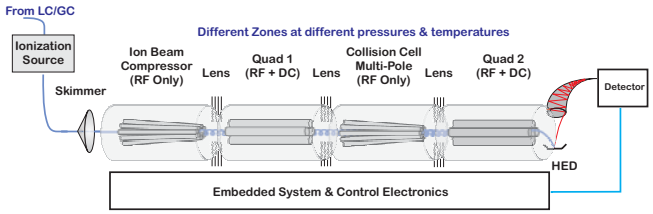

Fig. 1. A basic diagram of a quadrupole mass spectrometer.

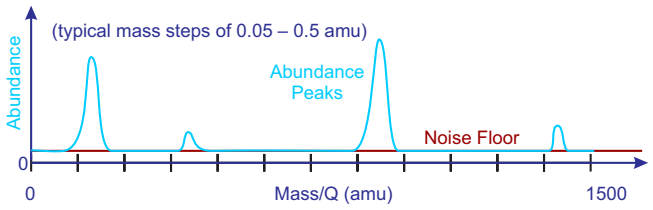

Fig. 2. A simple diagram of a mass scan or profile. The sweep of the mass spectrometer produces a map of ion intensity (abundance) versus mass-to-charge ratio (m/z). Ideally, the abundance bumps would be scaled Kronecker delta functions at a particular m/z value, but the convolution of these deltas with the measurement/instrument process3 produces a smoothed bump. We would like to localize these, or find the centroid.

A mass spectrometer is an analytical instrument used by scientists to characterize unknown chemical compounds and/or to quantify and qualify known compounds. While there are many variants of this particular idea, there are two broad classes of mass spectrometers: those that rely on the "time-of-flight" of an ion (TOF, where for a given electromagnetic impulse, lighter ions arrive at the detector earlier than larger, heavier ones) and those that rely on some sort of electromagnetic "mass filter" to pass ions that have a particular mass-to-charge ratio [2], [3]. In the latter group, there are various types of mass filters, including ion traps and quadrupoles. The latter group is analogous to electronic spectrum analyzers [4], while the former generally captures data in the same way as a real-time digital oscilloscope [5]. The diagram of Figure 1 shows a tandem or triple quad mass spectrometer, which has two quadrupole mass filters sandwiching a collision cell. The collision cell was originally implemented using quadrupole devices as well, but has since been replaced by devices with 6, 8, or more rods, which provide a better response at a lower cost. The term "triple quad" remains from this original configuration.

The quadrupole is a device which has four parallel conductive rods of alternating polarity, meaning that rods across from each other are at the same voltage, while adjacent rods are 180° out of phase with each other. For a given combination of DC and RF (sinusoidal) voltages at a given RF frequency, the quadrupole produces an electric and magnetic field in the center that allows only ions with a particular mass-to-charge ratio range to pass through [6]. The required DC and RF settings for each m/z range are calculated as a stable region of the solution of the Mathieu Equation [6]. Considerable effort is devoted to the design and manufacture of the quadrupole, as it is the key enabler of mass resolution (analogous to frequency resolution in a spectrum analyzer) [4].

Compounds are initially separated in a variety of methods, such as liquid (LC) or gas (GC) chromatography. The output of these instruments is fed into the source of the mass spectrometer. Ions are generated in the source and the ion beam is compressed as it enters into the ion optics path and sent to the first quad. Various electromagnetic filters are used along the ion optics path to control the effects of fringing fields that can adversely affect the measurement.

From the first quad the ions enter the collision cell, an RF device which holds an inert gas. The ions that make it through the first quadrupole (quad) impact the gas molecules in the collision cell and fragment. The product ions are then filtered by the second quad before arriving at the detector. The detector is a device such as an electron multiplier tube, that produces a cascade of electron emissions for each ion impact, amplifying the signal so that the resulting output can be sampled and digitized. This digitized signal is a measure

*Daniel Y. Abramovitch is a system architect in the Mass Spec Division at Agilent Technologies, 5301 Stevens Creek Blvd., M/S: 3U-WT, Santa Clara, CA 95051 USA, danny@agilent.com

of the number of ions at that mass-to-charge ratio (m/z).

By knowing the m/z settings for each quad – in Atomic Mass Units (AMU) or equivalently Daltons (Da) – as well as the collision energy, a map of source and product ions can be formed. Analytical chemistry then allows the scientist to back out the probable fracture points of different candidate molecules and divine the original input to the instrument. In this mode, it is a true compound identification system.

In mass spectrometers, the measured abundance data is registered against mass to charge ratios (m/z). The measured curve is typically considered to be the result of smearing of true m/z delta functions convolved with the physical and electronic response of the instrument (Figure 2). Identifying m/z lines is a significant part of identifying compounds. One of the main issues in complex mixtures is identifying lines when peaks significantly overlap each other. Even if an apex can be found, the width estimates needed for proper centroid calculations are often wrong. The algorithms described here are designed to address these issues.

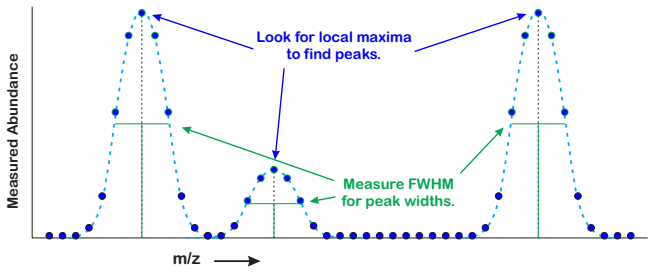

Fig. 3. Simple detection of well separated peaks. The blue dots are the sample points along the mass axis, while the dashed cyan line is the inferred abundance curve. The peaks are initially detected by searching for local maxima and their widths are then characterized, typically by searching for the full-width, half-max (FWHM) points.

There has been a fair amount of work on different peak finding methods. A nice comparison of them is found in [7], [8]. Some more advanced methods are discussed in [9]. Most of them involve finding some local maxima relative to surrounding points, and then qualifying those maxima, either with a threshold or through some peak curve shape requirements. Alternately, the slope of the abundance curve can be checked for negative zero crossings, indicating a local maxima [8]. In a well-tuned system, with isolated peaks (Figure 3), this can be accomplished fairly simply. Of course, as differentiation amplifies noise such methods are often combined with filtering. Another piece of pre-processing that gets significant attention in the literature is baseline correction, particularly in cases where the baseline is not a constant, but instead can be modeled as a smooth curve with some low order polynomial.

Simple peak detection involves searching for the local maxima or apexes of the abundance curve, labeling those above some minimal peak threshold, then determining the peak width, and center (via a centroid or other method). A very simple algorithm can be described as follows. Consider a set of N ordered pairs of data, $\{(m_i, a_i)\}$ in Figure 3 where m_i is the mass to charge ratio (m/z) at index, i, and a_i is the measured/detected ion level or abundance. We assume that the m_i values are monotonic in i, but this can be rising or falling. In a 3-point search a peak is detected if

$$a_{i-1} < a_i \text{ and } a_i > a_{i+1}, \tag{1}$$

for $0 < i < N - 1$ and if $a_i \geq$ peak threshold. In a 5-point search a peak is detected if

$$a_{i-2} < a_{i-1} < a_i \text{ and } a_i > a_{i+1} > a_{i+2}, \tag{2}$$

for $1 < i < N - 2$ and if and if $a_i \geq$ peak threshold. Clearly, the 3-point search will allow more potential peaks than the 5-point search, but in the presence of measurement uncertainty, the latter is more likely to have matched a real peak and not a noise artifact. We can also look for 3-point peaks and further qualify them as 5-point or higher peaks. We are looking for an apex the response (abundance) data in a computationally simple way (i.e. a way that can be readily implemented in an embedded system).

Because such simple searches are subject to noise and other measurement artifacts, it is reasonable to want to qualify the candidate peaks further. This is often done by searching for the width of the peak and using this to help calculate a centroid of the peak. As this centroid would generally involve more measurement points, one might consider it to add some noise immunity.

There are many possible ways of evaluating width, but a fairly common one is to search down the curve for the full width at half max (FWHM) [10], where the width is assumed to be the width where the curve reaches half of the maximum peak value. Simply put, for any identified peak apex, a_i, we search downhill on either side until we reach the points $(m_j, a_j), j < i$ and $(m_k, a_k), k > i$ where the measured values, $a_j, a_k < a_i/2$. The FWHM is then $|m_j - m_k|$. Since we are operating at discrete measurement points, some straightforward refinement can be achieved by interpolating the m_j, m_k values back towards the actual point where the abundance equals $a_i/2$. Further refinement may be achieved by interpolating a smooth curve onto the top few points of a peak and extracting an improved apex from there. This is standard practice. What happens when one or more peaks overlay our original one? That is, what happens when we are searching down for our points j and k from the peak center at index i and encounter another peak center at ii before we get to our unknown j or k points? The simple methods above would not take that "unmodeled dynamic" or "unmodeled feature" into account, resulting in a mis-estimate of the peak width and potentially the peak location and refined abundance level.

The methods in this paper are not about improving the search method specifically, but about being more careful with the classification methods. In other words, we sanity check the model and when we see that our underlying model assumptions are violated, we modify that model. It turns out that by segmenting the search, the verification steps can be made simple enough to not dramatically increase the computational cost.

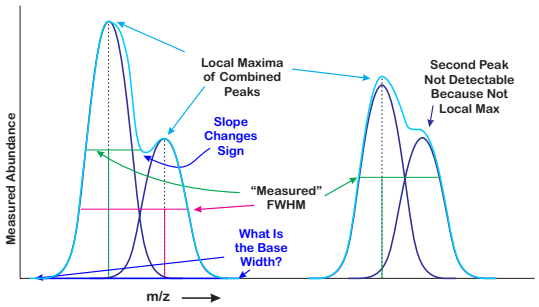

Fig. 4. Peaks that are too close to each other can interfere with detection methods.

When peaks are close together, as can happen with complex, unknown compounds, the underlying peaks may create distributions that interfere with each other. Two examples of this are sketched in Figure 4. On the left we see two overlapping peaks where one can find separate local maxima, but the calculation of the widths - based on a search of the abundance curve for the half-max point - are incorrect. As peak widths are used to calculate the centroid of the peak, which has more noise immunity than the apex point in determining the center of the peak, an erroneous width can result in poor mass assignment. On the right, the peaks are close enough together that there is no second apex and thus the second peak is not detected.

There has been much work on measuring and correcting for biases, for filtering out noise, for removing outliers, for modeling individual peak shapes, and for comparing libraries of known peaks to measured curves [8]. However, there has been little discussion on the problems shown in Figure 4, especially when discussing a full mass axis with substantial numbers of unknown potential peaks. Intuitively, the problem can be framed as a least squares fit, e.g. LevenbergMarquardt algorithm, of peak shapes to the abundance curve, but the computational complexity for this goes up as the cube of the number of candidate peaks [11]. This makes such an approach one more suited to off-line, post-processing, rather than for use in an on-line instrument.

We present an iterative method of successive peak identification and removal, a "divide and conquer" approach which results in considerable improvement in peak detection and center/width characterization in the presence of interfering side peaks. The methods here do not involve a substantial increase in computational load [12], [13].

The above problem is easily identifiable to a controls engineer as signal identification problem where the old method lacked the model elements needed for a correct identification. While the model in this case is not a dynamic system model, but instead one of the output of a dynamic system (the mass spectrometer), one can see that the extracted model (the peak centers, heights, and widths) is severely affected by the richness of the model which one assumes is producing the underlying data. Thus, this paper does not present new control methodologies. Instead it demonstrates how a seemingly unrelated identification problem can be solved using the mentality of a control engineer. Furthermore, the thinking put in to minimize the computational load derives directly from an understanding of real-time programming for control applications.

What we will demonstrate is that by using a simple decision process to switch between likely models, we can dramatically improve the results in the cases of these overlapped peaks. These switches allow a new model, a multi-peak cluster model, to be identified and removed using relatively simple operations. The residual measurement curve – left after the removal of the modeled peaks – is then searched again for new peaks to be modeled and removed. By segmenting the measurements into separate regions where non-trivial abundance has been found, the search space is reduced, making the iteration much faster, and thus more suitable for operation on real-time data.

II. Segmenting the Mass Axis

The set of algorithms presented here follows the same "iterative feature removal strategy" in some of the author's previous work [1], [14]. That is, on a measurement curve (such as those in Figures 2 or 3), a particular feature is localized, and fit to a low order model. That model is then used to generate a curve over the same domain as the measurement, and is "removed" from the measured curve. The residual curve is then scanned for the next feature to be modeled, fit, and removed. The process continues until there are no more significant features in the residual curve. The issue with this, is that if there are a significant number of sample points, the iteration time can become huge. For example in our mass spectrometry context, if the mass axis is 2000 AMU wide and the mass steps are at 0.2 AMU, then the curve has 10,000 points. The saving grace is that if we can break the mass axis up into smaller subsets and only iterate over those small subsets, then the computational cost can be significantly reduced. Again, the utility of taking such a step is clear to someone with sensitivity to time delay in real-time systems. We see how a real-time control perspective leads us here.

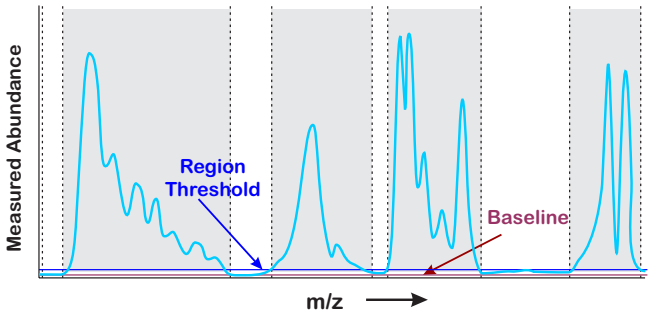

Fig. 5. Segmenting the mass axis into regions with interesting abundance and those without.

The first step to creating an iterative method is to segment the mass axis into alternating regions, holding interesting and non-interesting levels of abundance, respectively. This is diagrammed in Figure 5. We chose a threshold level above the noise baseline. We break the mass axis into regions

where the abundance is sustained above this threshold and those where it is not. Clearly, the latter regions serve no purpose in peak finding and can be ignored, but the former regions can be treated in separate peak searches. Thus, any sort of iterative search and removal algorithm iterates over a much smaller region (unless the abundance is high across the entire mass axis). In most experiments we have done, there are significant opportunities to significantly segment the mass axis, resulting in local search regions that have a few hundred, rather than a few thousand points. Thus, even though an iterative search is more computationally expensive than a single pass, the segmentation of the mass axis greatly reduces this effect.

III. Successive Dominant Peak Removal

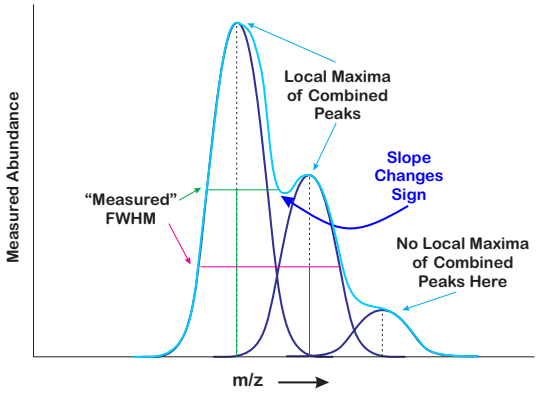

Fig. 6. Three slightly overlapping peaks.

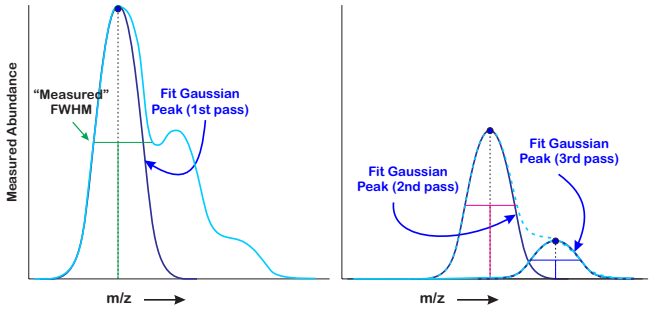

Fig. 7. Applying Successive Dominant Peak Removal to the peaks of Figure 6.

The example diagrammed in Figure 6 shows three peaks that overlap, but for which there is a dominant peak (the one on the left). The center apex can be identified, but because its FWHM point to its left is overrun by the larger left peak, any search-based determination of the center peaks width is flawed. The situation is worse for the far-right peak, which is so dominated by the spread of the center peak that we cant even see an apex.

The basic idea of Successive Dominant Peak Removal as applied to the diagram in Figure 6 is diagrammed in Figure 7. It is similar to an approach discussed in [15]. We use a canonical peak shape to model the largest peak in a region. A Gaussian shape is a convenient approximation because it can be estimated over the entire axis of a region by determining its center point and its width. Using the apex center and the dominant peak's width, we estimate the height as the difference between the peak apex and the baseline, and the FWHM by searching the peak curve. From these, we can calculate the μ, amplitude scale , and σ of a Gaussian curve. A unit height Gaussian curve centered at 0 looks like

$$G(x) = e^{-\frac{x^2}{2\sigma^2}}, G(0) = 1. \tag{3}$$

At half the maximum

$$G(x_{HM}) = e^{-\frac{x_{HM}^2}{2\sigma^2}} = \frac{1}{2}, \tag{4}$$

so that

$$\sigma = \frac{1}{\sqrt{2\ln 2}} x_{HM}. \tag{5}$$

As x_{HM} is the distance from the center point of the peak model to the half max point, we can average the measurements from both sides, if they are available. If there is a side peak on one side, then the measurement from the other side can be used or averaged with the side peak width. In any event, it is a simple calculation to compute a local Gaussian peak model from readily measured quantities about the peak in the residual measurement curve.

The peak model is shown on the left side of Figure 7. Removing this peak model from the abundance curve leaves the dashed curve on the right, which can then be searched to locate a second peak. The removal of the previous dominant peak has left us in a better position to estimate center peaks width, model it with a Gaussian, and remove it. The third pass, also on the right side of Figure 7, shows that after removing the first two peaks, we have revealed the previously hidden third peak. This method works extremely well, provided that the dominant peak of any particular iteration, is well determined or separated from any interfering side peaks.

IV. Issues Caused by Interfering Peaks

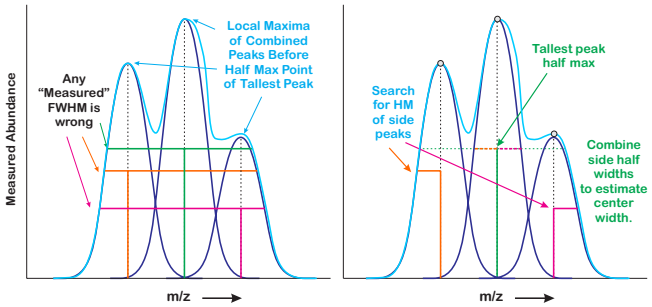

Fig. 8. When tallest peak is sandwiched between two others.

The main failure mode for Successive Dominant Peak Removal can be seen in the diagram of Figure 8. On the left we see that if interfering side peaks are some combination of close enough and large enough, the search for the half-max points of the so-called dominant peak will result in finding

locations associated with the side peaks. However, if we have tabulated all the apexes found in our current peak search iteration, we already have the indices of the side peaks. We can compare these to the mass axis indices of our half-max searches, and if we cross another peak index on the way to our half-max point, we declare a side peak on that side. On the right of Figure 8, we see that once we cross those side peaks, we switch to searching for their half-max points, rather than those of the dominant peak. Depending upon the number of side peaks found, the width of the center peak can then be estimated from either a successful half-max search on one side, or from combining the estimated widths of side peaks. Looking at the diagram on the right of Figure 8, we would end up with models for three closely spaced peaks.

V. Successive Multi-Peak Removal

We are now in a position to describe an algorithm that can better deal with interfering side peaks: Successive Multi-Peak Removal. The search algorithm starts the same way as Successive Dominant Peak Removal, but as we search down from the center of the largest peak to find the half-max points, we keep track of possibly crossing the indices of other found peak apexes. In the event that we do find a side peak before the half-max point, we model the side peak with a second Gaussian. Thus, instead of removing a single peak Gaussian model from the abundance curve, we might remove up to a three-peak model. We then start a new iteration of search and removal on the residual abundance curve. This type of iteration might be computationally expensive, had we not segmented the mass axis into small regions of significant abundance content, as described in Section II. The iterations are only done over those regions, dramatically decreasing the computational burden.

A few heuristic steps also improve the results:

- Since the superposition of individual Gaussian models scaled to the abundance height will not account for the effect of overlap, we scale the combined peak model curve so as not to be greater than the original curve.
- As side peaks may themselves have side peaks, we allow for each side peak (left and right) to have one more side peak of its own. The outside peak models are used to help scale our multi-peak model, but not actually included in the multi-peak removal. This provides a reasonable tradeoff between scaling and algorithm complexity.
- One of the properties of our assumed Gaussian peak model is that the magnitude of the slope is higher down the sides that at the top, so that if we detect an decrease in the slope magnitude before the half max point, and there is no side peak, a hidden peak might be found. Thus, we can detect some peaks even if their apexes were originally hidden by the skirt of a larger peak.

VI. Experimental Evaluation of Algorithms on Lab Data

In this section we will present an evaluation and comparison of three algorithms on laboratory data:

1) A simple apex and width search algorithm,
2) Successive Dominant Peak Removal, and
3) Successive Multi-Peak Removal.

We will display three plots of the same data: one for each algorithm over the same data range. This will demonstrate how each of the algorithms performs for different types of peak shapes.

The experimental data here is taken from an Agilent Ultivo Tandem Quad Mass Spectrometer [16]. The measured sample was Polypropylene Glycol (PPG), with an average molecular weight of 1000 AMU. The data had content from 10 to 1400 AMU. The data is saved in an unfiltered form and read in to Matlab. The data then has outliers removed and is smoothed using a Finite Impulse Response (FIR) filter to prepare it for peak finding. We apply each of the above algorithms to the same data set. The plot regions were chosen to emphasize the differences in results produced by these algorithms. The identified peaks are denoted by a black vertical line with the peak height denoted by an open circle. The estimates of FWHM are denoted by horizontal line segments with open circles at each end in the middle of the curve. The base width estimate is denoted by a similarly colored line segment near the bottom of the curve. We will only display curves with significant peak overlaps since this is most effective in showing the advantages of the new algorithm.

In particular, the plots for the simple apex and width search algorithm (Figures 9, 12, & 15) show broad horizontal line segments which are a direct artifact (and evidence) of the peak widths being mischaracterized due to the presence of other interfering peaks. The plots for Successive Dominant Peak Removal (Figures 10, 13, & 16) are immune to this, but miss peaks because the model is not accounting for close-in overlapping peaks of similar size. The plots from Successive Multi-Peak Removal (Figures 11, 14, & 17) show none of these artifacts and are pleasantly boring and logical in what they show. For a measurement system, boring and logical are an indication of success.

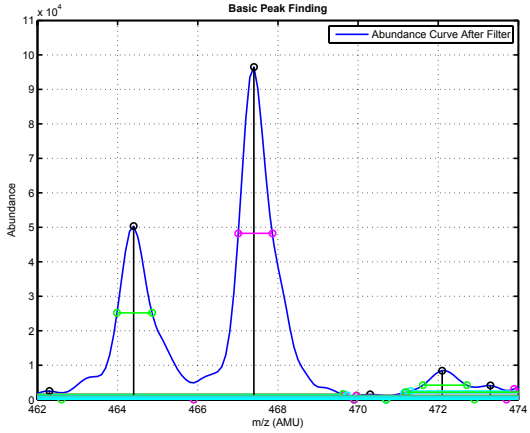

Fig. 9. Simple apex and width search applied between 462 and 474 AMU. The larges peaks are easy to find, and their FWHM are easy enough to calculate. However, even the small overlap between peak bases results in an inability to resolve the peak base widths.

The first data range is between 462 and 474 AMU,

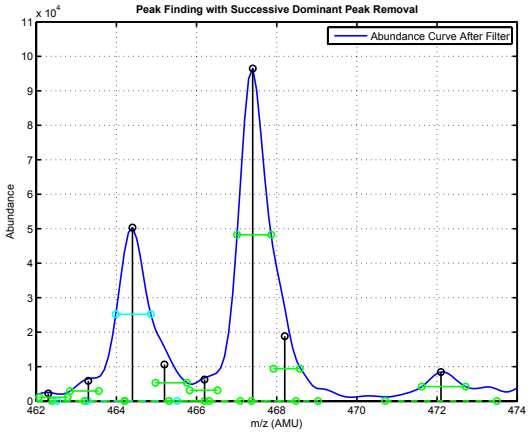

Fig. 10. Successive Dominant Peak Removal applied between 462 and 474 AMU. Using the model, we are able to estimate the FWHM of the larger peaks, but also find their base widths and several smaller hidden peaks.

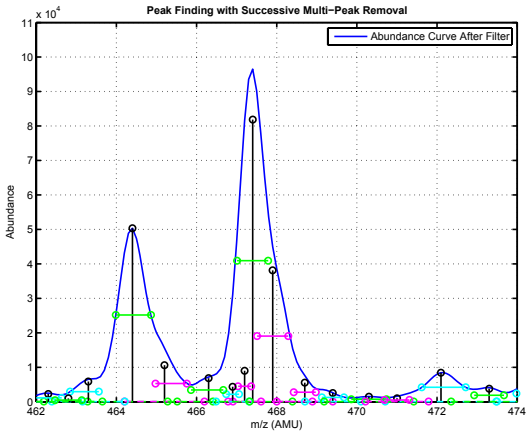

Fig. 11. Successive Multi-Peak Removal applied between 462 and 474 AMU. In this case, the algorithm is looking for side peaks from the start and thus is better able to find and characterize more of them.

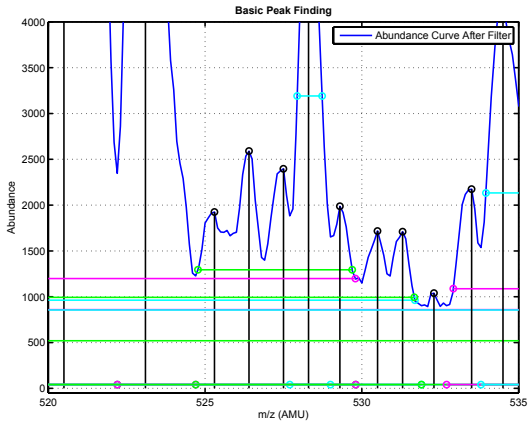

Fig. 12. Simple apex and width search applied between 520 and 535 AMU. Note how most of the peak half-max widths are mischaracterized, due to the tight clustering of peaks.

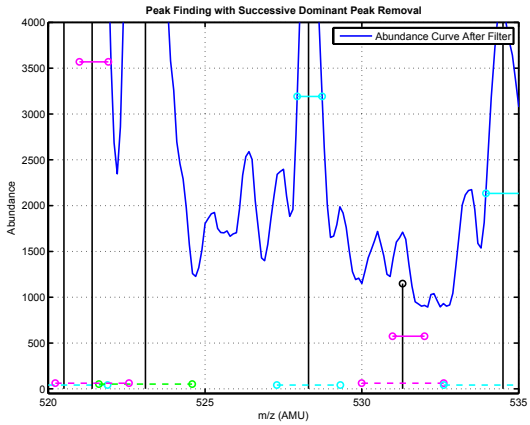

Fig. 13. Successive Dominant Peak Removal applied between 520 and 535 AMU. Peak widths are less likely to be overestimated, but a close side peak may result in that peak being missed, as happens at about 521 AMU. Furthermore, because the dominant peak is modeled without considering the side peaks, all the effects of peak superposition are assigned to the dominant peak, resulting in side peaks being underestimated.

displayed in Figures 9 – 11. In this region, the peaks are relatively well separated, and we see that at least for the three largest peaks, the height and width estimates can be accomplished via a simple algorithm, as shown in Figure 9 . In Figure 10, we see that Successive Dominant Peak Removal finds not only the three main peaks, but some smaller peaks that were previously hidden. It is a peak modeling question to determine if those smaller peaks should be considered part of the trailing skirt of a larger peak or an individual peak. In Figure 11, we see that the Successive Multi-Peak Removal method identifies even more potential peaks because the hidden peaks are considered part of a multi-peak cluster. This allows their contribution and that of the main peak to be calibrated together. Since the peaks near 467 and 468 are considered together, their relative contributions to the overall peak are better balanced.

Figures 12 - 14 show a mass range between 520 and 535 AMU. We see in Figure 12 that while a simple apex method can find a lot of apexes, the width calculations are completely confused by the overlap of the peaks. Figure 13 shows Successive Dominant Peak Removal applied to this range. It provides better estimates of the peak widths, but because it does not take side peaks into account, many potential peaks are ignored by this method. The results from Successive Multi-Peak Removal are shown in Figure 14. We see that more side peaks are found and their respective contributions to the abundance curve are more accurately accounted for. We have the best of both worlds, in which we see all the peaks from the original, simple method (and some hidden ones), but also have reasonable peak width estimates.

Finally, the mass range between 467 and 481 AMU is shown in Figures 15 - 17. In Figure 15 we see that the peak overlap causes really poor width estimates, most obviously in the peaks near 474 and 476 AMU. Those peaks also expose the issue with Successive Dominant Peak Removal (Figure 16), as the larger peaks at 474 and 476 are identified, but the side peaks are lost. In Figure 17, we see that Successive Multi-Peak Removal sees the side peaks, makes a reasonable estimate of their widths, and scales their height so as to take into account the effect of superposition of the peaks. It also finds previously hidden peaks that were missed by the other two methods.

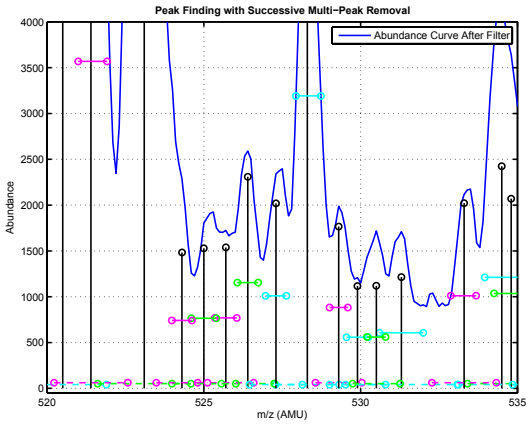

Fig. 14. Successive Multi-Peak Removal applied between 520 and 535 AMU. Note the more rational estimation of peak widths and heights. Since clusters of peaks are considered together, they are mutually scaled.

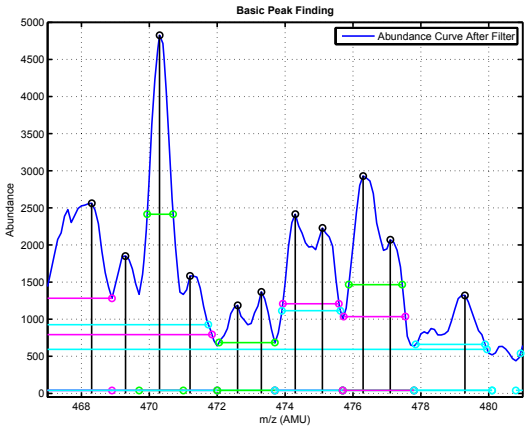

Fig. 15. Simple apex and width search applied between 467 and 481 AMU. Again, most apexes are detected, but the width calculations are unusable.

VII. Conclusions and Future Direction

As mentioned at the beginning, a simple method that identifies apexes in the abundance curve either by a peak search or by finding the positive to negative zero crossings of the abundance curve slope, then searches about those apex points for the full-width, half-max (FWHM) and the base width, works well when peaks are isolated and the signal to noise ratio (SNR) of the curve is high. The method of picking the largest peak in a region of interest and then iteratively modeling and removing the largest remaining peak, also works well when the largest peak is significantly larger than all the overlapping peaks so that is FWHM point can be determined before an interfering peak is encountered.

Instead, the issues of interfering peaks that show up before the width measurement point (e.g. FWHM) or peaks that are so close to a larger peak that they can only be detected by the bloated shape of the larger peak, are more difficult.

A reasonable question to ask is, "How common are these side peaks?" Statistics gathered on the data in these experiments are shown in Table I. We can see here that even in such a complex mix as shown in the previous section, the percentage of peaks with side peaks is less than 8%. Of

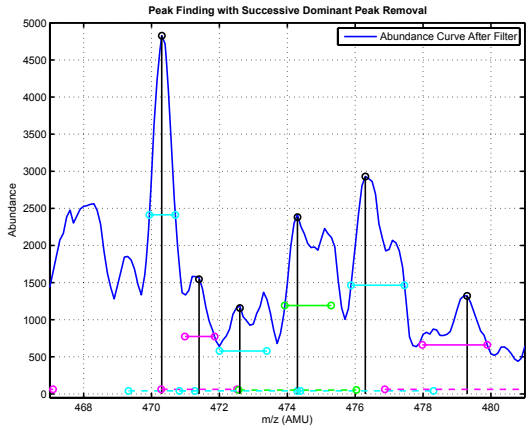

Fig. 16. Successive Dominant Peak Removal applied between 467 and 481 AMU. While the peak width calculations are better in most cases, the priority for the dominant peak means that the peak at 474 hides the peak at 475 and the peak at 476 hides the peak at 477. This also results in poor peak width estimates. The peak at 468 is ignored since it appears too asymmetric to be counted.

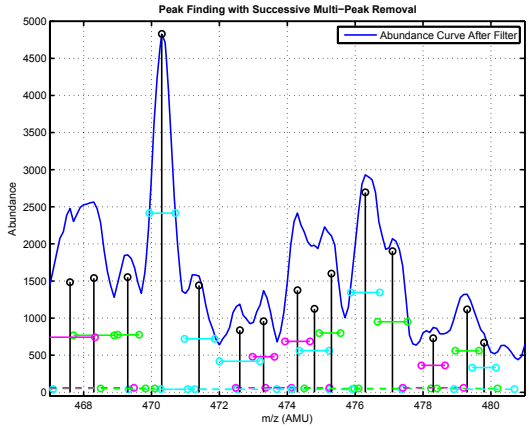

Fig. 17. Successive Multi-Peak Removal applied between 467 and 481 AMU. By considering multiple peaks together, we can now see closely spaced peaks (such as the ones at 475 and 477). We also can accept the peaks around 468 as being comprised of multiple overlapping peaks.

those, more than 1/4 have second side peaks. This means that one can identify most of the peaks in this test with simple methods, Successive Dominant Peak Removal, but for that small but significant fraction of the peaks, Successive Multi-Peak Removal is necessary to produce reasonable results.

It is reasonable to ask about whether an alternate peak model would produce better results, particularly in modeling the asymmetry often seen in width of peak skirts. To be useful, such a model would have to be easily parameterized from simple searches/calculations on the original measured abundance curve. This makes many of the functions used to abstractly model individual peaks less useful in a measurement environment.

In summary, the new algorithm is a combination of the following steps:

- The mass axis is segmented into different regions with interesting levels of abundance. These separate interesting regions are searched individually for peaks.
- Within each region, a search for local maxima is con-

Peak Type	Count	Percent (w.r.t center peak)	Percent (w.r.t left/right peaks)
Center	493	100	NA
Left	12	2.4341	NA
Left 2	1	0.2028	8.3316
Right	27	5.4767	NA
Right 2	5	1.0142	18.5185

TABLE I
STATISTICS OF THE CLUSTER LOCATIONS OF PEAKS FOUND IN EXPERIMENTS.

ducted and the locations and heights of these apexes are recorded. The largest of these peak candidates is identified as the starting point for width searches.

- If the search for the peak width crosses any side peaks, these are included into the model for peak removal.
- The data for the peak and any side peaks is fit to a model which is used to remove the abundance due to the peak(s). The rest of the peak width calculations are based upon the model(s) of the removed peak(s).
- The process is iterated in each individual region of interest until all peaks have been identified and removed, before moving to the next region of interest.

Through improved segmentation of the measurements and clever use of simple models, a dramatic improvement in peak finding is achieved, at least in regions where there is substantial overlap between peaks [12], [13]. In these situations, Successive Multi-Peak Removal provides dramatically improved results over the previous methods.

The "divide and conquer" nature of the algorithm dividing the mass axis into smaller, separate regions of interest and then applying an iterative methodology only on those smaller regions keeps the computational complexity of the new algorithm low. This means that improved peak detection and compound identification can be made directly on the instrument, rather than being relegated to specialized post-processing. Even if a more computationally intensive post processing methodology is applied, the segmentation and improved initial peak characterization give an advanced search or machine learning algorithm a set of smaller search spaces, and improved starting guesses in each of those spaces.

Again, while this is not a typical control system identification problem, system theoretic thinking allows us to see that we need to know when to switch models in real time. Experience with real-time systems pushes us in the direction of an algorithm that is computationally feasible with the Embedded System inside a commercial analytical instrument.

REFERENCES

[1] D. Y. Abramovitch, "Trying to keep it real: 25 years of trying to get the stuff I learned in grad school to work on mechatronic systems," in *Proceedings of the 2015 Multi-Conference on Systems and Control*, (Sydney, Australia), pp. 223–250, IEEE, IEEE, September 2015.

[2] J. T. Watson and O. D. Sparkman, *Introduction to Mass Spectrometry: Instrumentation, Applications, and Strategies for Data Interpretation*. Hoboken, NJ: Wiley, 4th ed., November 12 2007. ISBN-10: 0470516348; ISBN-13: 978-0470516348.

[3] E. de Hoffmann and V. Stroobant, *Mass Spectrometry: Principles and Applications*. Hoboken, NJ: Wiley-Interscience, 3rd ed., October 29 2007. ISBN-10: 0470033118; ISBN-13: 978-0470033111.

[4] R. A. Witte, *Spectrum and Network Measurements*. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1993.

[5] R. A. Witte, *Electronic Test Instruments: Theory and Applications*. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1993.

[6] P. H. Dawson, "Principles of operation," in *Quadrupole Mass Spectrometry and Its Applications* (P. H. Dawson, ed.), ch. 2, pp. 9–64, Woodbury, NY: American Institute of Physics Press, 1995.

[7] W. Yu, B. Wu, N. Lin, K. Stone, K. Williams, and H. Zhao, "Detecting and aligning peaks in mass spectrometry data with applications to MALDI," *Computational Biology and Chemistry*, vol. 30, pp. 27–38, February 2006.

[8] C. Yang, Z. He, and W. Yu, "Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis," *BMC BioInformatics*, vol. 10, no. 4, 2009.

[9] A. Antoniadis, J. Bigot, and S. Lambert-Lacroix, "Peaks detection and alignment for mass spectrometry data," *Journal de la Société Française de Statistique*, vol. 151, pp. 17–37, April 2010.

[10] Wikipedia, "Full width at half maximum," 2020. [Online; accessed March 11, 2020].

[11] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge: Cambridge University Press, third ed., 2007.

[12] D. Y. Abramovitch, "Method for finding species peaks in mass spectrometry," Patent Application Publication US 2019/0259592A1, Agilent Technologies, Santa Clara, CA USA, August 22 2019.

[13] D. Y. Abramovitch, "Improved peak detection for mass spectrometry via augmented dominant peak removal," in *Proceedings of the 67th Conference on Mass Spectrometry and Allied Topics*, (San Diego, CA), ASMS, ASMS, June 3–7 2018. Four page version available at: https://www.agilent.com/cs/library/posters/public/AGILENT_ASMS2018_TP292.pdf.

[14] D. Y. Abramovitch and C. R. Moon, "Automatic tuning of atomic force microscope," United States Patent 9,678,103, USPTO, Keysight Technologies Santa Rosa, CA USA, June 13 2017.

[15] D. A. Wright, "Methods of automated spectral peak detection and quantification without user input," Patent Application Publication PCT/US2010/036090, Thermo Fisher Scientific, May 25 2010.

[16] Agilent, *Ultivo Triple Quadrupole Mass Spectrometer*, August 31 2017.