4 PEAK FITTING

4.1 Introduction

When assessing large amounts of data it is beneficial to condense the most relevant information through analysis. It is therefore rare to present raw data. The analysis undertaken extracts the most useful information which can then later be presented in such a way as to understand the results without having to sift through the entire data set. Studies undertaken with field asymmetric ion mobility spectrometry (FAIMS) are often presented by reporting full dispersion field (DF) sweeps, which contradicts the above statement. However, when more than a single DF sweep, or comparison between several investigations is undertaken, it is much more common to describe the differences in terms of the variation in ion intensity, compensation voltage (CV) or full width at half maximum (FWHM) of ion responses with an appropriate error calculation. It is therefore the properties of the ion responses (which are nominally Gaussian peaks) found through the CV sweeps taken by FAIMS instrumentation that are ultimately used to understand the response.

The Owlstone FAIMS sensor is characterised by the small geometry used within the separation region. This small gap width enables high electric fields without breaching the electrical breakdown limit of the supporting carrier flow but it can also limit the resolution since ions experience a smaller residence time compared to larger designs. As a result a number of ion responses may overlap one another, referred to here as a mixed signal.

4.1.1 Mixed signals

Ions of interest must strike the detector, most commonly within a stand alone FAIMS system this is a Faraday cup. The signal obtained from the Faraday cup is the summation of the charge from ions of a single polarity striking the detector within a discrete time period. After this period the Faraday cup records ions of the opposite polarity and the response for the previous configuration is logged and the summation of charge begins again. Owing to separation of ions being dependent upon drift velocity and constant interaction with the neutral carrier flow, constituents of an ion population are spread across a range of CV. An example of this spread of a single ion species is shown in Figure 4.1, where the data is centred at a CV value of -1 V.

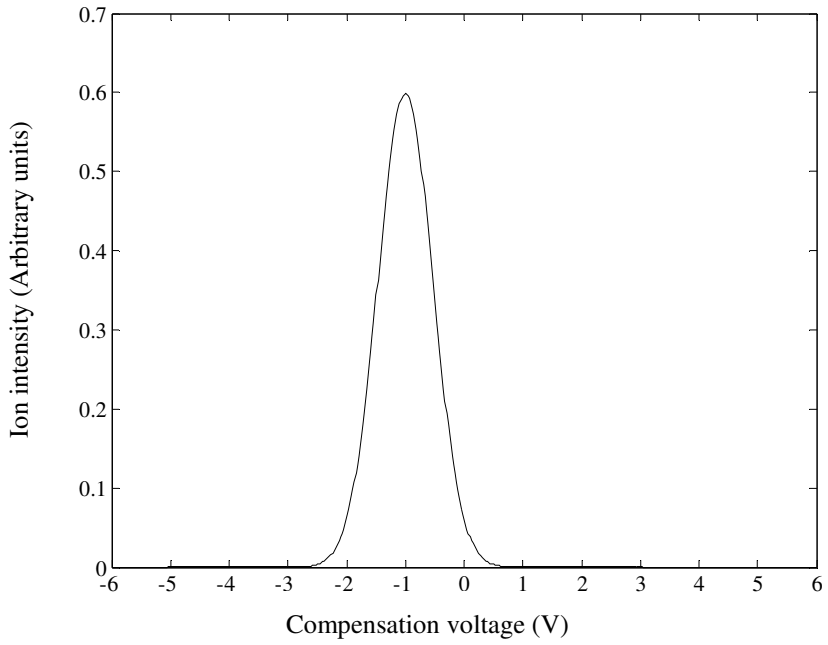

Figure 4.1 Above is a simulated plot representative of a CV sweep with only a single ion species suitable for transmission through the separation region of the FAIMS system. The area, position and FWHM of the peak are arbitrary.

When several ion species are present within the separation region there can be a mixing of the individual Gaussian peak responses, even if the ion species present have different mobility characteristics. Since the Faraday cup only returns a summation of the ions detected within a time period there is no differentiation, other than their polarity, between

ions from different ion species by the detector. This, in turn, means that it is possible for a result from a single ion species to be lost. In this scenario the ion species are said to be unresolved.

It can be extremely difficult to infer the properties of the peaks present from only considering the sum total. For instance, in Figure 4.2, it is possible that only two responses would be inferred from the sum total (solid line). Since this spectra was created artificially it is known that three independent responses make up the signal.

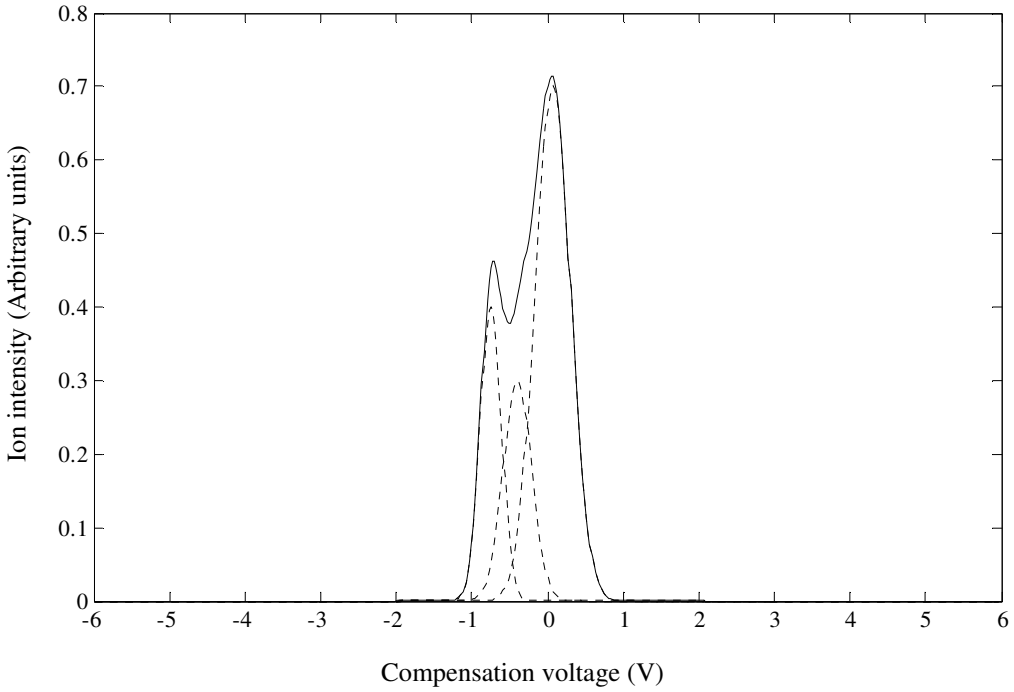

Figure 4.2 This is a simulated CV sweep typical of unresolved ion species peaks. The individual contribution from three ion species is depicted (dotted lines) with the summed output from the Faraday cup (solid line).

The mixing of the Gaussian peaks means that the true positions (used for identification of ion species) and area (used for determining abundance of ions) of individual responses will be masked by the summation of the signal.

4.1.2 Molecular-ion response

A further complication is that even a single Gaussian peak from a FAIMS detector may be due to multiple ion species. An example of how the response from several ion species may appear as a single Gaussian peak is shown in Figure 4.3.

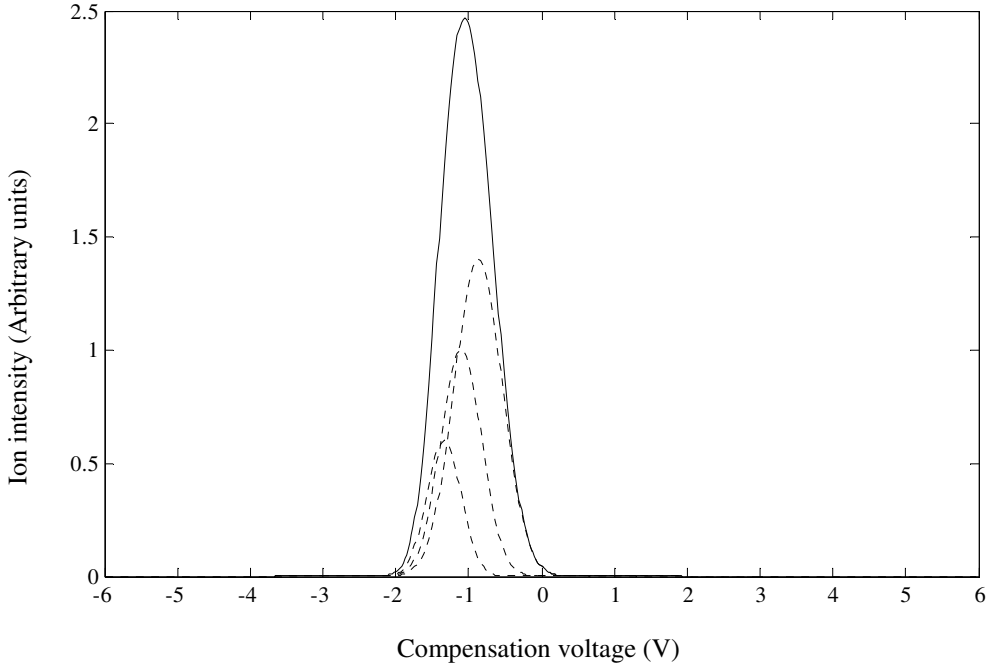

Figure 4.3 Above is a manufactured response representative of three ion species (dotted lines) and the sum of these three responses representative of the Faraday cup output that would be resultant (solid line).

Unless the constituents entering a FAIMS device are well understood it is impossible to infer the exact composition underneath a Gaussian peak from a single CV sweep. It is, however, possible to infer additional information following further CV sweeps at different dispersion fields. Alternatively, producing a library of CV spectra under controlled conditions enables identification of an unknown sample through comparison.

Proof that the situation depicted within Figure 4.3 can arise has been obtained by replacing the Faraday cup detector with a mass spectrometer. Figure 4.4 was originally two figures from Ells *et al*. [1] that show the observation of several ion species at a single point from a CV sweep. The set-up used was a cylindrical FAIMS placed between electro spray

ionisation (ESI) and a quadrupole mass spectrometer for the detection of chlorinated and brominated compounds formed due to the disinfection of drinking water. Several ion species exist under what appears to be a single peak response within the CV sweep.

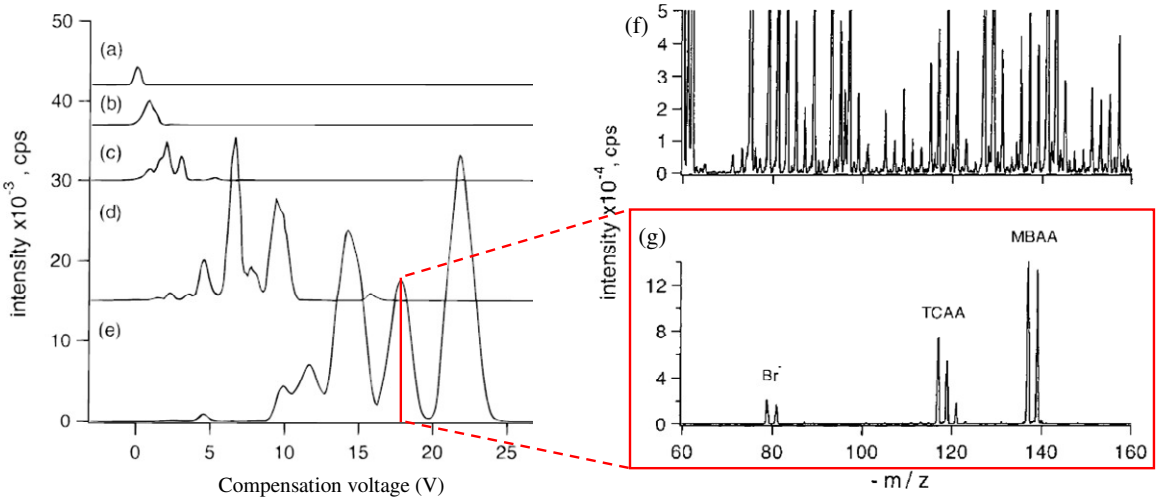

Figure 4.4 a) to e) are CV spectra of the five US environment protection agency regulated haloacetic acids and bromochloroacetic acid in a solution of 9:1 methanol/tap water (v/v) containing 0.2 mM ammonium acetate at dispersion voltage (DV) values of 0, -1300, -1700, -2500, -3300 V respectively. f) ESI-MS of the same solution. g) ESI-FAIMS-MS of the same solution at DV = -3300 V and CV = 18.0 V. originally from Ells *et al* [1].

Whilst it is difficult to definitively determine discrete ion species of an unknown sample through CV spectra alone, useful information can be obtained through accurate peak fitting.

For a given CV sweep, the greater the number of resolved Gaussian peaks that can be determined, the greater the points of reference there are for a comparison with spectra obtained under controlled conditions; facilitating identification of the species present.

4.1.3 Information from a single CV sweep

Important information can be gained from a single CV sweep of the ions that have been able to traverse the separation region of the FAIMS system.

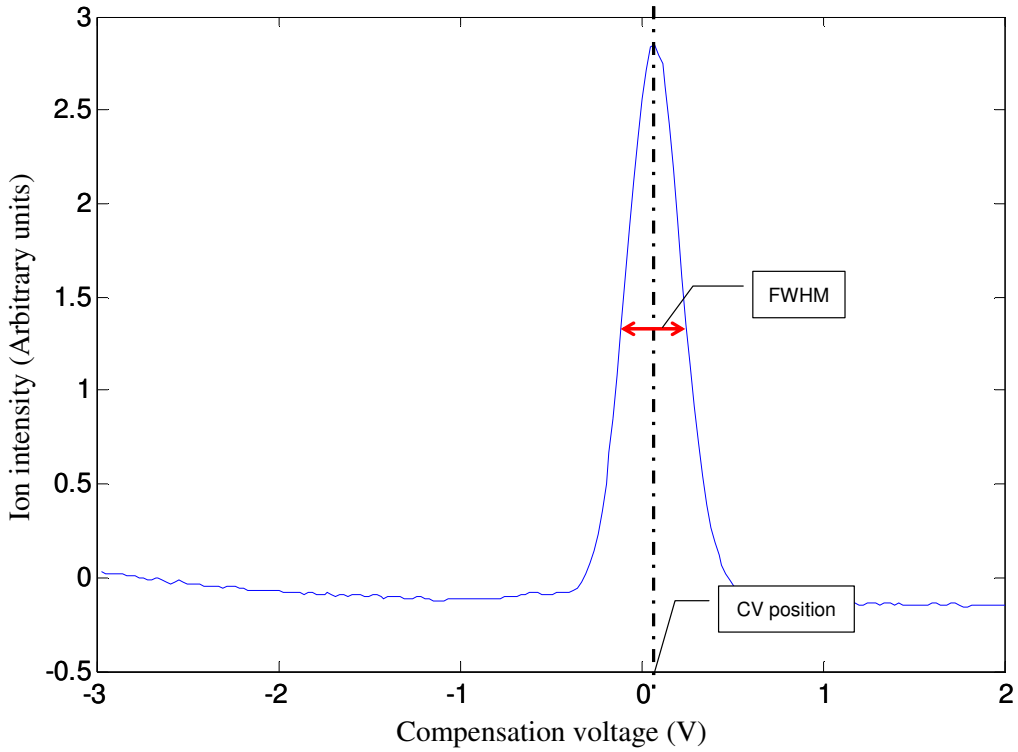

Figure 4.5 Above is a CV sweep of the reactive ion peak (RIP) of air in the positive mode using an Owlstone Tourist [2].

The area beneath the peak is representative of the total charge and hence proportional to the number of ions incident upon the Faraday cup, while the compensation voltage value of the peak's maximum is reliant on the drift velocity of the ion swarm. The magnitude of the FWHM of the Gaussian peak is dependent upon the interactions that the ions have undergone within the separation region (Section 2.9.1). These interactions, in turn, are affected by the residence time of the ions within the separation region and additional factors related to the identity of the ions and neutral gas molecules present. The baseline is also not nominally zero. This is a consequence of the time it takes for electrical charge on the FAIMS sensor to dissipate following a change from previously recording in the

opposite polarity. Baseline correction will be explored in Section 4.2.2. The units of the ion intensity and compensation voltage stated in Figure 4.5 are common to all the figures within this chapter.

4.1.4 Improving the limit of detection

As the dispersion field is increased within the separation region there will be increased separation of the different ion species, but also an unavoidable increase in the losses of ions of interest due to diffusion (Section 2.9). Successful peak fitting allows the response of ion species to be resolved from one another at electric field strengths lower than required for baseline resolution. It is then possible to obtain limits of detection that may not have been possible had no peak fitting occurred. This mitigates against the loss of sensitivity.

Peak fitting is valuable in FAIMS systems where the ion responses are more likely to be mixed because of a short residence time of ions within the separation region, as with the Owlstone FAIMS sensor and complex real world samples.

It is, however, important not to overestimate the capability of peak fitting as the results should be carefully balanced with known or anticipated chemical reality [3]. Any manipulation of data increases the likelihood of introducing additional errors which were not originally present. That said, a peak fitting algorithm may still be useful without resolving every ion species in an ion response. Applying a procedure to CV sweeps throughout an investigation can provide evidence of trends, which can then be exploited to gain a greater understanding of the analyte under study. Since peak properties (amplitude, area, CV position and FWHM) are recorded for every fit, plotting these parameters against

time will enable the evolution of a signal to be observed. Due to the number of individual peak fits the isolation of likely erroneous results can be filtered and intensity of a peak traced with respect to another property such as CV. This can be applied whenever FAIMS spectra are collected, such as from an EDF or GC-FAIMS system, as utilised through this body of work.

4.1.5 Assumptions and limits

To ensure that the methodology used to perform peak fitting is effective it must reflect the character of the ion responses to be fitted. This is accomplished by restricting the peak fitting by using assumptions that are related to behaviour common to all ion responses within CV sweeps. Predominantly, two assumptions were required.

The first assumption was the ion response from a single ion species has a Gaussian distribution. This is stated from theory [4, 5] and appears to be confirmed by experimentation as Gaussian peaks fit many of the spectra well.

Secondly, through the course of this thesis data collection from the FAIMS systems was through the use of a Faraday cup. The result of this is that the response from several ion signals may be summed together to produce the final output.

These two assertions mean that the raw data is made from the summation of individual Gaussian peaks. Any fit of ion response should be made up of individual Gaussian peaks which when summed together, do not exceed the signal returned from the Faraday cup. This sets a boundary condition which confines the possible fit. Comparison of the summed fitted Gaussian peaks to the raw data also provides a method to judge the accuracy of any

fit. Owing to the possible complexity underlying a single ion response fits are only attempted on broad features. Consequently, if a large number of Gaussian peaks sum to perfectly match the raw data the result could be disregarded since such detail can not be confirmed. Additional responses, which are apparent in the raw data, such as asymmetric ion signals, will however be considered as potential extra information suitable for peak fitting.

4.2 Procedures prior to peak fitting

Before peak fitting can begin some management and minor corrections to the data may be required. Common procedures are data organisation, baseline modification and defining what is noise. These important steps are considered within this section.

All computational work (in this chapter and throughout the thesis) was carried out with original programs constructed using Matlab (R2007b) available from The Mathworks Inc.

4.2.1 Data organisation

As may be anticipated, the very first stage of dealing with the raw data is to organise it so that it is in a format that can be processed. Depending on the system used the data may be easily accessible, or it may have to be exported and later placed in a peak fitting procedure. The raw data files may also include additional information such as user notes that must be discarded, so that they do not interfere with the peak fitting process.

The programs created for the work discussed within this chapter required data to be separated into positive and negative modes and stored in the consecutive order that they were recorded by the instrumentation. This meant that only positive or negative data was

dealt with at any one time, since some processes were dependent upon the polarity of the raw data. Organising the data into chronological order also made it easier to organise the data for summary.

4.2.2 Baseline modification

Once the data is correctly organised the individual CV sweeps are then loaded in to the program so that further manipulation can occur. An example CV sweep is shown in Figure 4.6.

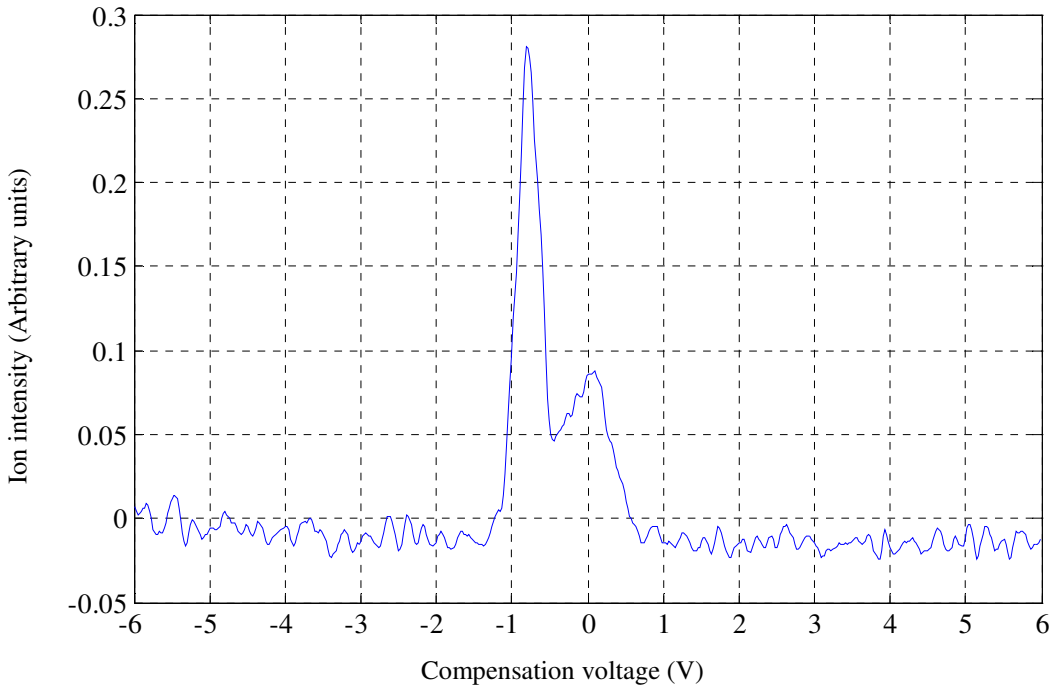

Figure 4.6 Example of a single CV sweep from an experimental run which is completely unmodified.

The grid lines have been included in Figure 4.6, so that it can be clearly seen that the baseline is not consistently equal to zero. The baseline drift is due to the decay of charge following a switch from detecting ions of opposite polarity by the Faraday cup. It is taken that this baseline drift is not representative of the true signal from the ions and so must be compensated for before anything can be inferred from the spectrum. The principal impact

due to the baseline drift would be the introduction of a difference between the observed and actual ion intensity.

To compensate for the baseline drift a curve is fitted, which describes the drift, and the raw data is converted accordingly. The baseline modification is performed automatically without user input. To accomplish this, the program must know what is exclusively representative of the baseline and of the signal. Such a distinction must be made as fitting a curve to the full data would lead to incorrect baseline modification, as the fit would be skewed by the large signal contribution. Instead a ‘window’ is defined which describes the region of the CV spectra that contains the ion responses. This window is arbitrary and is preset by the operator in the initial conditions of the program so that it is applied to all CV sweeps being considered. The limits of the window are decided upon following some initial investigation using test samples and the maximum dispersion fields to be utilised in the full study. For the current example, the data points that are defined as within this window and the data points that are considered only due to the noise are described in Figure 4.7.

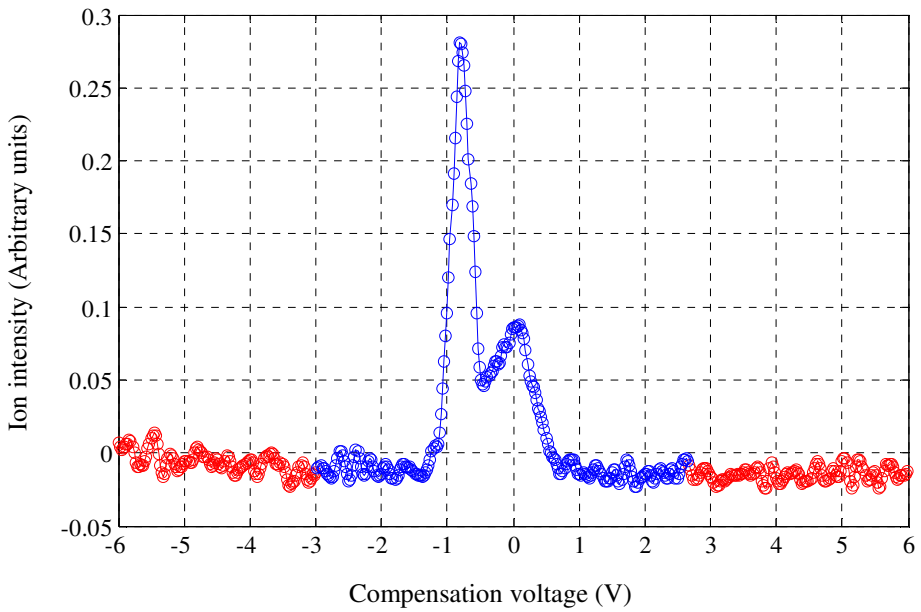

Figure 4.7 The blue circles and solid line denote the window of data points designated as due to the signal response. The red circles and solid line denote data points solely due to noise. It is the red data points that are used to create the curve to complete the baseline modification.

From Figure 4.7, it is clear that the window encompasses a good deal of the signal only dependent on noise, but this is by design. From initial investigations the likely positions of the peaks are known and it is prudent to include additional data points in case an event occurs that is not anticipated.

The data points which are not within the window region are now taken as purely resulting from noise and the drifting baseline. A polynomial curve is fitted to these points only. It is desired that the non-window region should include as much output dependent upon noise as possible. This is to minimise any localised structure affecting the baseline fit but care has to be taken that no signal response is included. The order of the polynomial should be small enough that the baseline fitted is not dependent on small structure within the noise but also large enough so that legitimate drift is accounted for. A polynomial order of four is taken as the form of the curve fit within this chapter.

The non-window data points are used to fit the polynomial. The values of the resultant curve are then used to modify the raw data. The curve fitted as the baseline to the example data and the resultant baseline modified data is displayed in Figure 4.8.

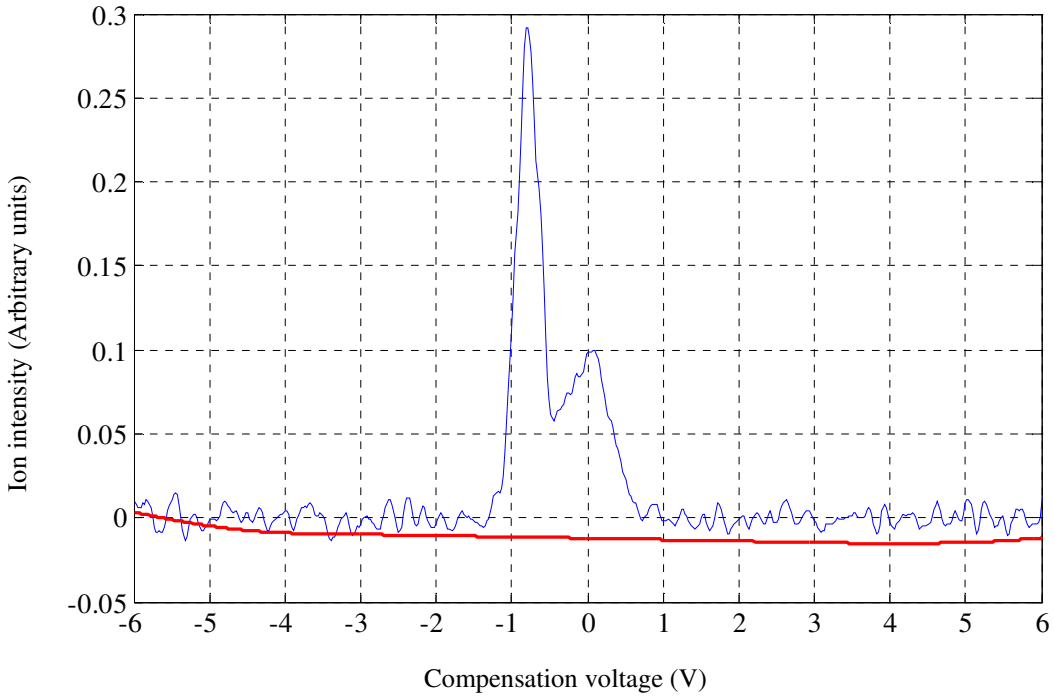

Figure 4.8 The fitted curve (solid red) as found from the non-window region depicted in Figure 4.7 and raw data which has been compensated for baseline drift (solid blue).

As can be seen in Figure 4.8 the baseline correction appears to have been successful because the noise now fluctuates around zero ion intensity. There also appears to be no contribution to the noise modification from the ion signal which is a validation of the window limits that were used in the application of the program.

Finally, the fitting of a polynomial of order four appears to have been a good choice as the gradient of the baseline is not restricted to either being positive or negative and is not affected by the minutiae of the noise.

4.2.3 Defining a noise threshold

The next stage in fitting peaks requires identification of what is to be treated as noise. In this work a high pass filter was used to re-define the ion responses signal and noise. To determine a threshold the standard deviation of the data points in the non-window region, as defined in Section 4.2.2, was found. The standard deviation value was then multiplied by a user defined constant. This constant could be a set value, so that the user input is not required, but it was included as a variable within the program for simple modification and greater user control. Using the same example data as previously, Figure 4.9 shows the baseline modified raw data with the noise threshold limit imposed over the top.

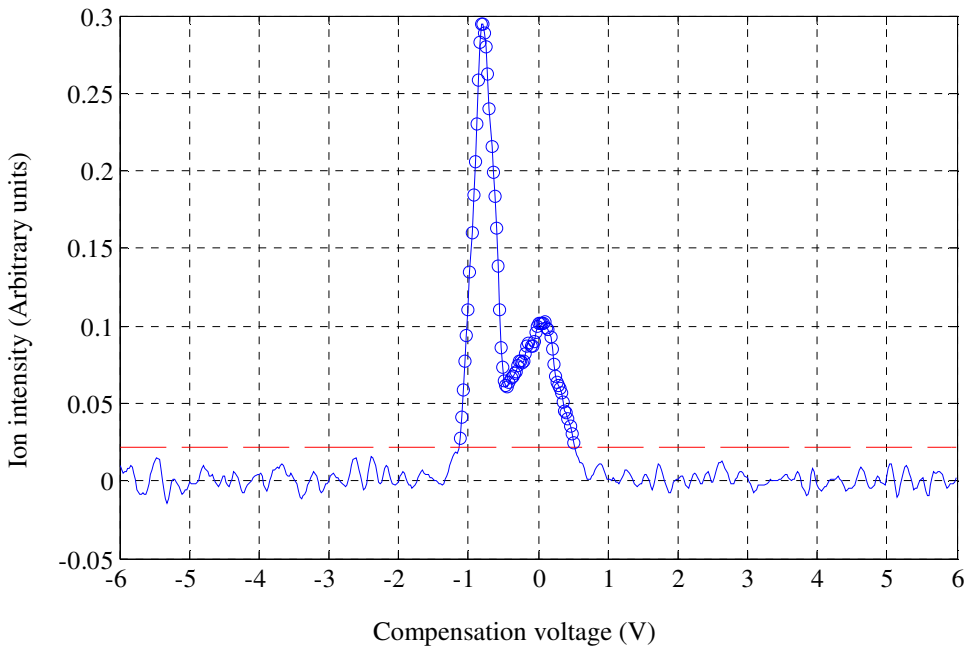

Figure 4.9 The noise threshold (dashed red) which has been calculated from the standard deviation of the non-window region and multiplied by a user defined constant (this case four). The baseline corrected raw data (solid blue) is provided along with the data points greater than the noise threshold (blue circles).

The noise threshold in Figure 4.9 appears appropriate and clearly separates what will, from this point onwards, be treated as the signal from the Faraday cup and what will be treated as noise. The noise threshold is calculated anew for every CV sweep of data passed through the program. This means that sweeps with a decreasing signal-to-noise ratio (S/N), such as the case of increasing DF field strength, will be catered for.

4.2.4 Initial peak isolation

Following the previous steps the CV sweeps have been baseline corrected and the data points within each CV sweep have been characterised either as a result of noise or ion response. The final peak fitting operation in all the scenarios considered within this chapter will be completed through unconstrained nonlinear optimization (described later in Section 4.6.1). Before this step an initial estimation of the number of peaks and their properties must be found to enable the final optimisation. Three methods of finding those estimations have been investigated; isolation of maxima, differential, and Successive Gaussian.

4.3 Isolation of maximums

The isolation of maximums method looks for maxima in the ion signal. After such a point has been located the procedure utilises simple expressions and criteria to enable fast computation and transparency throughout operation to fit Gaussian peaks to the data set.

4.3.1 Initial estimate of peak intensity and position

The isolation of the peaks within a CV sweep was tackled one peak at a time. To identify the initial peak, the maximum value of the data greater than the noise threshold was found. From the assumption that the ion response is described by a Gaussian peak, the maximum value indicates the central position of a peak. The ion intensity and position (described by the term number of the CV sweep) of the maximum value was recorded.

It should be noted that taking the maximum point as the peak intensity and position can be affected by non-smooth ion responses. For instance if the ion response is quite ‘spiky’ (common at low S/N) the maximum point may not be representative of a true peak. The

ion intensity may still be useful but the position of the maximum may be off centre. In such cases it is beneficial to smooth the data so that small variations have less of an effect. Smoothing processes are explored in greater detail within Section 4.9.1.

4.3.2 FWHM of the initial peak

Whilst the values of ion intensity and position, which are believed to be representative of the greatest ion response, have been discovered the spread of the ion response is still to be estimated.

The spread of an ion response is described by the FWHM of the peak. To discover an estimation of the FWHM the position of the first data point (which has an intensity less than half of the maximum) leading out from the position of the maximum is recorded. The difference between the position of this point and the position of the maximum is doubled and used as an approximation for the FWHM of the initial peak.

While this method is only intended as an approximation it will provide an overestimate if the FWHM of the peak investigated is part of a mixed ion response. To reduce the frequency that this occurred the two sides of the peak were compared to assess which provided the smallest FWHM. This was completed through comparing the data points a fixed distance from the maximum. The side of the peak that had a lower intensity was the side which the FWHM was calculated from, since this would be the side least likely to have mixed with another Gaussian peak. Figure 4.10 shows the information important for discovering an estimation of the first peak from the example case that was previously baseline corrected in Section 4.2.2.

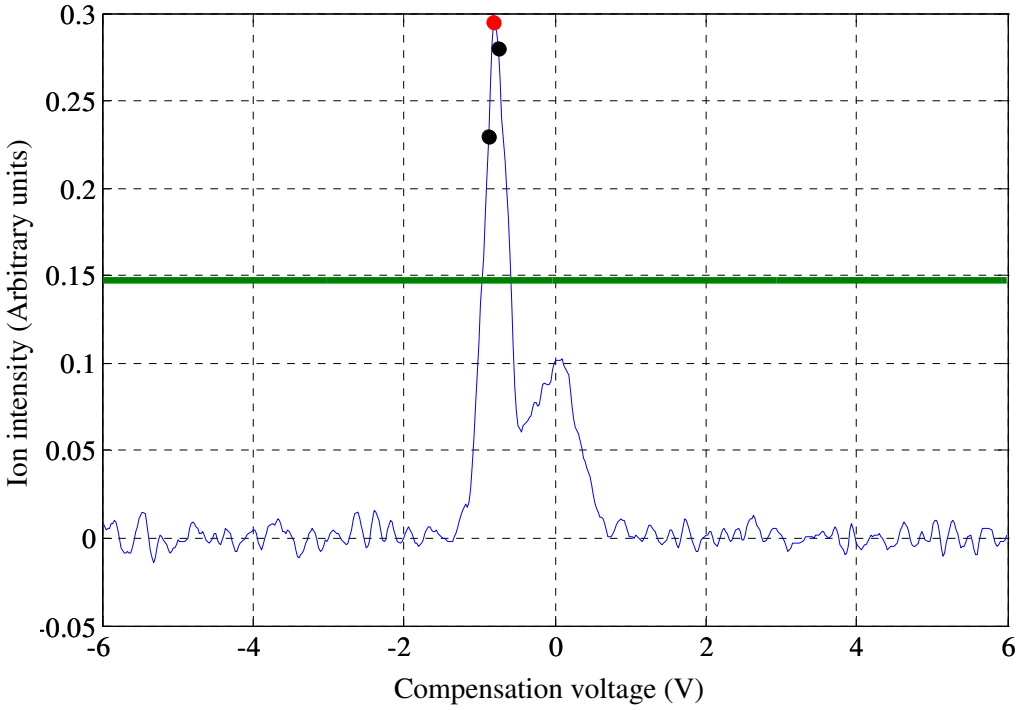

Figure 4.10 Baseline corrected example data (solid blue) is presented along with the maximum point (red circle), two data points equally distant in the x-axis from the maximum (black circles) and the ion intensity equal to half the maximum value that is used to discover the eventual estimate for the FWHM (solid green).

From Figure 4.10 it can be seen that it is the left hand side of the peak which decreases in intensity the quickest. This means that it will be the left hand side of the peak which is tracked to discover when the data points drop below the half maximum of the guess peak. The position of the first data point below the half maximum and the position of the maximum are then used to estimate the FWHM of the initial peak as described previously. An increase of ~ 20 % was also made to overestimate the ion response to prevent triggering too many peak fits.

4.3.3 Constructing and storing the initial peak

The three properties required to generate a Gaussian peak (amplitude, position and FWHM) have all been approximated for the most prominent initial ion response through

Sections 4.3.1 and 4.3.2. To construct the full Gaussian peak based upon these properties Equation 4.1 is implemented.

$$y_i = peak \times \exp\left(\frac{-(x_i - position)^2}{\left(2 \cdot \left(\frac{FWHM}{2\sqrt{2\ell n2}}\right)^2\right)}\right) \qquad 4.1$$

Where the subscript i is used to denote separate terms of a series. y_i and x_i are the ion intensity and term number at the point designated by the subscript i, respectively.

An approximate Gaussian peak has now been created that is representative of the initial peak of the test data. Figure 4.11 shows the resultant Gaussian peak overlain on the test data.

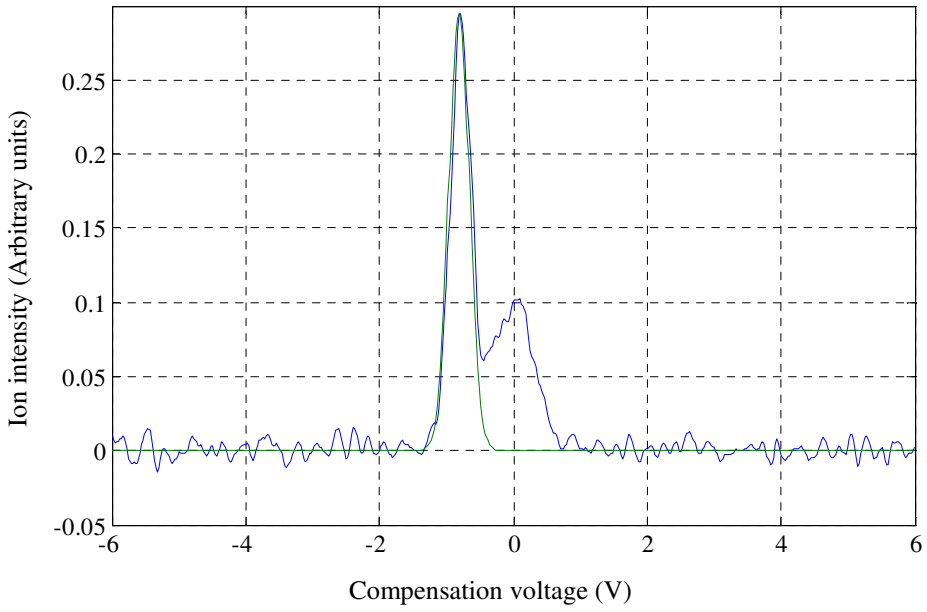

Figure 4.11 The baseline corrected example data (blue) and the initial Gaussian peak (green) as estimated from taking estimations from the example data.

Figure 4.11 shows that the guess peak recovered from the example data is a good approximation of the major response within the example data. However, there is clearly remaining ion response that requires similar treatment.

4.3.4 Further Gaussian peak fits

Now that a single Gaussian peak has been fitted to a CV sweep the remaining ion response requires fitting. The ion response which is now represented by the first Gaussian peak is subtracted from the baseline corrected CV sweep and the same procedure that was used before is repeated. This continues until there is no ion response greater than the noise threshold. A sequence is shown in Figure 4.12 of the successive fitting of Gaussian peaks to the example data.

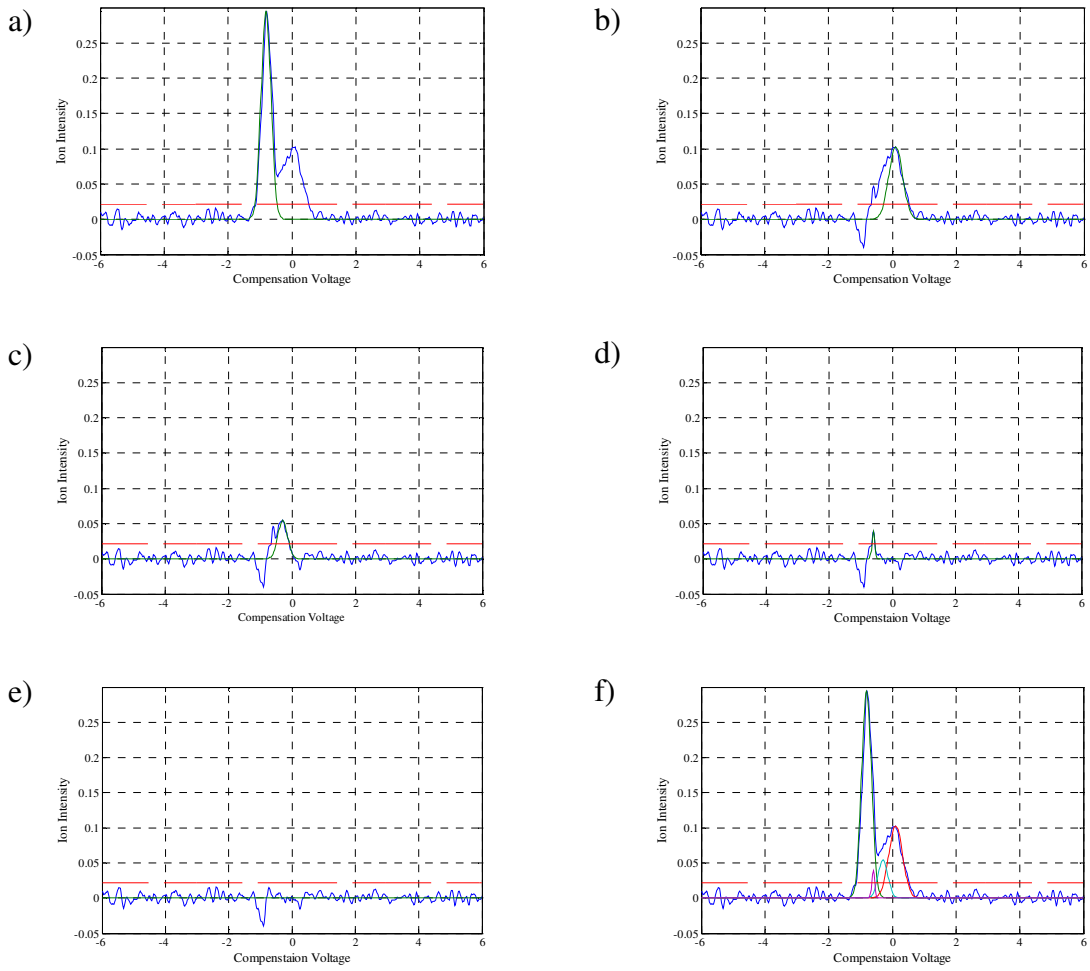

Figure 4.12 a) The initial peak is fitted, b) the second peak is fitted on the baseline corrected data minus the initial peak, c) to e) this process is continued until there is no ion response greater than the threshold level. f) The baseline corrected data is plotted alongside the four Gaussian peaks fitted to the baseline corrected data.

These initial fits have provided several peaks. While broadly these estimated peaks are useful and could be plotted over time, they do not represent the best fit. Due to the method implemented the peaks are fitted to ion responses of greatest amplitude before any other.

Following a peak fit the result is subtracted from the original signal. It is therefore possible that part of another unfitted ion response has been lost if there was any mixed ion response between species.

Summing the estimated Gaussian peaks observed here results in a total ion response greater than the response obtained from the detector, this conflicts the consequence of the second assumption made for accurately fitting peaks (Section 4.1.5). More refinement is therefore required before the procedure can be useful. This will be described in Section 4.6.1 following a discussion on the two remaining methods of peak estimation explored through this work.

4.4 Peak estimation through differentiation

Discovery of possible peaks through differentiation is potentially more successful than observing maximum points because points of inflection, which suggest substantial mixed responses, can still be isolated even though a peak maximum is not evident. There is, however, increased computation associated with finding data representative of peaks.

4.4.1 Methodology of peak isolation

The maxima and minima within a series can be found from calculating the differential throughout the series. Differentiation is a well recognised and implemented mathematical tool for finding peaks and troughs of a data set. The locations of where the differential of the series equals zero is a maximum, minimum or point of inflection of the original series. Since the objective is to locate and isolate peaks within a data set this mathematical procedure is directly relevant.

To take this further, the second order differential of the original test series is found. The second order differential not only describes the location of peaks but also their identity. From the second order differential maxima, minima and points of inflection of the original series are described by minima, maxima and zero points respectively.

The differential of a series was calculated by dividing the difference between two successive points of the ion intensity and two successive points of their location. This was carried out upon data after baseline correction through applying Equation 4.2,

$$y_i = \frac{(b_i - b_{i-1})}{(i-(i-1))}. \qquad 4.2$$

Where i is the current term number, y_i is the differential value at term i, b is the ion intensity value (subscript denotes term). This procedure could be repeated on successive series to discover higher order differentials of the original data set.

Since the minima of the second differential describe the maxima of the original data set it was this information that was used to isolate the peak positions. The intensities of the original data set at these isolated positions were passed on as the peak amplitudes of the Gaussian peaks to be fitted. The FWHM value was calculated as dependent on the peak position (an arbitrary value was reduced with respect as to how far the peak position was from zero displacement). This method of obtaining the FWHM was used instead of the process described in Section 4.3.2 because it was more successful with highly mixed ion responses. The disadvantage was that the estimated FWHM would be less representative of the data.

The differentiation method proved effective for detecting peaks in smooth data sets and mixed ion responses. The method, however, is very sensitive to ‘spikey’ (normally low

S/N) data sets so it has been necessary to include smoothing procedures if a data set was found to trigger too many peaks to be fitted. Methods for smoothing a data set are described in more detail within Section 4.9.1.

4.4.2 Effect of smoothing data set

A demonstration of the effect of smoothing, when initial peak estimation is completed by the differential method, is given in Figure 4.13. Parts (a) and (b) are completed without smoothing of any of the data set while parts (c) and (d) are completed with an automatic smoothing step dependent upon on how many peaks would be triggered if no smoothing occurred. The second differentials presented within parts (a) and (c) are each multiplied by twenty so that they can be easily observed alongside the raw data. The final peak fits, (b) and (d), are a result of the initial peak estimations followed by the final fit described in Section 4.6

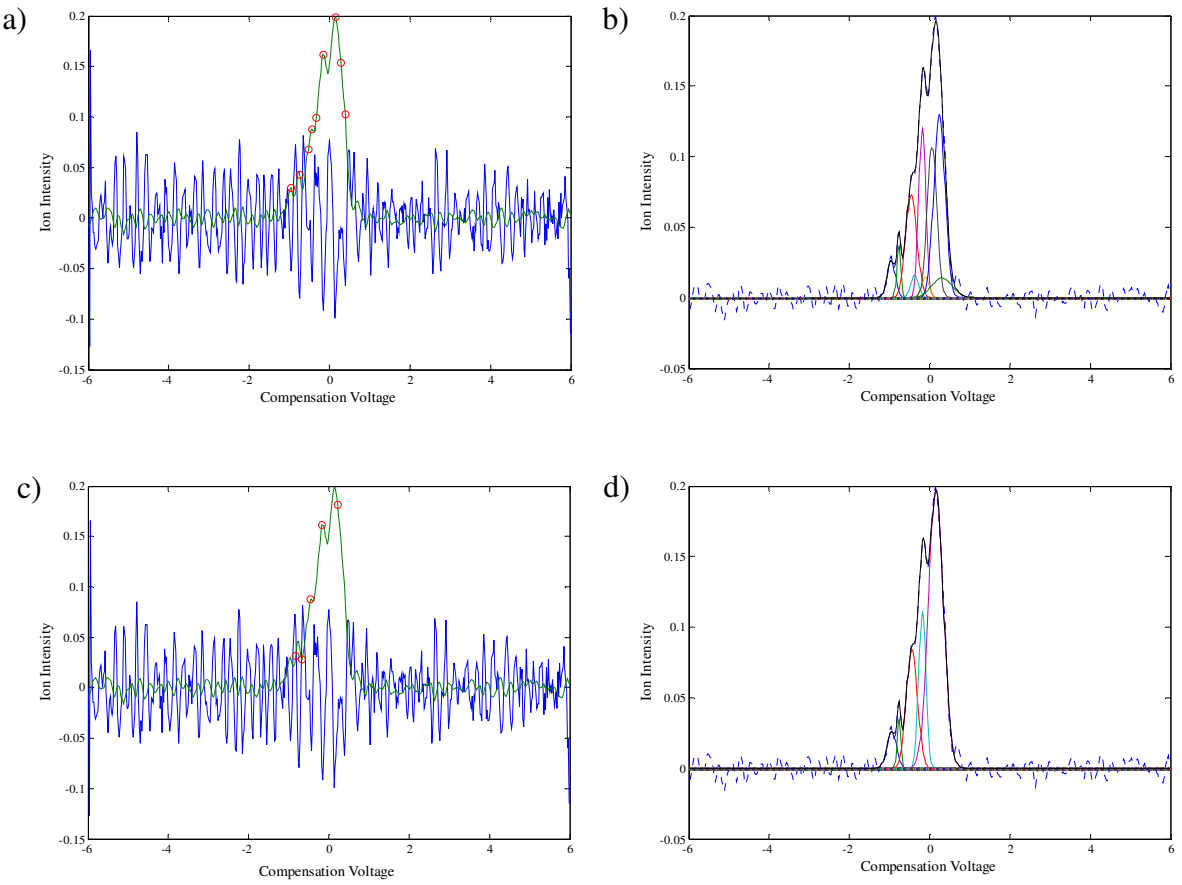

Figure 4.13 a) second differential (solid blue) and initial estimation of peaks (red circles) made from unsmoothed data set, b) resultant peak fit using initial peak fits shown in *a*, c) second differential (solid blue) and initial estimation of peaks (red circles) made from smoothed data set, d) resultant peak fit using initial peak fits shown in *c*.

The smoothing that was employed led to a more representative result of ion responses similar to what is expected from a FAIMS sensor. While the fitted peaks produced within Figure 4.13 (b) provided more peaks there is no indication that they couldn't have been triggered by small variations in the data set. The smoothed response prevents fits being made to small fluctuations and encourages only the major features to be fitted.

Again, as before, it is not suggested that the fitted peak responses on the smoothed data describe all the individual ion responses present. Instead they describe the major ion responses but in greater detail than could be realised previously. This assists in the possible tracking of a particular ion response over subsequent CV sweeps.

4.5 Peak isolation through Successive Gaussians

The final method investigated of isolating peaks within a data set was that of Successive Gaussians. This method involves comparing successive Gaussian peaks with the raw data. Successive Gaussian's were employed as the procedure is less sensitive to the perturbing influences of 'spikey' data and highly mixed responses encountered with previous techniques.

4.5.1 Methodology of Successive Gaussians

It is assumed that every data point from the Faraday cup is the result of a Gaussian distribution. A full CV sweep can then be broken down into its independent magnitudes and a Gaussian peak created, equal in area to each data point. The number of these Gaussian peaks was therefore equal to the number of elements recorded within the CV sweep. Also, the total sum of the individual Gaussian peaks would be equal to the total area of the full CV sweep.

The purpose of creating these individual Gaussian peaks was so they could be summed, one by one, and compared to the full data set after each addition. A calculation was carried out as to how well the summed sample matched the same range of the full data set. This calculation was in effect a confidence test to understand how well the summation of individual Gaussian peaks matched a range of the original data set. A confidence test of zero represents a perfect match between the summed Gaussian peaks and the sample data over the assessed range.

Outlined here is the single process that is carried out on all ranges within the data set. To ensure every range is sampled the confidence test is applied to a sample found from a rigid

procedure. Initially only the first point of the original data set is considered. This single point is the entire sample which the confidence test is carried out upon. The second range is constructed of the sum of the first and second data points of the original data set. The range is increased in this way until the entire original data set has been sampled. Following this the process is repeated but this time the first data point is omitted. Again this process is cycled until the final data point of the original data point is the entirety of the range under study. Each time the confidence test is carried out the result is recorded and placed within a matrix where its position is dependent upon the position of the first and last points of the range under study within the original data set.

The outcome is an array where the values closest to zero represent the co-ordinates of curves which most accurately fit a Gaussian profile. From here on the method described within this section will be referred to as the Successive Gaussian method.

The Successive Gaussian method is extremely slow compared to the other methods presented within this chapter (takes ~ 1800 times longer than the two previous methods) owing to the exhaustive cycling of ranges. However, it offers a technique that is more robust to variations (such as noise) within a data set. Previous techniques have also resulted in a bias when fitting peaks due to the particular way in which they operate. Through the testing of every possible range, the procedure provides an opportunity to resolve mixed signals with no overt prejudice other than how well it describes the data. With other methods investigated the ability to determine more complicated signals often results in an increase in the number of procedures required. Each additional procedure requires more information which may, instead of making the procedure more general, result in it being suitable for a decreasing range of spectra.

4.5.2 Results from Successive Gaussian method

As mentioned, the result of the Successive Gaussian method is an array of confidence values. This array is usefully presented as a contour plot, an example of which is shown in Figure 4.14.

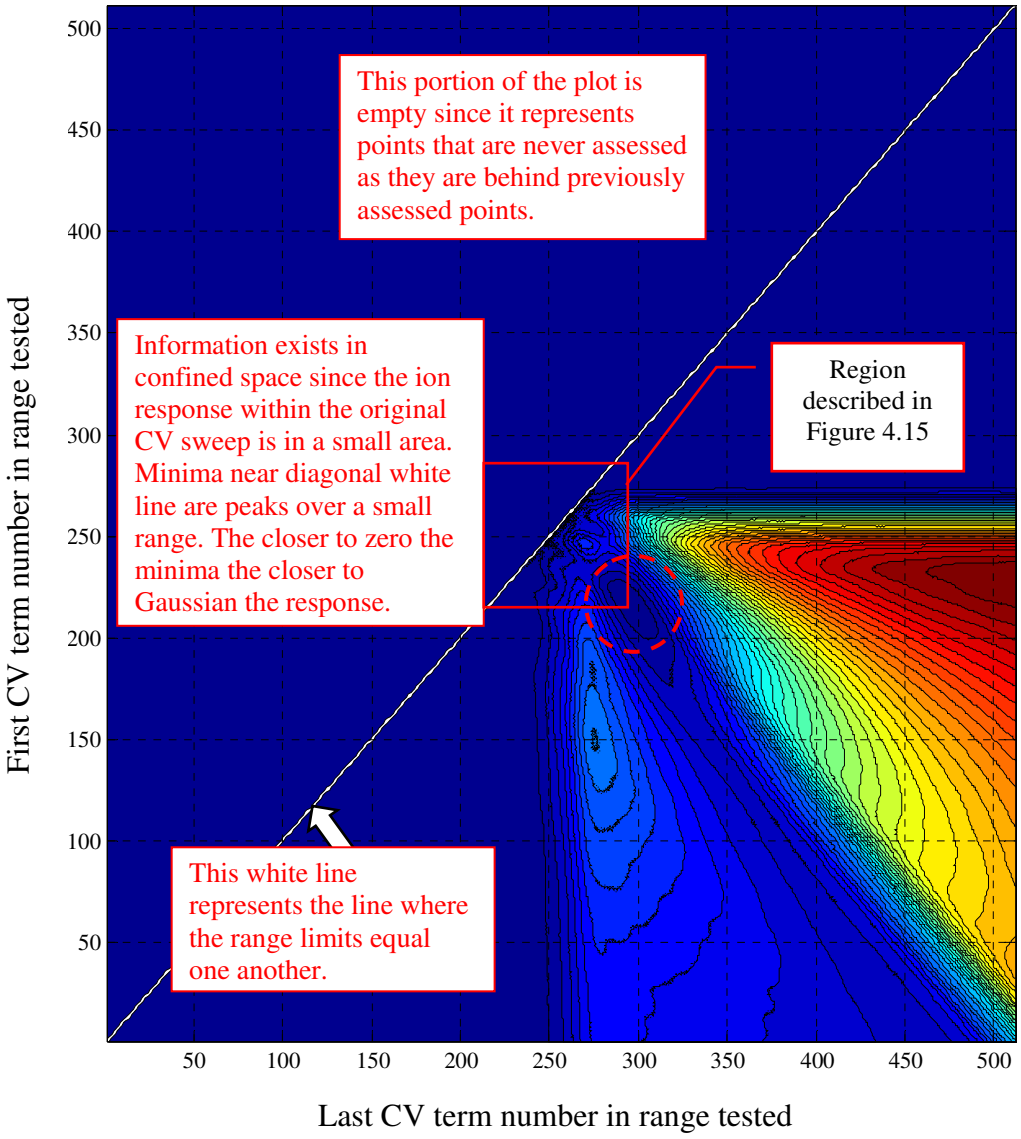

Figure 4.14 A contour plot of the confidence array produced through the Successive Gaussian method.

The information within the confidence array which describes the peaks isolated from the original data is expressed by minima within the confidence array. The contour plot allows

for an easy inspection of the confidence array but more rigorous means are normally required to truly isolate the desired information. Figure 4.15 shows the same contour plot but with the relevant minima labelled.

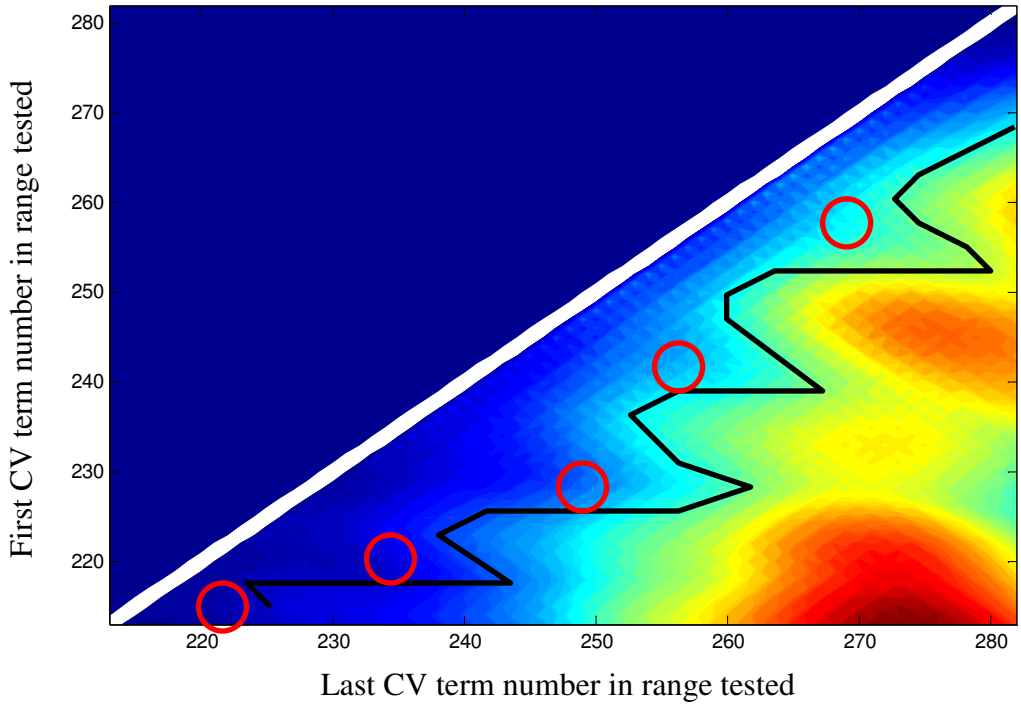

Figure 4.15 This plot is the highlighted region of Figure 4.14 with the confidence values representative of peaks depicted with red circles. The solid black line has been added and traces regions of rapidly changing gradient.

The co-ordinates of the minima between regions of changing gradient describe the range that a section of the original CV spectra closely matches the profile of a Gaussian peak. For example, one such point displayed within Figure 4.15 was found at CV term number 255 (along the *x* axis) and CV term number 242 (along the *y* axis). It is between these two term numbers that a peak has been isolated. Figure 4.16 takes the co-ordinates described by the minima discovered and plots the ranges on top of the original data.

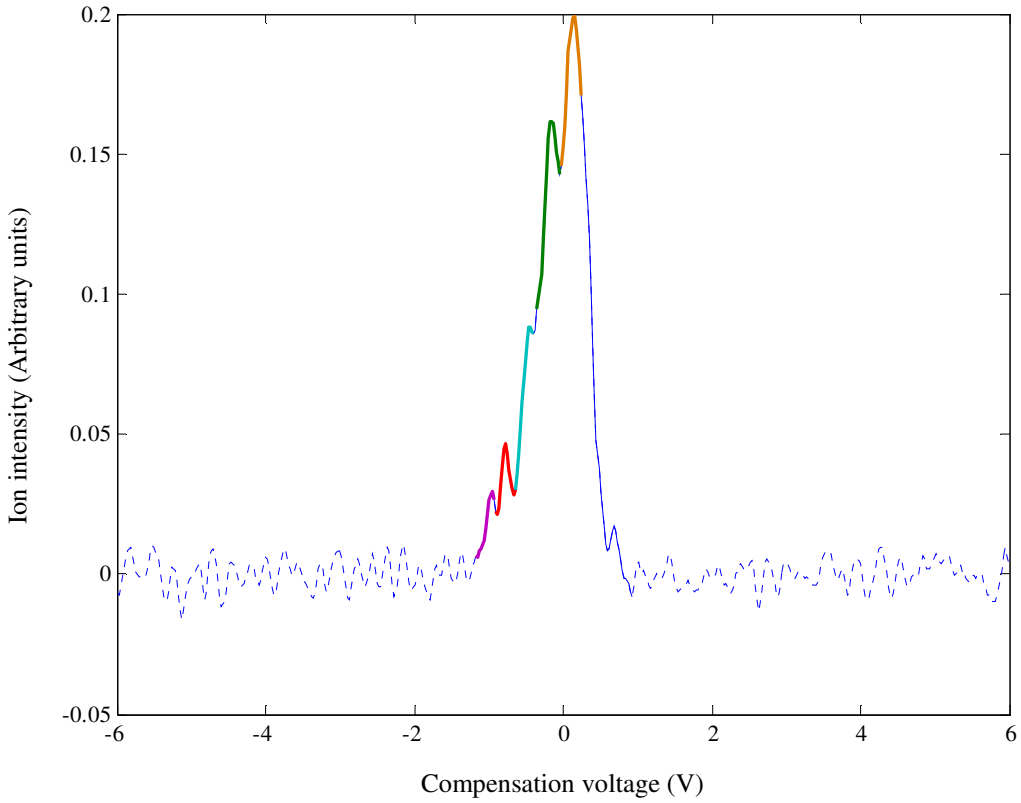

Figure 4.16 The original CV spectra (blue dotted) shown alongside the parts of the spectra which correspond to the minima of the confidence array (purple, red, light blue, green and orange solid lines). The part of the spectra which is described by the dotted red circle from Figure 4.15 is also shown (solid blue line).

It can be seen that the method has successfully isolated the same individual peaks that the previous methods would have been able to accomplish but with only a single automated process that cycled through the whole data set. Information could be taken from the ranges discovered and these would be passed on as the initial peak fits for the final non-linear least squares as utilised by the other methods described in this chapter.

4.5.3 Prioritisation within Successive Gaussian method

Although the new method relies on a single general procedure there is the potential that the method could be influenced by prioritisation towards some features more than others. The two points where this could occur are through the confidence test and isolation of the minimums from the parameter space.

The confidence test used takes into account the fit of the test data set with the raw data and the summed data from all the individual Gaussian peaks. A particular weighting towards one data set over another has not yet been discovered but it has been noted that the technique is not as sensitive at detecting mixed Gaussian peaks as the other methods presented in this chapter. The confidence test can be modified depending on the particulars of the data but the one used in this investigation is,

$$confidence = abs(testSum - GaussSum) \times abs(ySum - GaussSum). \qquad 4.3$$

Where *confidence* is the confidence value returned after carrying out the test, *testSum* is the sum of the individual Gaussian peaks being tested, *GaussSum* is the Gaussian peak calculated from the total area of the summed individual Gaussian peaks and *ySum* is the sum of the raw data over the range being challenged. *abs* states that the absolute value is taken.

It is important to include all three quantities (*testSum*, *GaussSum*, *ySum*) since they all represent slightly different quantities. Taking the absolute of the two terms on the right hand side of Equation 4.3 and multiplying them by one another amplifies any differences and ensures the lower the returned confidence value the more likely the range under study describes a single ion species response. Figure 4.17 shows the three quantities present within the confidence test.

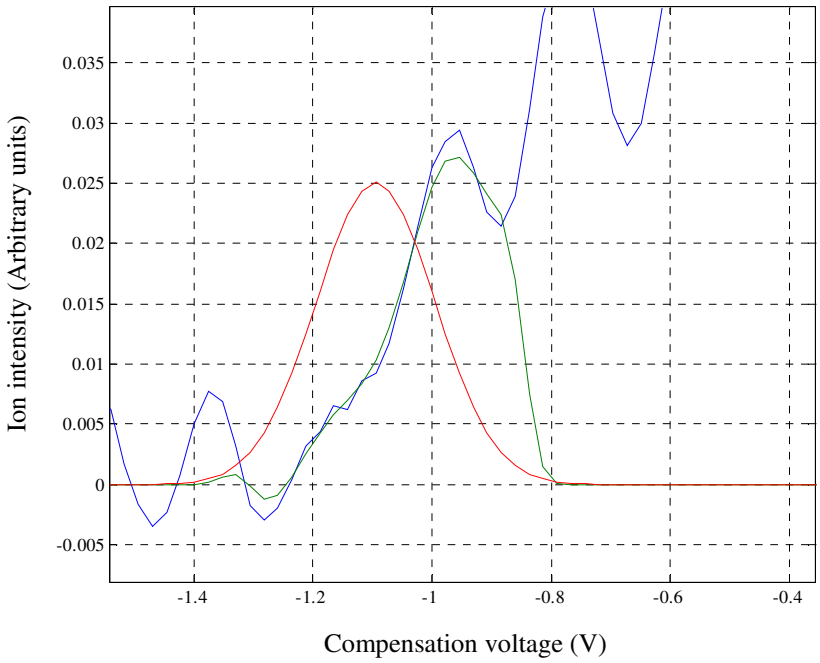

Figure 4.17 An example of the Successive Gaussian method in operation between CV term numbers 200 and 220 (corresponding to CV range of -1.3289 to -0.8602 V). Baseline corrected original data (blue), sum of individual Gaussian peaks (green) and Gaussian peak fitted by taking properties from original data between stated range (red) are all displayed.

The isolation of minimum confidence results and minima is the simplest way of finding the best fit of the curves but there may be additional information present, depending on the shape and depth of depressions. Certain trough shapes in the confidence array may in fact allude to mixing Gaussian peaks but this has not been explored in depth.

The method of Successive Gaussians has been presented here as an answer to some of the issues found through the methods of isolation of maxima and differentiation. Practically, however, the process is too slow to be used through out a full investigation. As a result only the two previous methods of peak estimation have been used elsewhere in the analysis of data within this thesis.

4.6 Final peak fit

The final peak fitting provides discrete Gaussian profiles which sum to equal the raw data from a CV sweep. First the likely numbers of peaks, their intensity, position and FWHM were required. This was accomplished through the initial peak fits described within Sections 4.3 - 4.5. This information was then passed on to a procedure which modified the properties of the approximated peaks so that the properties of the broad ion responses underlying the mixed ion response were discovered.

4.6.1 Final peak fit procedure

The technique implemented to accomplish this adjustment is the FMINSEARCH procedure which is part of the MATLAB software package. FMINSEARCH finds the minimum of a scalar function of several variables, starting at an initial estimate. This is generally referred to as 'unconstrained non-linear optimisation'. The minimum scalar function is set to minimise the difference between the sum of fitted peaks and the baseline corrected ion response from the Faraday cup. The initial estimates used are the values required to construct the initial peaks presented earlier.

The result of using the FMINSEARCH procedure is shown within Figure 4.18. The example data and estimated peak fits are the same as those that were previously used within Section 4.3.

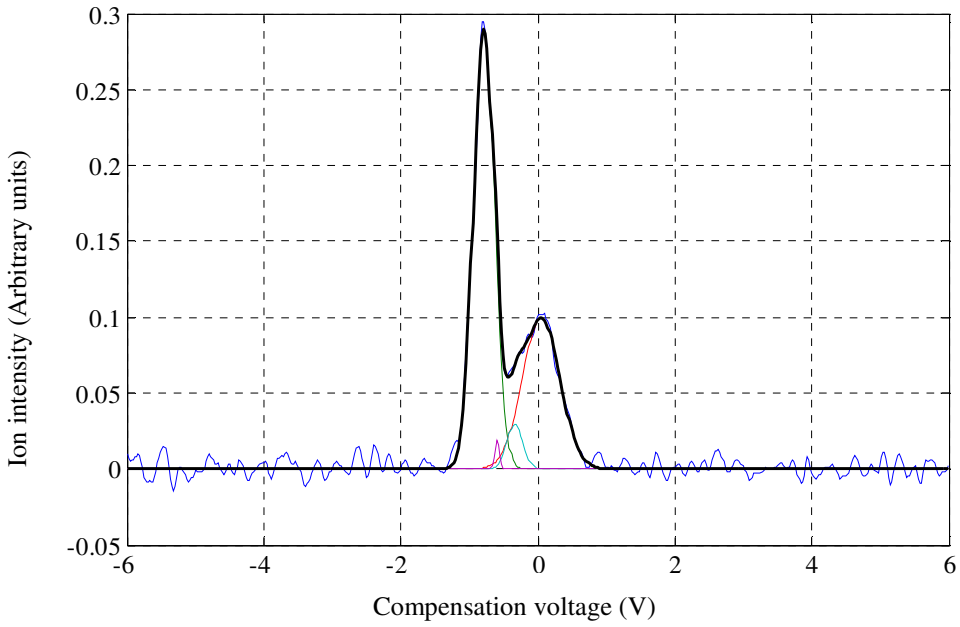

Figure 4.18 These peaks have been fitted using the FMINSEARCH. The sum of the individual peaks is overlaid (black line) providing an impression of how good the fit is to the baseline corrected response form the Faraday cup (blue line).

In Figure 4.18, it can be seen that the modification to the peaks initially fitted is not dramatic. However, the use of the additional procedure has made some important amendments. The first is that the peak width is a much closer match to the baseline corrected data. The initial peaks had, by design, over estimated the true signal. Another important amendment is that where there is an overlap between Gaussian peaks priority is not given to the response with the greatest amplitude and instead it is shared between the peaks that mix. The black line in Figure 4.18 shows the sum of the fitted peaks and it is rare that it exceeds the blue line which represents the output from the Faraday cup. This is a good indication that the fitted peaks could accurately describe the response from the Faraday cup.

Two ion responses dominate the result while two remaining fitted ion responses could either be discarded (especially the smaller response since it is as great as the noise) or considered as an indication of asymmetry of one of the major ion responses.

4.6.2 Recording and storing of final peak fit

To monitor the evolution of peaks over time the observed peaks should be recorded with an identifying time stamp. This enables results of the peak fitting to be evaluated with respect to the sweeps preceding and following the current sweep and saves the data allowing later analysis.

4.7 Limitations of presented peak fitting

As has been previously stated, unless the results of a peak fitting can be independently confirmed they should never be treated as proof of what is really occurring without greater information. With regards to the chemistry and physical processes underlying the response conclusions will be qualitative without complimentary analysis. It may be possible to discern several peaks under an ion response at low dispersion field strength and then later separate the same number of peaks fully resolved at higher field intensity. This would be an example where greater confidence could be given to the peak fits made at the lower DF.

Peak fitting is best used to provide an appreciation of how the major ion responses change over time. Such a treatment also enables a larger data set to be sampled allowing easier identification of erroneous fits. The evolution of ion responses can also be monitored, which can be a powerful tool in explaining what has occurred within a study. Additionally, it is normally straight forward to automate peak fitting, allowing a large amount of data analysis to be completed autonomously for later reviewing.

While the final process of peak fitting was common to all approaches investigated there were three separate methods of approximating the properties of discrete ion responses.

Each of these methods has their own benefits and drawbacks and these are considered within Sections 4.7.1 - 4.7.3.

4.7.1 Limitations: isolation of maximums

The method of isolation of maxima is sensitive to the noise threshold. As mentioned in Section 4.2.3 a user defined constant is required and it is through this quantity that the sensitivity for a particular data set can be tailored. A suitable constant is usually well characterised by the S/N of the sample data. A high S/N not only means that a smaller constant is required but also that the majority of the true ion response will undergo peak fitting. With a low S/N the constant should be increased to ensure that only the true ion response is passed on for peak fitting.

Another drawback to the isolation of maxima is the method of how the FWHM is estimated (Section 4.3.2). The procedure is short and simple which helps increase the eventual total speed of calculation but highly mixed ion responses can pose a considerable problem. If a maximum is found in between two other large ion responses the tracking of a point to below half of the maximum will include the width of the mixed ion responses, grossly overestimating the FWHM. An example of this situation is shown in Figure 4.19.

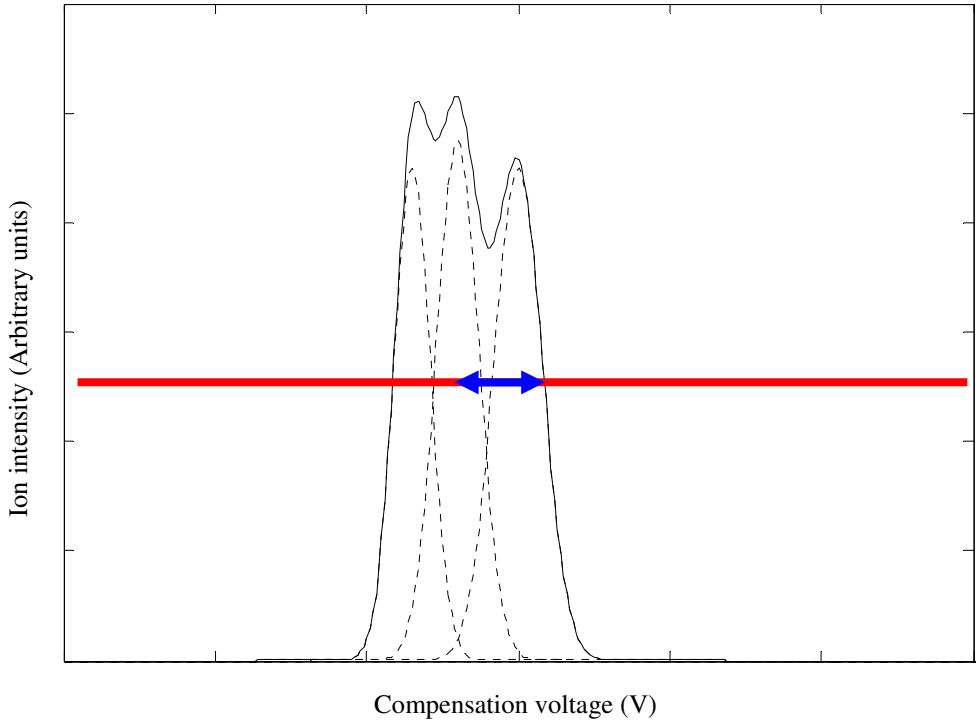

Figure 4.19 Simulated response from three ion species. Constituent peaks (dashed black) and sum total (solid black) are shown alongside line of half the maximum (solid red) and returned erroneous FWHM (blue arrow). Isolation of maxima will first attempt to fit to central ion response. Procedure means that FWHM of underlying ion response is overestimated.

To prevent this from occurring, instead of the initial peak fit being made to the point with greatest ion intensity the first maxima (point greater than two neighbours) as reading the data from low to high CV was used. This meant that the FWHM could always be approximated from the left hand side as there should not be any other ion response present to have mixed with. Unfortunately points of inflection caused this approach to fail and resulted in a higher failure rate than the situation depicted in Figure 4.19.

4.7.2 Limitations: differential

To counter a large number of initial peak fits due to a low S/N smoothing of the data set was investigated (Section 4.9.1). However, smoothing has a direct impact upon the gradient within a CV sweep so that the differential method of isolating peaks can become

compromised. Too much smoothing and genuine ion responses would not be detected while too little would result in a peak fit being made to noise. The level of smoothing was eventually linked to the number of initial peak fits, a large number of initial fits led to greater smoothing.

The differential method of peak isolation was found to be the most sensitive technique to mixed ion responses. This led to it being employed with DF sweeps because peaks would become successively more resolved with increasing electric field. Discrete peak ion responses were therefore isolated earlier. The differential technique was used throughout Chapter 5 where DF sweeps were commonly employed through data collection.

As will be considered in Section 4.9.2 the differential method is also sensitive to the use of interpolation. As a result interpolation was not utilised when obtaining initial estimates of the discrete ion responses.

4.7.3 Limitations: Successive Gaussians

The requirement to test every perturbation means that the method of Successive Gaussians is a lot slower than either isolation of maximums or the differential method. However, since only a single general procedure was implemented the method is less susceptible to bias. At present the technique is less capable of discovering mixed ion responses than the differential method.

4.8 Tracking peaks

Successful peak fitting provides the properties of representative ion responses. Within a full investigation successive CV spectra are made. Being able to monitor the properties of major ion responses across an experiment can provide information not otherwise easily obtained. Examples of this may be the effect of a changing *E*/*N* environment within a DF sweep or the elution of chemical compounds from a gas chromatography column when coupled to a FAIMS sensor.

Peak fitting was used to understand the change in properties of the ion responses found through an exponential dilution experiment, a detailed description is provided within Appendix H alongside the procedure for fitting a mathematical expression for the recorded change in properties.

4.9 Modification of the original data set

While the main bulk of this chapter has been used to describe the separate peak fitting procedures there are additional methodologies that can be employed to compliment the work presented so far. The two methods considered here are the process of smoothing data and extending a data set, both through interpolation and extrapolation.

4.9.1 Smoothing of data

It has been found empirically that some data produced by the FAIMS system has been difficult to fit peaks to. An example of the problem encountered is from a CV sweep taken from a GC-FAIMS system shown in Figure 4.20.

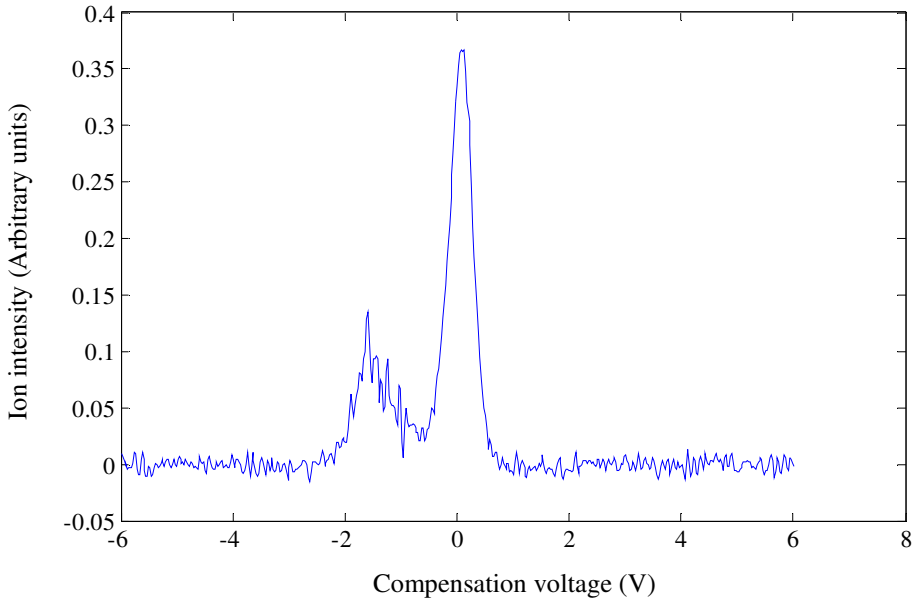

Figure 4.20 Baseline corrected raw data of a CV sweep from a thermal desorption - gas chromatography - FAIMS experiment.

The window region of the CV sweep within Figure 4.20 contains an ion response between -2 and -1 V which is very spikey. Attempting peak fitting on this region is likely to be problematic, smoothing of the data set is therefore applied.

The method of smoothing applied is based on the Kernal nearest neighbour smoothing [6] where the size of range averaged over, the greater the degree of smoothing. If too great a range is selected then important information may be lost, such as a mixed signal. In this work the range over which information was averaged was determined by the number of maxima within the unmodified sample.

Figure 4.21 shows the effects of extending the range over which the data is averaged. Specifically, Figure 4.21 (a) shows the baseline corrected data of a CV sweep. The S/N is not particularly large and it appears that the signal incorporates peaks as a result of the noise present. Fitting peaks straight to this data would result in fits due to the noise as opposed to the true ion signal. The fitting of erroneous peaks would lead to an inability to easily identify possible trends between CV sweeps.

Figure 4.21 (b) displays the same data but this time it has been smoothed over a range of two data points, the original data is also displayed as a dotted line for comparison. Already it can be seen that the minor peaks are becoming less evident while the broad shape and intensity of the ion signal remains.

This approach is taken further in Figure 4.21 (c), where the signal has been averaged over ten data points. Previous CV sweeps are again shown as dotted lines. The averaging method employed has removed minor fluctuations but also seriously attenuated the ion intensity observed. These figures demonstrate that the level of averaging required is actively dependent upon the data set.

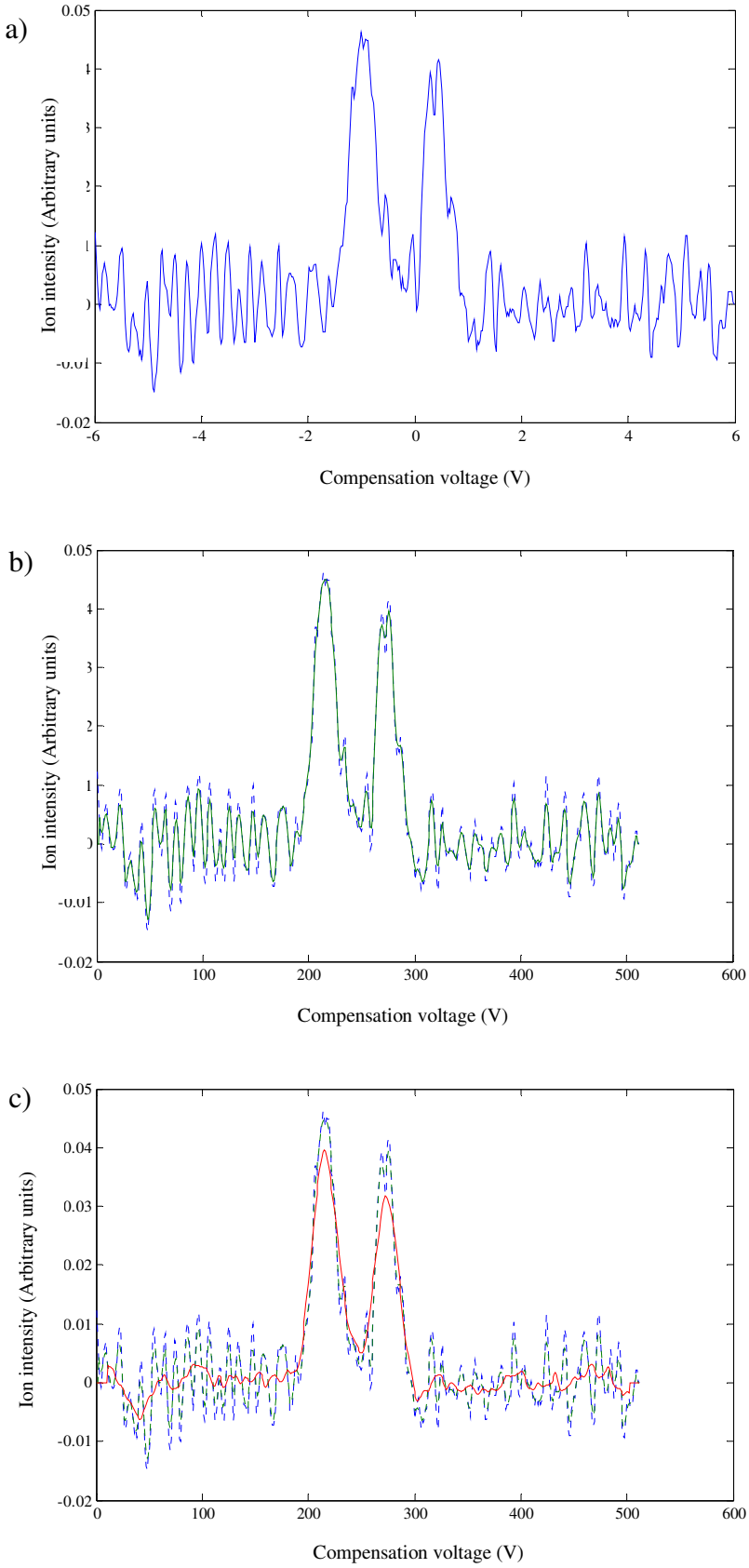

Figure 4.21 a) CV sweep of baseline corrected data b) and c) CV sweeps of the same data set with successively greater levels of smoothing

There are different smoothing techniques available (including spline [7] and Savitzky-Golay [8]) but they have not been investigated within this body of work.

4.9.2 Extending the data set

Two processes of increasing the number of data points within a set are considered. The first, interpolation, is the process of introducing additional data in-between already existing data. This differs from extrapolation which is concerned with extending a data set beyond the data points already possessed.

Interpolation was considered as some peak fits were being made using a small number of data points. It was hypothesised that including more data points would enable the non-linear least squares fit to more accurately determine peak fits. Figure 4.22 presents data where both interpolation and extrapolation has been performed.

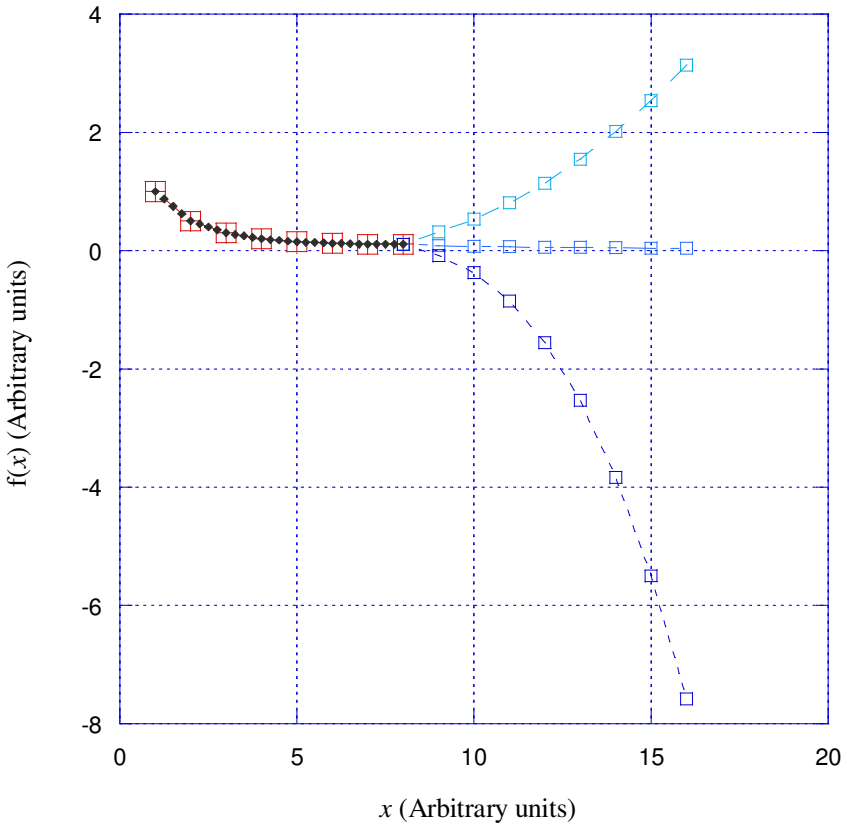

Figure 4.22 The original data points (red squares) are displayed alongside additional points (black diamonds) fitted through linear interpolation. The blue squares are extrapolations of the original data points via a second order polynomial (light blue), power law (medium blue) and third order polynomial (dark blue).

The interpolation displayed within Figure 4.22 is linear. Linear interpolation is where the data points fitted between the existing data points are consistent with a straight line between original data points. Care must be taken when differentiating linearly interpolated data as a constant gradient will be introduced between original data. The three extrapolation curves are resultant from three separate methods and aptly demonstrate how extending a data set can be flawed. The original data set is based upon a power law and if this information was always available extrapolation via the power law would undoubtedly always be chosen. If this was not known prior to analysis it would seem logical to choose the method of extrapolation that best suits the original data set. A polynomial would fit the original data well but it can then provide a widely varying result beyond the known values.

Owing to the potentially high errors achievable with using an inappropriate extrapolation method they were not employed within this work. An exception was the determination of the limit of detection from an exponential dilution experiment (Appendix H) because the underlying character of the data was well understood.

Two methods of interpolation were considered, linear and Fourier transform. Linear interpolation has been described previously. Fourier transform interpolation uses a more involved technique that requires performing Fourier analysis to discover periodic functions which describe the data.

The use of both interpolation techniques was restricted since the differential technique of discovering peak guesses looked for a gradient change. The input of data points with a constant gradient between data points, as in linear interpolation, triggered many false initial peak fits. Also since the Fourier transform method looks for a periodic signal there is modification of the resultant differential of the data set, again resulting in the triggering of false guess peaks. Figure 4.23 demonstrates the difficulty in applying the use of interpolation to a data set through either interpolation method.

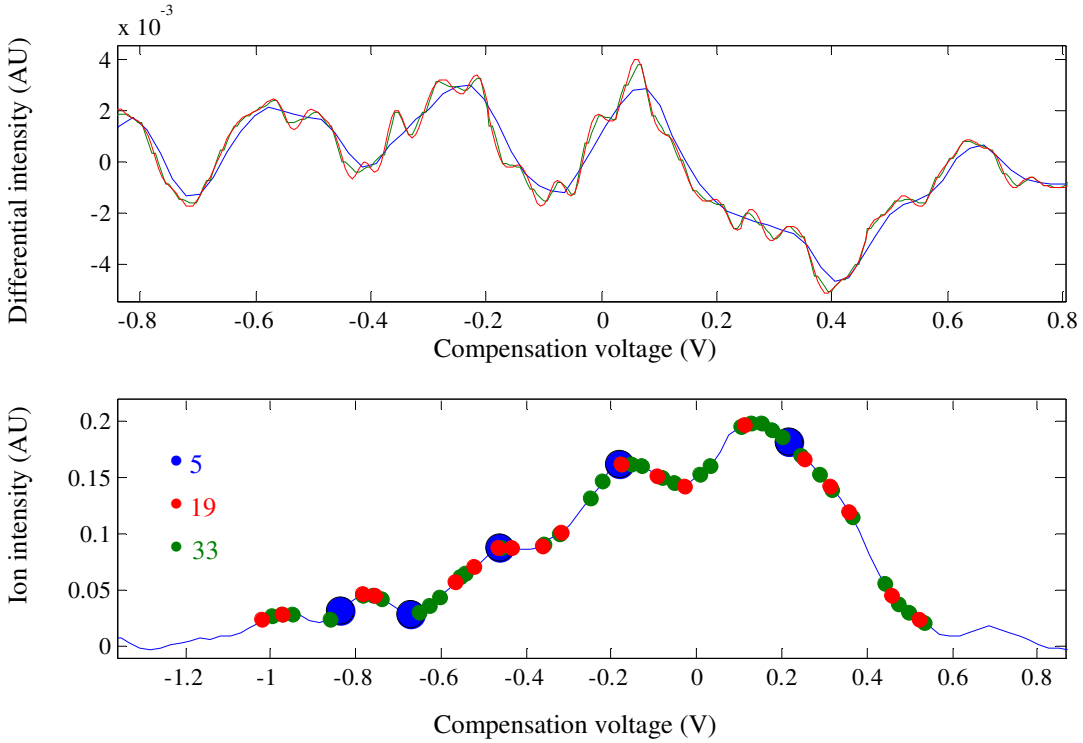

Figure 4.23 Top graph shows the differential obtained through non-interpolated (blue), linearly interpolated (green) and Fourier transform interpolation (red). Non interpolated data intensity has been scaled with the interpolation factor to aid comparison. Bottom graph shows the peak fits resulting from each differential, blue circles for non-interpolated data, green circles for linearly interpolated data and red circles for Fourier transform interpolation. Initial peak fits obtained through differential peak guess method

It can be seen from Figure 4.23 that interpolation triggers many more guess peaks compared to non-interpolated data. By observing the differential obtained following interpolation compared to the original signal it is clear that these extra guess peaks are a result of the interpolation techniques themselves, as opposed to revealing underlying information in the original data.

There has been more success in applying interpolation where the extended data set has been passed on to the non-linear least squares fit procedure of the peak fitting with only the peak guesses from the non-interpolated data. The extra data points available due to interpolation improved final peak fitting in some cases. The benefit was not consistent and similar data sets obtained better fits with different levels of interpolation. The technique with regards to the main peak fitting method and differential peak isolation techniques was

deemed to be inconsistent and resulted in too little improvement for it to be included in final peak fitting programs.

4.10 Conclusions

Peak fitting has been utilised within this thesis and the different procedures implemented have been detailed in this chapter for transparency. The separate steps required to accomplish peak fitting (*e.g.* baseline correction) have been presented in the order that they were undertaken, including three different methods of obtaining estimations of major ion responses prior to the final unconstrained non-linear optimisation.

The methods of estimating the peaks within a single CV sweep (isolation of maxima, differential and Successive Gaussians) were each addressed in turn. This treatment revealed that the isolation of maxima was the simplest method for discovering likely peaks and their properties. As the method did not require each element of a sweep to be handled analysis time of the procedure is well suited to large data sets where analysis time is restricted. Additionally, the methodology is appropriate to relatively smooth and un-complicated data sets but does demonstrate bias to signals of greatest ion intensity. Some of these short comings can be addressed through smoothing but analysis time will be increased.

The initial estimation of peaks through the differential method was the most accomplished at dealing with low resolution between major ion species. The procedures implemented meant that every element of a series was handled at least once, resulting in an increased analysis time for larger data sets. The technique is not aided through interpolation while

smoothing of a data set, over an appropriate range, is typically beneficial. There is bias within the method and it is critical that only signals resulting from ion response is assessed.

The Successive Gaussian method is the most robust to fluctuations in the data set without smoothing and shows little bias. The analysis time of the technique, however, is normally considerable and sensitive to the size of the data set. It is therefore a technique that could only be applied after data was collected while the two previous could potentially be completed within the timeframe of a single CV sweep, permitting near real-time peak fitting. The method of Successive Gaussians was developed as an answer to some of the limitations of the two other techniques and is most effectively deployed on data sets where they have failed.

Throughout the chapter it has been highlighted that the peak fitting presented is not currently considered a mature enough tool to be able to discover un-realised information from a single CV sweep. It does, however, have the potential to enable greater levels of information to be acquired concerning mixed ion responses and the true number of major ion species present. This is particular true if the peak fitting is applied to evolving data (*e.g.* DF sweeps) as it permits self assessment of returned results.

4.11 References

1. Ells, B., Barnett, D.A., Froese, K., Purves, R.W., Hrudey, S., and Guevremont, R., *Detection of chlorinated and brominated byproducts of drinking water disinfection using electrospray ionization-high-field asymmetric waveform ion mobility spectrometry-mass spectrometry.* Analytical Chemistry, 1999. **71**(20): p. 4747 - 4752.
2. Owlstone_Ltd. *The Tourist, Field Asymmetric Ion Mobility Spectrometry* 2010 [cited Aug 2010]; Available from: http://www.owlstonenanotech.com/PDF/Tourist_2pager.pdf.
3. Davis, D.M., Harden, C.S., Shoff, D.B., Bell, S.E., Eiceman, G.A., and Ewing, R.G., *Analysis of ion mobility spectra for mixed vapors using Gaussian deconvolution.* Analytica Chimica Acta, 1994. **289**(3): p. 263 - 272.
4. Shvartsburg, A.A., Tang, K., and Smith, R.D., *Modeling the resolution and sensitivity of FAIMS analyses* Journal of the American Society for Mass Spectrometry, 2004. **15**(10): p. 1487 - 1498.
5. Spangler, G.E. and Miller, R.A., *Application of mobility theory to the interpretation of data generated by linear and RF excited ion mobility spectrometers.* International Journal of Mass Spectrometry, 2002. **214**(1): p. 95-104.
6. Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning*. 2001: Springer.
7. Hastie, T.J. and Tibshirani, R.J., *Generlized Additive Models*. 1990: Chapman and Hall.
8. Savitzky, A. and Golay, M.J.E., *Smoothing and Differentiation of Data by Simplified Least Squares Procedures.* Analytical Chemistry, 1964. **36**(8): p. 1627-1639.