Improving Data Analysis by Accounting for Noise Variation

Robert Alecio [1], Stuart Ray [1], and Tony Ferrige [1], Keith Waddell [2].
[1] Positive Probability Ltd, Isleham, U.K. [1] PerSeptive Biosystems, Framingham, MA,

Overview

Advanced data reconstruction methods are used to extract more information from spectra than traditional methods but they have relied on the average data noise level to determine the convergence point, even for spectra with a strongly varying noise level. This work shows the benefits of correctly accounting for any noise variation.

A. Superior, evenly filtered results are obtained from advanced non-linear filters.

B. The ability of deconvolutions to detect features in the data is not compromised by an inappropriate noise level.

C. The amount of intensity recovered for very weak features is substantially improved.

D. Superior deconvolutions and quantified error bars are obtained through correctly fitting the data regardless of the noise level.

A. Filtering Noise

Introduction

Linear filters – e.g. Savitsky-Golay, Fourier smoothing and triangular averaging – have a uniform effect regardless of the S/N because they are designed to remove high frequencies at the expense of broadening signals. However, non-linear filters – e.g. ***Enchant™*** - are able to filter noise with only minimal peak broadening.

Until recently, the ***Enchant™*** program has used the average data noise level but this is inadequate when the noise level changes markedly across the spectrum. In these cases, noise that is well above the average value is treated as signal and is not filtered. Conversely, weak signals with a low noise are treated as noise and become broadened.

Regardless of any applied instrument or user filter, ***Enchant™*** computes the noise throughout the data before going through its iteration cycle. By accounting for the underlying noise level, noise is uniformly filtered without broadening signals, irrespective of their S/N.

Experimental

The data are the MALDI spectrum of a 22 kDa polymer. The signals of interest are on a falling baseline and the noise at low m/z is more intense than the signals. This is clear in ***Figure 1***. The top trace shows the raw data and the lower trace is the baseline corrected result using an advanced baseline correction algorithm.

The baseline corrected data were first filtered using the ***Enchant™*** algorithm but using an average noise level. The computation was then repeated but taking into account the varying S/N across the spectrum.
Finally, the baseline corrected data were deconvolved using the ***ReSpect™*** algorithm using both the average and variable noise levels so as to determine their effect on the distribution of noise features in the deconvolved result.

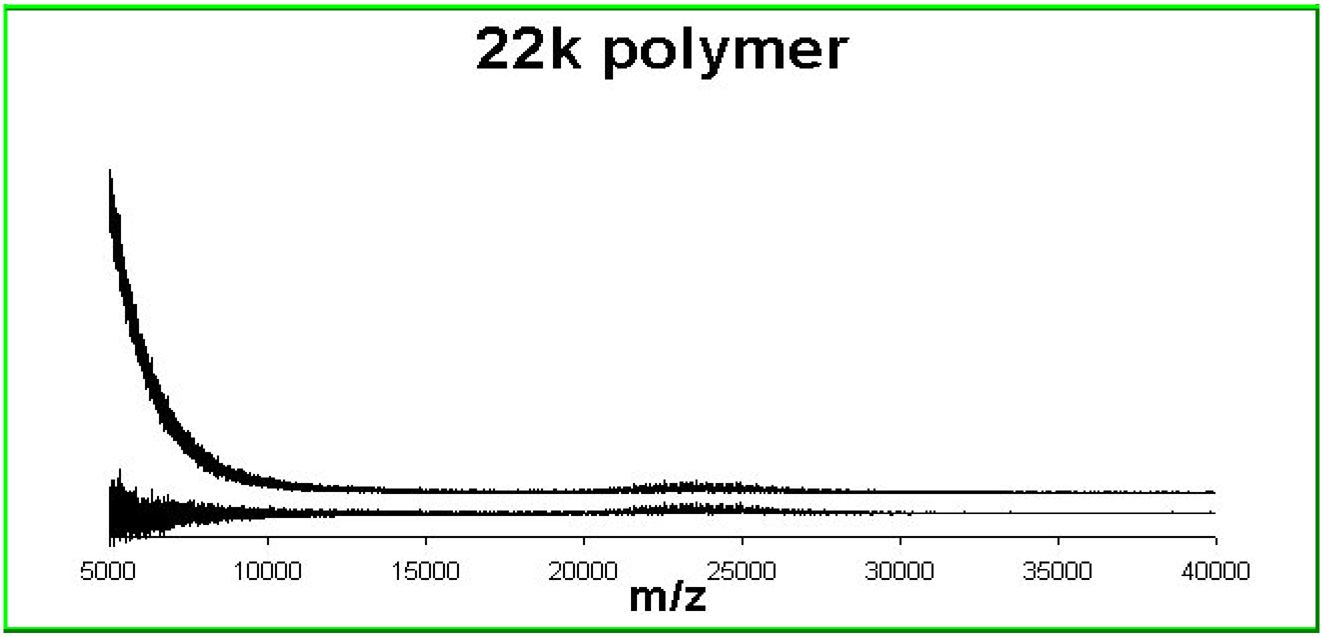
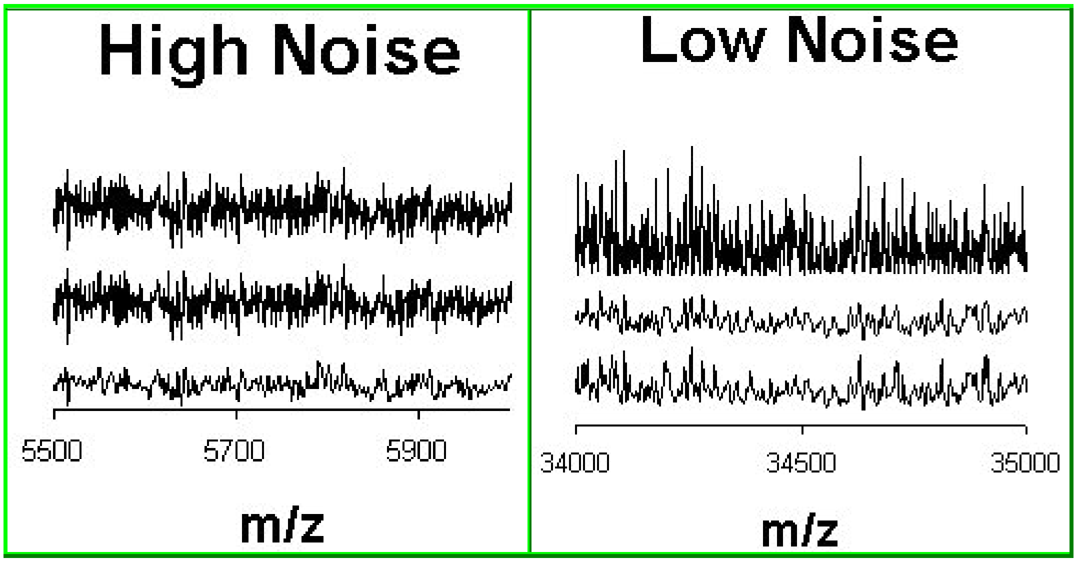


Figure 1 *Top: Data, Bottom: Baseline corrected*

Results & Discussion

The top trace of ***Figure 2a*** shows a small region of noise at low m/z. The centre and lower traces show the ***Enchant™*** results using an average and variable noise. Because the noise in this region is greater than the average for the whole spectrum, it is treated as signal and only minimal filtering occurs. By accounting for the actual S/N variation, the filtering is much more effective. In ***Figure 2b***, the top trace shows a small region of noise at high m/z. The noise is lower than the average and filtering is too extreme using the average noise. By accounting for a varying S/N the filter is less severe and the noise characteristics now match those at low m/z.



Top: 2a: Left - noise for high baseline. 2b: Right - noise for low baseline.
Centre: Effect of non-linear filter using average noise level.
Bottom: Effect of non-linear filter using variable noise level.

Note: Of necessity, the vertical scale between these two figures is relative.

__Figure 3__ compares the effect of average and variable noise levels for the signals. The noise across the peaks shown is greater than the average value so that intense noise features on the side of, or near the top of peaks are treated as signal. However, by accounting for the varying S/N, almost all these features are correctly filtered without peak broadening (bottom trace).

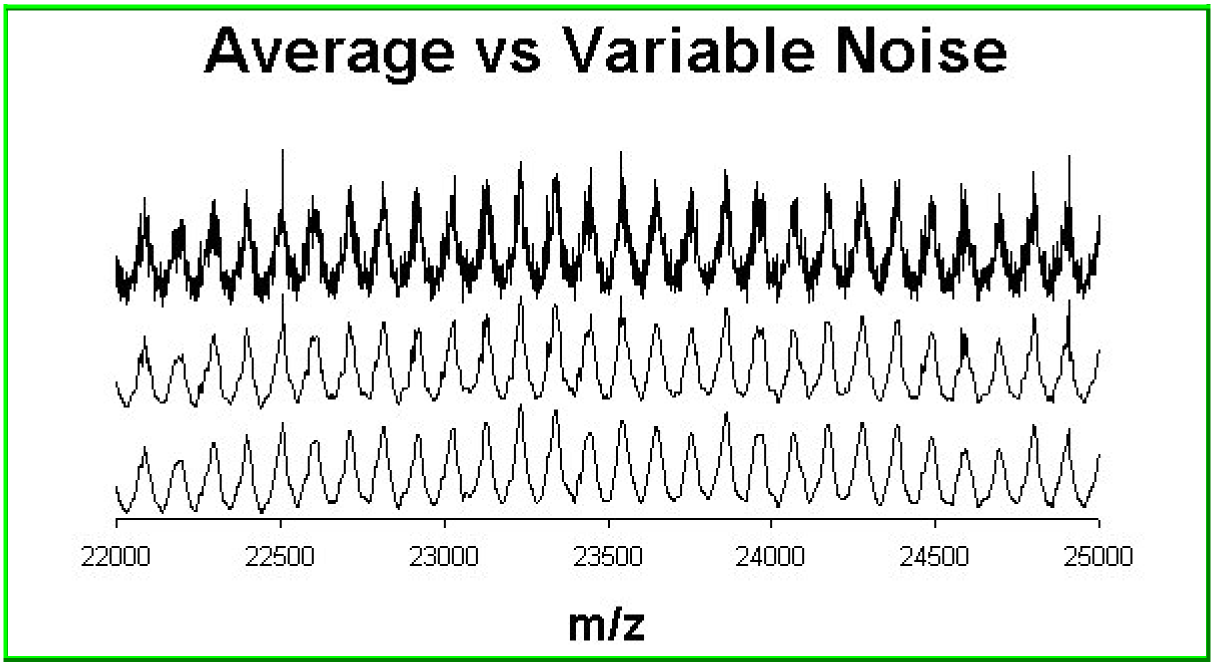


Figure 3 ***Top: Data Centre: Average noise filtering Bottom: Variable noise filtering***

Conclusions

The results obtained by applying non-linear filters that are designed to reduce noise without broadening signals are dramatically improved by taking into account any variation in the data noise level.

B. Noise Distribution

Introduction

Probabilistic deconvolution programs are iterative and an important convergence parameter is the data noise level. Like other algorithms, the ***ReSpect™*** deconvolution algorithm has, until recently, used the average noise level. However, mass spectra frequently have a substantially varying noise level. Therefore, where the noise level is higher than the average value, features that are consistent with the model will be reported as signal. Conversely, noise features that are much weaker than the average noise level will be severely suppressed because they are treated as noise. Such features will have a reduced intensity or may go undetected. Therefore, there will be a very unrealistic distribution of found noise features across the spectrum.

By taking into account the varying noise level, only those significant noise features with respect to the local noise level are detected and reported.

Experimental

The baseline corrected data for the 22 kDa polymer shown in *__Figure 1__* was first deconvolved using the average noise level. In a second computation the varying noise level was used.

Because the m/z increment is not uniform, the number of detected noise features was counted for each 5000 data points in regions where genuine signals were absent.

Results & Discussion

Table 1 shows the number of significant noise features (the low frequency noise components that fit the model) detected in each block of 5000 data points. The region m/z 16744-34124 was omitted since this contained genuine polymer signals.

When the average noise level is used, many more significant noise features are detected where the noise is high compared with where it is low. Indeed, at high m/z no signals are detected because the average noise value is much greater than the noise in this region. Using a variable noise level for the deconvolution, the distribution of detected features is, as expected, much more uniform.

It is important to note that the less than expected number of features using variable noise in the region m/z 34124-40000 arises from the fact that the data were "clipped". Therefore, this region contains numerous tiny pockets of several successive zero intensity values which reduces the number of features that may be detected.

Conclusions

By taking into account any varying noise level within the deconvolution calculation, all significant features are detected according to their S/N, regardless of their absolute intensity. Noise features are therefore evenly, rather than unevenly distributed.

Table 1: Distribution of Detected Noise Features

m/z	Average Noise	Variable Noise
5000 - 7936	63	44
7936 - 10872	55	39
10872 - 13808	48	41
13808 - 16744	43	40
34124 - 37061	4	22
37061 - 40000	0	19
Totals	213	205

C. Recovering Peak Intensities

Introduction

All properly designed reconstruction methods fit their results to within the noise level. Therefore, as the S/N of a peak is reduced, a point is reached where it becomes indistinguishable from the noise and is assigned zero intensity. Consequently, intensity errors will only be small when peaks have a reasonable area with respect to the noise standard deviation.

If the noise level changes significantly across the data, it is possible that the intensities of weak peaks in regions of low noise will be underestimated. Therefore, it is important that the deconvolution should take into account any variation in the noise level throughout the data in order to recover the peak intensities correctly.

Experimental

Figure 4 shows the baseline corrected MALDI spectrum of poly methylmethacrylate with a nominal molecular weight of 3000 Da. The zoomed regions (***Figures 5* & *6***) clearly show the presence of minor components. There are therefore numerous signals, a high dynamic range and a noise level that changes substantially across the spectrum. These data are therefore ideally suited to explore the effect of taking the noise into account when deconvolving the data to obtain peak intensities. An estimate of the true peak intensities was obtained by filtering the noise with a linear filter designed to reduce noise without broadening signals and then followed by curve fitting. The estimates are subject to some error where peaks overlap other peaks or overlap noise peaks. The data were then deconvolved with the ***ReSpect***™ algorithm using the computed average noise level. The deconvolution was repeated taking into account the underlying, varying noise level.

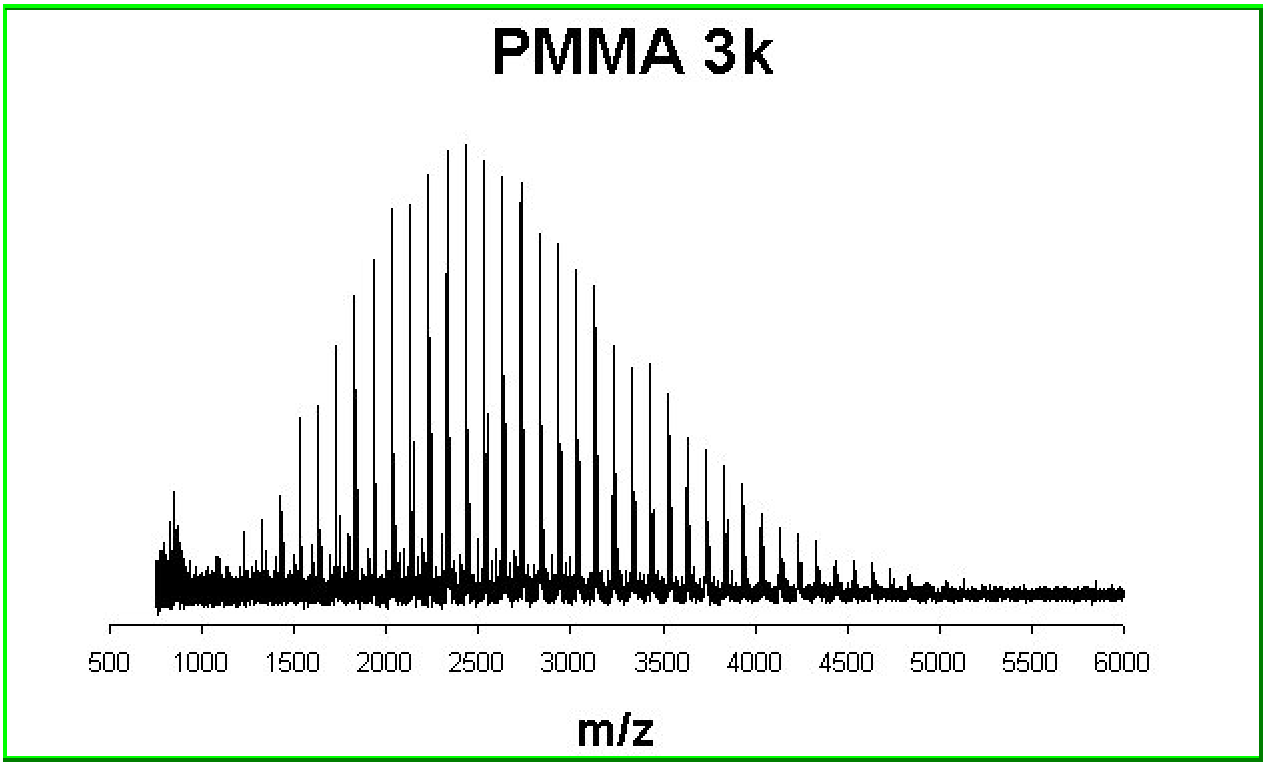


Figure 4 ***Baseline corrected data***

Results & Discussion

The table accompanying ***Figure 5*** shows how much of the intensity for each peak was recovered by using the average noise level and by taking the underlying noise level into account. There is little difference in the results but more intensity is recovered for weak peaks by correctly accounting for the noise. The effect increases as the peak S/N falls, as shown in the table accompanying ***Figure 6***.

Figures 7* & *8 show how the recovered peak intensity is correlated with the estimates of the true peak intensities. ***Figure 9*** compares the results for using an average and a varying noise level. It is clear that much more of the true peak intensity is recovered for weak peaks when any noise variation is taken into account. As expected, the recovered intensity markedly accelerates towards zero as the assumed noise level is approached. This effect is dramatically reduced when the noise variation is taken into account.

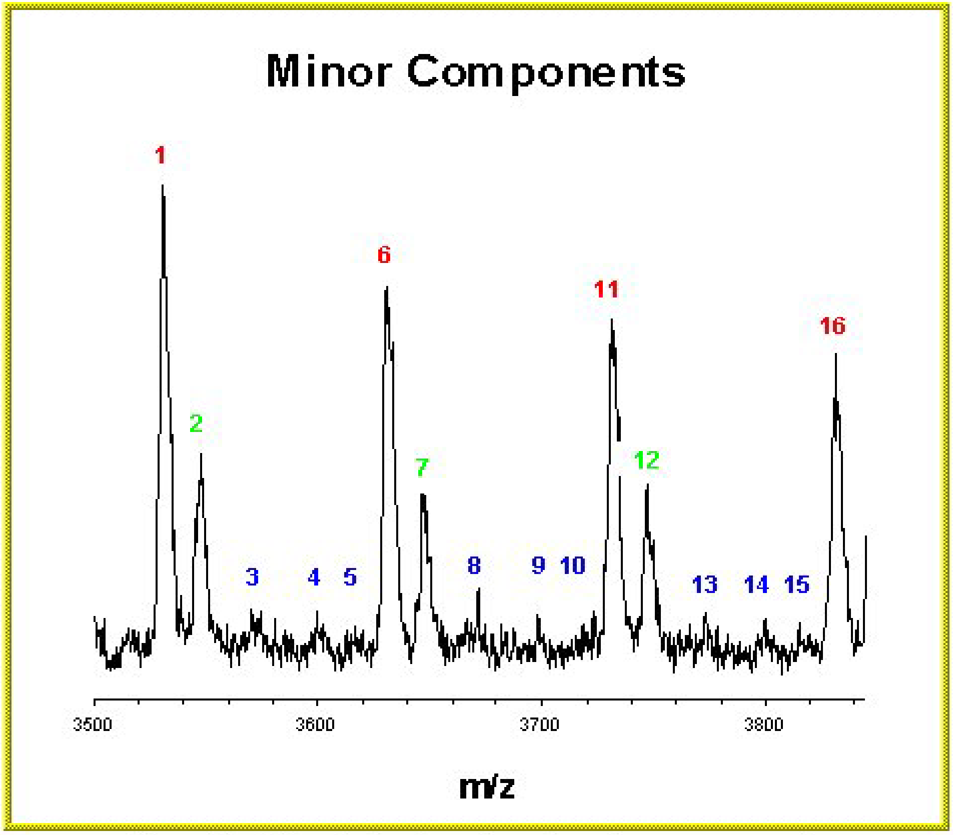


Figure 5 ***Selected region showing minor components***

% Recovered Intensity

No	m/z	Ave N	Var N
1	3531.2	99.7	99.6
2	3547.0	99.8	99.7
3	3571.5	94.0	94.2
4	3599.7	91.6	91.8
5	3616.5	90.0	91.2
6	3631.4	99.4	99.5
7	3647.2	99.0	99.2
8	3670.4	95.6	96.0
9	3698.3	97.6	98.1
10	3716.0	90.3	92.9
11	3731.6	99.5	99.6
12	3747.2	98.3	98.4
13	3772.8	92.5	93.1
14	3799.7	97.7	98.2
15	3816.9	95.1	95.5
16	3831.4	99.2	99.2

Percentage of the intensity recovered using an average noise level (Ave N) and a variable noise level (Var N) for the deconvolution.

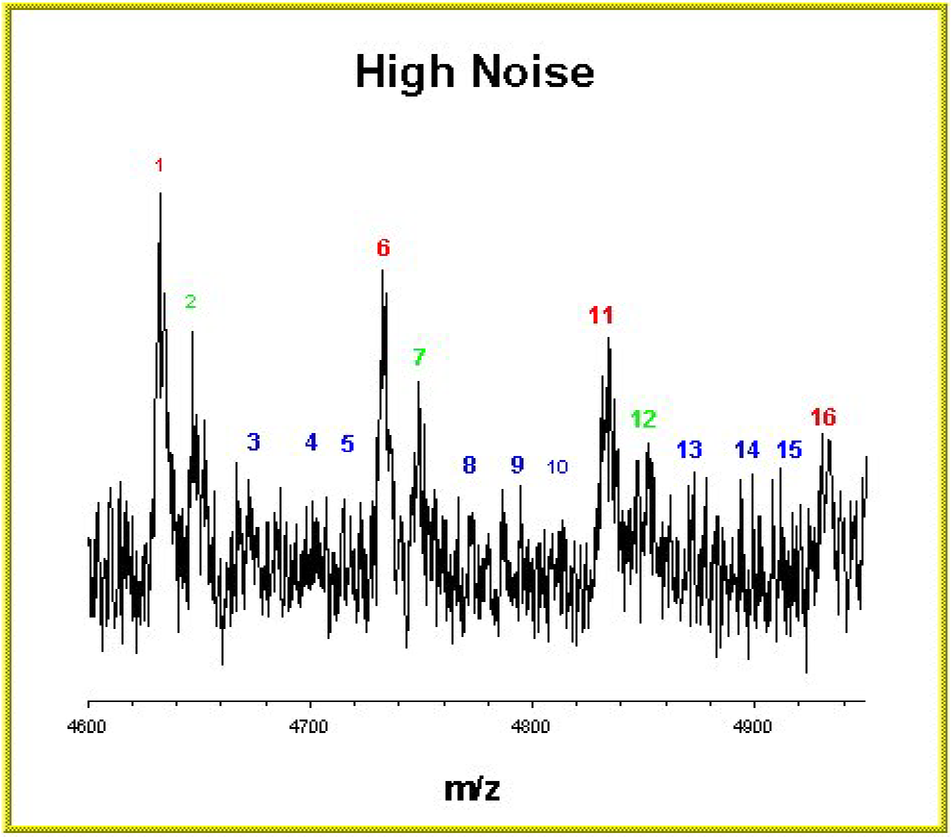


Figure 6 ***Selected region showing poor S/N for weak components***

% Recovered Intensity

No	m/z	Ave N	Var N
1	4632.6	96.0	96.8
2	4647.9	93.6	95.4
3	4671.6	89.1	92.6
4	4700.1	75.8	81.7
5	4716.2	78.0	85.0
6	4732.9	96.1	97.2
7	4748.3	94.9	96.6
8	4772.6	88.9	93.1
9	4796.4	58.3	73.5
10	4813.9	62.5	75.1
11	4833.2	94.8	98.9
12	4850.0	93.1	95.7
13	4872.0	82.0	88.9
14	4899.8	71.4	89.4
15	4916.9	66.5	86.6
16	4932.1	91.8	94.1

Percentage of the intensity recovered using an average noise level (Ave N) and a variable noise level (Var N) for the deconvolution.

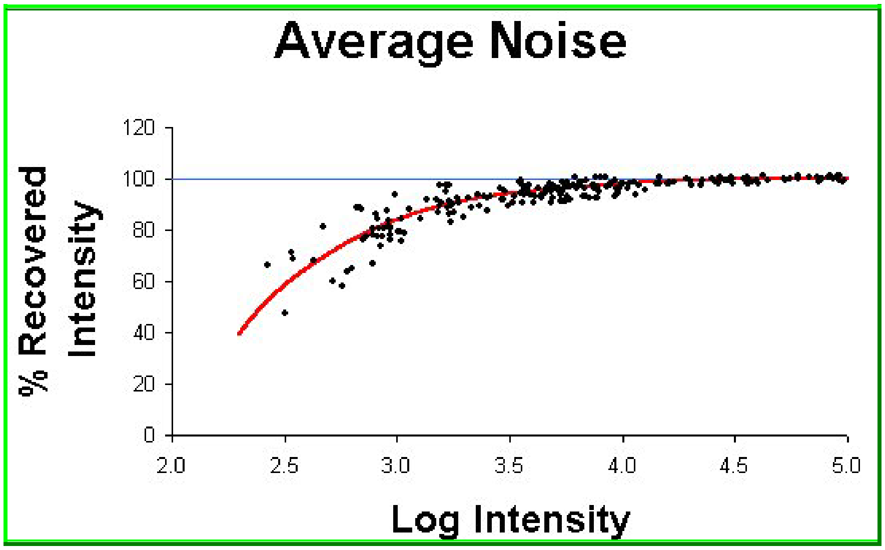


Figure 7 *Recovered intensity profile using average noise level*

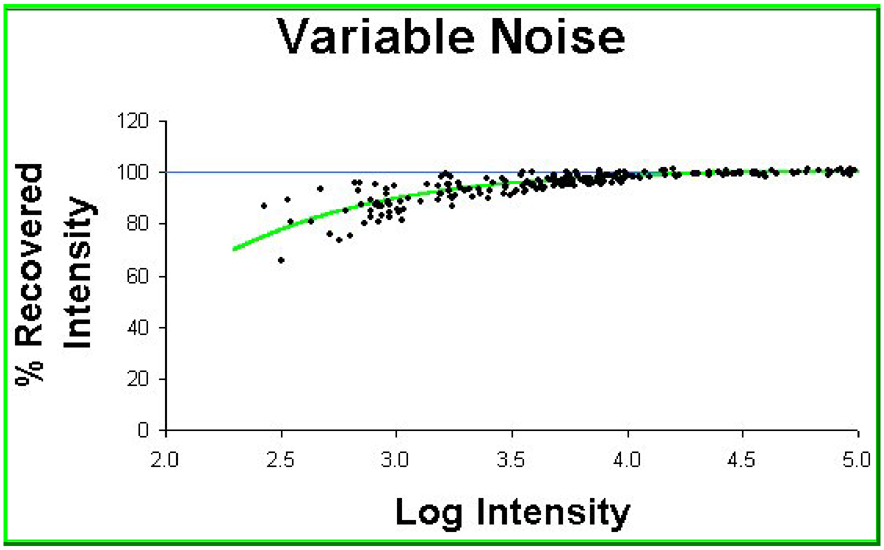


Figure 8 *Recovered intensity profile using variable noise*

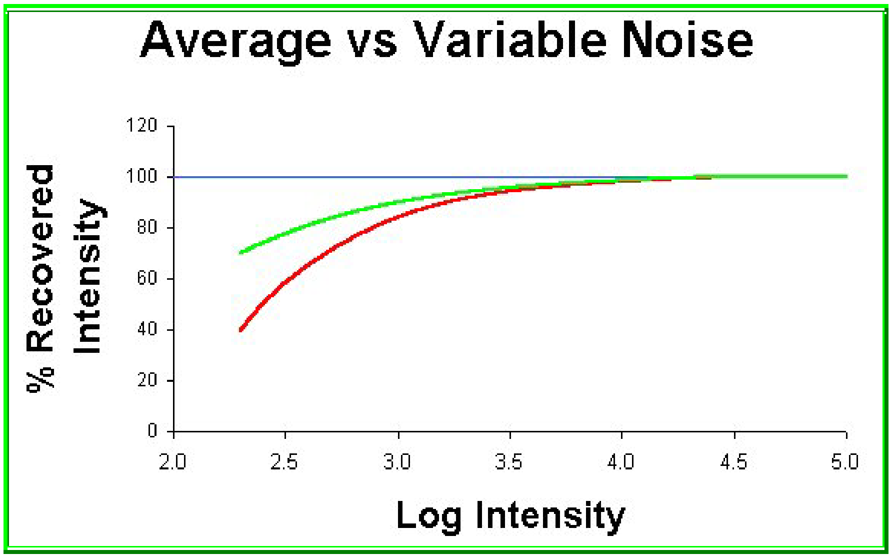


Figure 9 *Red: Average noise recovered intensity; Green: Variable noise recovered intensity*

Conclusions

For data where the noise level changes across the spectrum, much more of the true intensity of weak peaks is recovered when any variation in the underlying noise level is correctly taken into account.

D. Improved Deconvolutions & Significance Levels

Introduction

What applies to noise (Section B) applies to signals. If an average noise level is used for deconvolving data where the noise level varies, it follows that where the noise is higher than average, signals will be over-fitted and where it is lower they will be under-fitted. Also, weak signals with a noise level below average will have less significance. The deconvolution is less efficient and severely overlapped peaks are resolved less. By accounting for noise variations the deconvolution efficiency is the same throughout the spectrum. The ability to resolve overlapping peaks is not compromised and weak peaks with low noise are efficiently resolved. The computed m/z and intensity errors are consequently more reliable.

Experimental

The top trace of ***Figure 10*** shows typical MALDI-TOF DNA sequencing data. The noise level changes substantially across the data. Although peak overlap becomes severe at high m/z, the peak width in data points is relatively constant and these data are well suited to explore the efficiency of deconvolutions when the noise varies across the spectrum. The baseline corrected data shown in the lower trace of ***Figure 10*** were first deconvolved using the average data noise level and the program default values for convergence. The deconvolution was then repeated but taking into account the changing noise level.

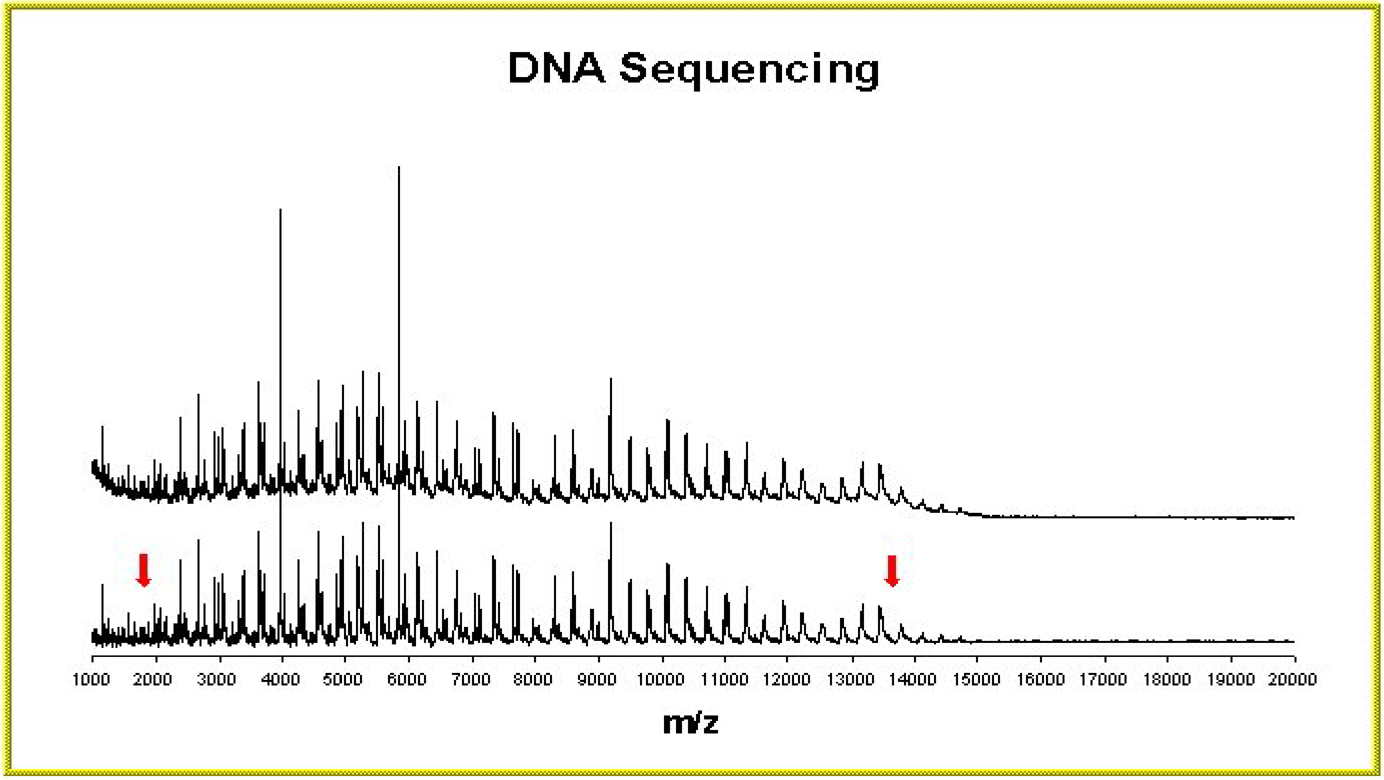
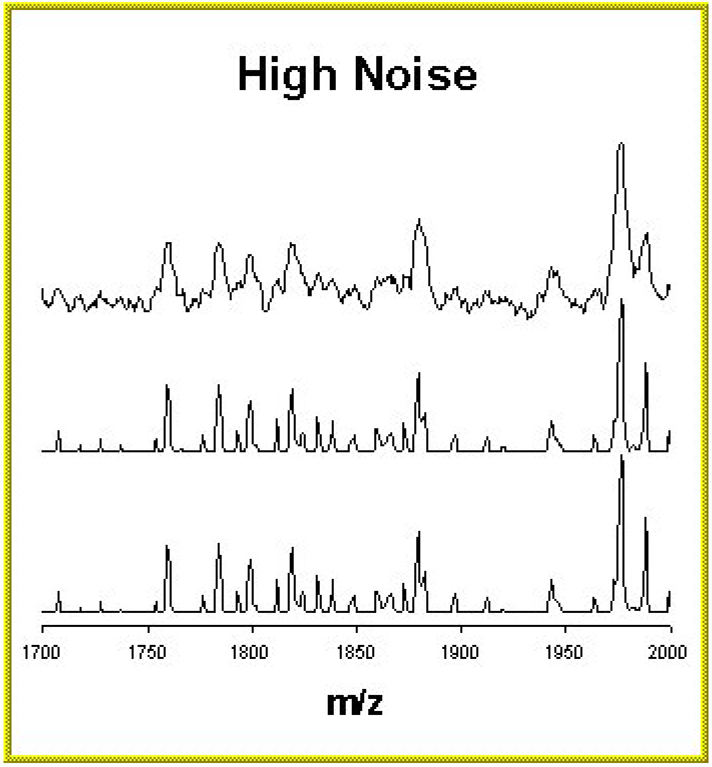


Figure 10 *Top: DNA sequencing data Bottom: Baseline corrected*
Arrows indicate regions shown in Figures 11a & 11b

Results & Discussion

The top trace of ***Figure 11a*** shows some weak peaks where the noise is high along with the deconvolved results using average and variable noise. Because the deconvolution is not compromised by the average noise estimate being too high, there is little difference in the results. ***Figure 11b*** shows weak overlapped peaks at high m/z where the noise is low. Here, the average noise level is too high and the resolution of overlapped peaks is inferior compared with using variable noise.



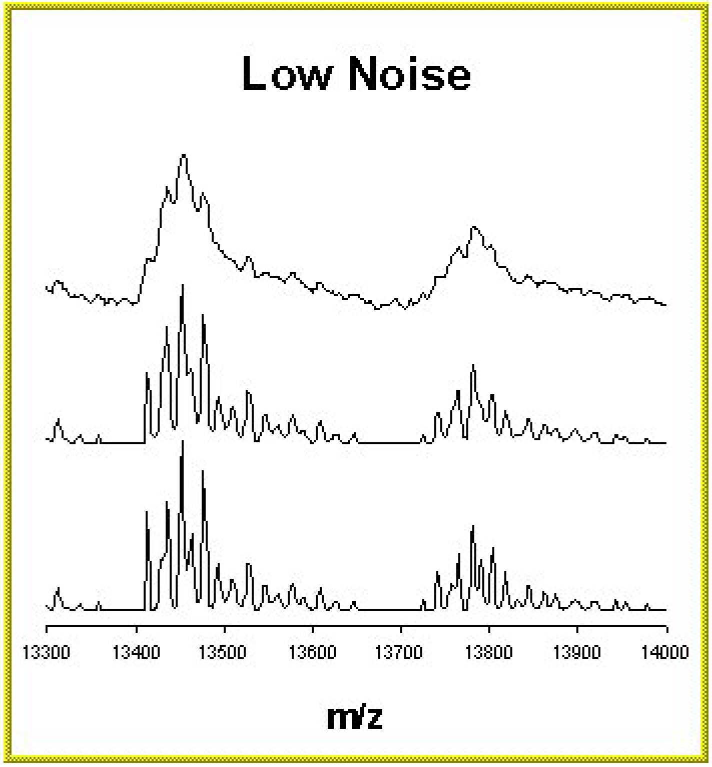


Figure 11a **Figure 11b**

Top: *11a: Weak peaks where noise is high. 11b: Weak peaks where noise is low.*

Centre: *Deconvolved result using average noise level.*

Bottom: *Deconvolved result using variable noise level.*

Table 2 shows the found m/z values and their errors to be virtually identical where the noise is higher than the average since moderately over-fitting the data will have little effect compared with its correct fitting. Conversely, where the noise is lower than the average, the data are under-fitted, resulting in higher error estimates. The same is true for the intensities and, proportionally, the errors are lower where the noise is low and more of the intensity is recovered using variable noise because the data are fitted correctly. Slightly more intensity is recovered where the noise is high due to over-fitting.

For this particular sample, the sequence contains three sites of degeneracy where the DNA synthesis products varied at three particular sites. One site has a G/T split, the second another G/T and the third a C/T split. The higher mass peaks are therefore clusters of 6 peaks corresponding to variations that contain combinations of C, T and G. Since the mass difference between T & G is 25.01 and C & T is 15.01, it is possible to predict the mass differences between the adjacent theoretical peaks in the clusters.

Table 2: Comparison of Quantified Errors

Average Noise Level					Variable Noise Level				
m/z	**Err**	**Int**	**Err**	**Int/Err**	**m/z**	**Err**	**Int**	**Err**	**Int/Err**
1759.6	0.3	3700	171	21.7	1759.6	0.3	3628	169	21.4
1784.2	0.3	3570	171	20.9	1784.2	0.3	3491	171	20.4
1799.3	0.5	2750	217	12.6	1799.2	0.5	2660	211	12.6
1819.2	0.3	3414	153	22.3	1819.2	0.3	3344	153	21.8
13527.1	0.8	2277	62	37.0	13527.1	0.7	2276	49	46.0
13547.3	0.5	1502	63	23.8	13547.1	0.5	1506	54	27.8
13561.1	1.4	966	67	14.5	13560.8	1.2	1013	55	18.6
13577.3	1.2	1471	79	18.7	13576.9	0.5	1400	59	23.7
13590.0	2.2	542	78	7.0	13589.5	1.8	661	69	9.6
13608.5	0.8	892	66	13.5	13608.5	0.7	898	54	16.6

Note: The intensity divided by the error provides an estimate of the significance of the peak.

The deconvolution does not separate all peaks in the high mass clusters when the average noise is used. However, the superior deconvolution obtained using a variable noise level resolves all peaks so that a full analysis is obtained. The variable noise deconvolution results are summarised in ***Table 3***.

Table 3: Comparison of Mass Differences

Theory	Cluster 1	Error	Cluster 2	Error
0.00	-	-	-	-
15.01	15.51	1.96	15.26	1.97
25.01	23.06	1.64	23.80	1.72
40.02	39.26	1.58	39.78	1.71
50.02	50.00	1.73	49.97	1.51
65.04	64.46	1.57	62.28	1.80

Note: Cluster 1 is centred around m/z 13450 and Cluster 2 is centred around m/z 13780.
Differences are measured from the first peak of each cluster.

Conclusions

The quality of the deconvolved results and the corresponding quantified masses and intensities are compromised by using the average data noise level, particularly in regions where the noise is lower than the average value. By correctly taking into account any variation in the noise across the spectrum, superior deconvolutions are obtained and peak overlaps are more readily resolved. This improved fitting to the data also provides robust quantified errors and absolute intensities that are correctly fitted within the noise throughout the data. Moreover, in the example presented here, the improved deconvolution allows a full analysis that could not be achieved when the average noise level was used.

***Enchant*™ and *ReSpect*™** are trademarks of Positive Probability Limited (PPL).