Tina Memo No. 2016-007
Preprint submitted to JAAS, available here as Green Open Access.

The Statistical Properties of Raw and Preprocessed ToF Mass Spectra.

A.P. Seepujak, N.A. Thacker, P.D. Tar and J.D. Gilmour.

Last updated
21 / 8 / 2016

The Statistical Properties of Raw and Preprocessed ToF Mass Spectra

A.P. Seepujak,[1*] N.A. Thacker,[1] P.D. Tar,[1] and J.D. Gilmour[2]

Single–valued summaries, such as signal–to–noise ratio, are insufficient to fully describe the input and output characteristics of time–of–flight mass spectra preprocessing methods. A detailed understanding of uncertainty, biases and correlations is essential for selecting appropriate methods, and for drawing confident scientific conclusions from preprocessed data. We apply a range of diagnostic tests to mass spectra, allowing statistical and systematic sources of uncertainty to be assessed throughout the typical stages of a preprocessing pipeline. Baseline correction, alignment and peak detection are reconsidered, with an emphasis on producing outputs with statistical properties compatible with an independent Poisson ion counting process. Benchmarking is also performed against the popular preprocessing suite SpecAlign. In contrast to other preprocessing methods, new techniques are presented which provide improved statistical stability. The benefits are demonstrated using simulation and also data from the RELAX (refrigerator–enhanced laser analyser for xenon) mass spectrometer. A factor of approximately two improvement in accuracy of Xe peak measurement over the original method for the same dataset is observed.

1 Introduction

Mass spectrometry separates ionised species by their mass–to–charge ratio in order to identify and quantify molecules or isotopes. Whilst all mass spectrometers contain the essential components of an ion source, a mass analyser and a means of ion detection,[1] there exists a wide variety of each of these three components[2]. Mass spectrometry has applications across the range of biological and physical sciences[3, 4, 5, 6]. In this work, we present preprocessing techniques that are relevant to mass spectrometers equipped with time–of–flight (ToF)[7, 8, 9] mass analysers. In particular, the methods are suited to, but not limited to, noble gas isotope analysis.

An **ideal** mass spectrum can be viewed as a histogram, with each bin representing a specific mass range. The frequencies recorded in each bin, being independent counts of detected ions falling within that range, should be consistent with discrete Poisson counting statistics [10, 11]. In practice, however, the output signal from a typical mass spectrometer is not a discrete ion count; it is a voltage, produced by the time–varying output current from a detector as it passes through an impedance. [2] A step change in voltage will be associated with each count, therefore a translation between voltage and ion detections is required. Electrical and thermal effects introduce additional uncertainty in the form of Gaussian–type noise on the continuously monitored output [12]. Charging and discharging effects can introduce a non–zero baseline for ion counts, which may vary as a function of the signal. Mass calibration may also drift, which can misalign what should be equivalent bins, and so lead to the broadening or misidentification of peaks. These issues do not necessarily invalidate analysis approaches based on Poisson statistics, so long as appropriate preprocessing can be performed. The aim of our preprocessing is to minimise these effects so that data analysis based upon ideal Poisson statistics can be applied, such as Linear Poisson Modelling.[25, 26, 27]

Preprocessing is an essential step for the analysis of mass spectra,[13] but represents a challenging problem. Poorly performed preprocessing may prevent meaningful analysis of signal.[14] Previous methods for preprocessing include principal component analysis, independent component analysis, agglomerative hierarchical clustering, sequential approaches based on Gaussian scale–space theory and hidden Markov model–based methods.[10, 13] Bayesian approaches have also been suggested.[13] Some of the above problems with mass spectra data have been addressed by others in the form of various baseline correction[15, 16] and alignment methods.[17, 18, 19, 20] With increasing reliance on mass spectrometric quantitation, e.g. in proteomics[21] and structure determination of protein–protein complexes,[22] knowledge of uncertainty is becoming vital. Determination of error arising from fluctuations in the $\frac{m}{z}$ value has been attempted.[23] Empirical estimates of error on values computed from mass spectra have also been undertaken, using the principle that a χ^2 statistic can be forced to adopt an expected value for a specified number of degrees of freedom by fitting appropriate variance terms.[3] Additionally, analytical techniques such as error propagation[24] have been applied in order to understand errors on summaries such as metabolomics data analysis.[28]. However, the evaluation of methods often focus on narrow properties, such as signal–to–noise, or true/false positive rates. They also often make untested assumptions of independence, or neglect effects at low signal.[23] It is rare for algorithms to be assessed more comprehensively in terms of multiple statistical properties, e.g. changes in noise as a function of signal, actual shapes of noise distributions, correlations within noise, or

[1] *Imaging Science and Biomedical Engineering Division, Medical School, University of Manchester, Stopford Building, Oxford Road, Manchester, M13 9PT, UK. Email: a.seepujak@manchester.ac.uk; Tel: +44 (0)161 275 5131*

[2] *School of Earth, Atmospheric and Environmental Sciences, The University of Manchester, Manchester M13 9PL, UK*

preservation of signal normalisation and shape (bias); which are exactly the tests required to confirm idealised histogram behaviour.

To address the above issues, we develop methods of baseline correction, peak alignment and peak integration specifically designed to provide output characteristics that are closer to those of idealised independent Poisson histograms. In addition, we present methods for assessing a wide range of statistical properties. These include checking the functional dependencies between signal and noise, via Bland–Altman (BA) plots[32, 31] (also known as funnel plots and Tukey mean–difference plots[10]). We assess noise correlations, using a standard correlation coefficient. We also assess how well the shape of signal is preserved, via χ^2 statistics, and check that total signal normalisation is preserved, using Pull distributions.[44] These metrics of success are selected in preference to any other evaluation approaches, as these directly quantify the statistical properties of interest. We evaluate our methods using synthetic Monte Carlo spectra and experimental spectra from the refrigerator–enhanced laser analyser for xenon (RELAX) mass spectrometer.[3] Our methods have already been applied successfully to organic mass spectra that are far more complex,[30] where resulting spectra were used for quantifying lipid samples. Here, simpler spectra are selected so that fundamental statistical properties of the data can be more easily assessed. Results are compared to the existing preprocessing suite SpecAlign, by Wong[33, 34]. This software was selected as a standard owing to its popularity and ease of availability as an open–access resource.

2 Methods

The three preprocessing steps evaluated in this work consist of baseline correction, alignment and integration of peaks:

- Baseline correction is performed to mitigate against the non-zero Gaussian noise background, thereby providing an improved zero-point for the counting of ions;
- Alignment is performed to ensure that shifts along the m/z axis are minimised, allowing mass bins to be interpreted consistently across multiple spectra;
- Peak integration is performed to reduce spectra into simpler histograms with one peak per bin, which would otherwise be spread over multiple mass bins.

To assess these methods, a Monte Carlo spectra generator and six metrics of success are applied:

- From Bland-Altman analysis, the *power-law relationship between signal and noise* is determined, on a mass-by-mass basis;
- From Bland-Altman analysis, the *scaling of errors in comparision to Poisson errors* is determined, on a mass-by-mass basis;
- Using χ^2 statistics, the *overall shape of spectra* are assessed against known reference shapes;
- Using Pull distributions, *bias in total ion counts* is assessed against known values of total signal integral;
- Using Pull distributions, the *statistical spread of total ion counts* is compared to expected Poisson variability;
- A correlation coefficient is used to check *correlations between noise on adjacent masses*, on a mass bin by mass bin basis.

Between these metrics, bin-by-bin behaviour and total signal behaviour can be comprehensively understood, with well-defined predictions for all expected results.

2.1 Data sources

In order to assess the preprocessing functions, a source of spectra with known problems and associated ground truth is required. A Monte Carlo simulation was therefore employed, based upon an experimentally–acquired RELAX air calibration spectra. [3][42] RELAX is an ultra–sensitive resonance mass spectrometer for measuring xenon isotope ratios of extraterrestrial materials. The instrument consists of a resonance ionisation ion source with a cryogenic sample concentrator, a ToF mass analyser, and a detector consisting of a pair of chevron–mounted microchannel plates. Data are acquired in analogue mode using an Acquiris DP240 8–bit digitiser card sampling

at 1 ns. Noble gas analyses including those employing RELAX are conducted statically. Thus, the signal intensity measured using the RELAX instrument exponentially decays as a function of time as xenon is implanted into the detector. For calibration purposes, a reservoir attached to the instrument supplies a known composition of xenon derived from the earth's atmosphere. These allow determination of the sensitivity of the instrument (from the signal produced from the known amount of calibration xenon) and of its mass discrimination (by comparing measured isotope ratios to the known values of atmospheric xenon). These air calibrations are used to configure the parameters of Monte Carlo data, and also act as a source of real data, against which a conventional calibration and our pre-processed calibration can be checked.

The air calibration sample also includes contributions from hydrocarbons, identified as instrumental contamination. The presence of hydrocarbons can be identified from the broadening of some peaks and the presence of visible peak at mass 126.

Our Monte Carlo simulation generates peaks with a Gaussian profile. Mass spectra with independent Poisson noise were created by drawing frequencies for each mass bin, H_i, (with i being a particular mass range) from a Poisson random number generator.[40] The expected value of the Poisson variable, $< H_i >$, is determined by the Gaussian signal profiles at each mass, i. Simulated spectra with misalignment are created by systematically adding a small random offset to the mean of each Gaussian peak. Finally, a non-zero baseline is added to each mass bin using a Gaussian random number generator with finite positive mean. 500 spectra are created per experiment, with normalisation decaying exponentially, to match real data. The properties of the spectra were assessed to configure the Monte Carlo to give comparable signal-to-noise, background, levels of misalignment, and number of spectra per dataset. Nine peaks were identified (see Table 1). The background noise (between peaks) was measured at 1.6 counts with a mean background level of 5.0 counts; the amplitude of the highest peak was 500 counts, decaying to zero by the 500th spectrum; and a peak misalignment $\frac{m}{z} < 3$ bins was measured. Fig. 1 and Fig. 2 show example simulated spectra.

The decaying quantities across the 500 spectra and the varying widths and relative peak hights presents a range of challenges for the preprocessing. The low peaks at 124, 126 and 128 are close together, ideal for testing baseline correction and peak detection. The other peaks are a mixture of distances apart and are comparatively large, ideal for testing alignment and overall signal integral.

Peak	mean	standard deviation
1	124	15
2	126	15
3	128	35
4	129	5
5	130	5
6	131	5
7	132	5
8	134	5
9	136	5

Table 1. List of spectrum replicate parameters; the replicated spectrum consisted of nine xenon isotope peaks.

2.2 Baseline correction

The continuous voltage produced from mass spectrometers is an indirect count of ions, as opposed to being an idealised direct Poisson counting process. The voltage corresponding to zero ion counts is not necessarily zero, moreover, this voltage is not necessarily constant across an entire spectra set. As a result, peaks appear upon a varying noisy background. The correction method proposed herein iteratively estimates this non-zero background, B_i, on a spectrum–by–spectrum basis, subtracting it from the original, $H'_i = H_i - B_i$, in order to provide baseline-corrected output. This provides a zero-point that actually corresponds to zero ions counted. This must be performed without biasing the signal and without introducing noise correlations.

The algorithm identifies peaks using hysteresis thresholding, which classifies bins into peaks or background.[36] Mass bins with values above an upper threshold, $H_i > t_u$, are identified as belonging to a peak. Bins adjacent to previously–identified peak bins are also designated as part of that peak if above a lower threshold, e.g. $H_{i-1} > t_l$ and $H_{i+1} > t_l$ if i is part of a peak. This thresholding can be interpreted as a pair of statistical significance tests, where peaks are expected to be some number of standard deviations above the noise, with a weaker test being

permitted on bins neighbouring identified peaks so that their tails can be identified with more sensitivity. The upper threshold is computed to be 3 standard deviations above the noise, with the lower threshold starting at 1 standard deviation.

The algorithm does not assume any fixed level of noise on the background. Baseline noise is iteratively estimated on a spectrum-by-spectum basis until the best-fitting background is found. To accommodate this, all steps (including those below) are repeated until a stable standard deviation (SD) around the baseline is achieved. The residuals between background bins and zero, $(H_i' - 0)^2$ (for i that are currently believed to be non-peak masses), are used to compute a sample SD that will change as successive iterations improve determination of peak locations. Once this SD converges to a fixed value, the process terminates.

For the current estimate of non-peak locations, bins are used to make locally–weighted estimates of the mean background. The weighting is achieved through the use of a Gaussian smoothing function

$$G(\beta) = \alpha.exp(-\frac{\beta^2}{2w}) \tag{1}$$

where w is the width of the profile, and α a normalisation constant chosen such that $\sum_\beta G(\beta) = 1$. At points where gaps exist in the background due to peaks, the normalisation of the smoothing is adjusted to compensate. The locally–weighted estimate of the mean background can then be described by the convolution

$$B_i \approx H \otimes G \tag{2}$$

The resulting smoothed background is subtracted from the original spectrum, before the RMS around the baseline is computed. This process, from peak identification, iterates with updated hysteresis thresholds, with $t_u = 3 \times SD$ and $t_l = 1 \times SD$, until the measured SD converges. Setting the lower threshold to zero on the final iteration ensures that peaks are fully extracted to their base. The positive and negative residuals around the tails approximately cancel, ensuring that no net bias is introduced as the tails enter the noise floor.

2.3 Alignment

Slight drifts in timing can cause mass peaks to move up or down the m/z scale, increasing instabilities of bin values, especially near the edges of peaks. Removing this source of variation can be achieved through an alignment process. This process should make efforts to not change the shape or normalisation of the signal, or introduce unwanted noise correlations. We assume a simple model of misalignment, which can be remedied using a constant offset. Whilst this may limit applicability to certain mass spectra datasets, in the present experimental RELAX data, such an offset was observed, thus the simple model is adopted. Additionally, related work on organic mass spectra found this offset approach to be sufficient.[30] The alignment algorithm applies two steps. Firstly, a simple alignment to the nearest whole mass bin; secondly, a sub–bin resolution alignment in the Fourier domain.

The average of a set of related spectra is used as a reference spectrum, $\bar{H}_i = \frac{1}{N}\sum_i H_i$. Alignment to nearest whole mass bin begins by applying the Anscombe square-root transform[37] to each recorded bin frequency, which transforms the Poisson counts of the bins to more Gaussian–behaved quantities with constant variance. The spectrum is then shifted through a predetermined range of bins, sufficiently large to encompass the full effects of any misaligned spectra. At each whole-bin offset, δ, the sum of squared residuals is computed between the spectrum and the reference (scaled by a normalisation term s, to match the spectrum being aligned) which is consistent with a Likeihood solution for Gaussian-shaped residuals. At the minimum, we can be confident that the best sub-bin solution (see below) is less than 1 bin away from the current whole-bin estimate:

$$\arg_\delta \min \sum_i (\sqrt{H_{i+\delta}} - \sqrt{s\bar{H}_i})^2 \tag{3}$$

Secondly, to align to a sub–bin resolution, a technique similar to sinc interpolation[38] is applied. The Anscombe transform is performed on the reference and misaligned spectra. Shifting is achieved by manipulation of the Fourier terms. Each term corresponds to a frequency component in the spectrum. The coefficients are updated by adding phase shifts in a way that achieves a fixed shift of all components in the mass domain. The solution is given by computing sine and cosine terms, with a discrete Fourier transform for the square-root reference spectrum, giving a set of J coefficients, $\bar{a}_j$ and $\bar{b}_j$:

$$\sqrt{\bar{H}_i} = \sum_{j=0}^{J} \bar{a}_j \sin(\frac{2\pi ij}{J}) + \bar{b}_j \cos(\frac{2\pi ij}{J}) \tag{4}$$

Similarly, the spectrum requiring alignment, H_i, will have coefficients a_j and b_j. The phase of a component is then given by

$$\phi_j = atan\frac{a_j}{b_j}$$

such that $a_j = m_j \sin\phi$ and $b_j = m_j \cos\phi$, where $m_j = \sqrt{a_j^2 + b_j^2}$. Updated coefficients can be computed for a relative shift of δ along the original function using:

$$\theta_j = \delta j$$

for $0 < j < \frac{n}{2}$ and

$$\theta_j = -\delta(n - j)$$

for $\frac{n}{2} < j < n$ (where the index, j, follows those in [**?**]). Applying the phase shifts gives new coefficients, $a'_j = m_j \sin(\phi + \theta)$ and $b'_j = m_j \cos(\phi + \theta)$. The δ parameter is then adjusted to minimise the distance between the reference coefficients and the updated ones:

$$\arg_\delta \min \sum_j (\bar{a}_j - a'_j)^2 + (\bar{b}_j - b'_j)^2 \tag{5}$$

A golden ratio search[40] is used to determine the best sum squared result, rather than the brute-force search, as we expect only 1 minimum to exist locally. Once a solution is found, the inverse Fourier transform is applied, followed by inverting the Anscombe transform to restore the spectrum to the original domain. The final mass bins should be independent (assuming the source spectra were independent) and the signal quality should be preserved, via Parseval's theorem.

2.4 Integrating peaks

The noisy gaps between peaks and the multiple mass bins which span the width of peaks spreads the useful information within a spectrum across an unnecessary number of bins. The width of peaks may also vary, depending upon signal strength and choice of x-axis (e.g. time domain, mass, or m/z). The area of a peak is a more representative measure of a species' abundance than the peak amplitude, with experimental spectra exhibiting a distribution rather than a delta function.[13] Additionally, a peak which spans multiple bins may posses highly variable bins near to the maximum (owing to the steep sides), if that maximum can shift a couple of bins in either direction. Integrating a peak into a single bin may reduce these shift artifacts. Lower resolution binning should also reduce any effects due to correlations between bins, as adjacent correlations are eliminated once they are added together. A peak detection and integration method can thus be applied, which should be designed to reduce noise correlations and preserve total signal integral, and also result in a simpler histogram format.

Given a set of related spectra, the average spectrum is computed, $\bar{H}_i$. Using hysteresis thresholding, all maxima above a threshold in $\bar{H}$ are marked as peaks, which avoids spurious noise spikes being allocated to their own bins. The tails of each peak are scanned, being summed into a new histogram, with one bin per peak, ceasing when the tail drops below zero or begins to rise again. This is in contrast to the thresholding performed during baseline correction, which is performed on a spectrum-by-spectrum basis. Instead, new histograms are created, with each bin being assigned to the detected peak locations from the *mean spectrum*, so that each histogram has a common binning. Signal within each peak are integrated by simply summing bin values, giving each an equal weight so as not to change the total integral. Assuming the Gaussian–like background noise has as many negative as positive instances, which should be the case after baseline correction, these contributions to the bin errors should largely cancel in the integration, further helping to avoid bias.

2.5 Bland–Altman analysis: mass bin error behaviour

Ideal Poisson spectra have the property that the SD of a mass bin should be equal to the square root of the expected bin value, i.e. $SD_i = \sqrt{< H_i >}$. This assumes that H_i is a direct count of detected ions, not voltage or ADC units. If noise in mass spectra bins is dominated by Gaussian perturbations, or artifacts from misalignment

of spectra are present, then this link between SD and bin value will be broken. A range of dependencies can be expressed as a power–law and a scaling

$$\sigma = a\left(\frac{H_i}{a}\right)^{\frac{0.5}{b}} \tag{6}$$

where the term a is a scaling factor on the variance, and b is a scaling factor on the power–law dependency (with Poisson being the reference at unity). These parameters can be estimated from replicate spectra data in a Bland-Altman (BA) plot, and then fitting the above power–law model using likelihood. The x–axis of a BA plot covers the range of expected values, $\langle H_i \rangle$, whilst the y–axis covers observed residuals, $\langle H_i \rangle - H_i$. Fig. 4 shows the difference between Gaussian and Poisson residuals.

As the expected bin values, $\langle H_i \rangle$, are unavailable in real data, they are estimated as the average of equivalent bins in the two spectra adjacent in the time sequence, thus,

$$\delta_{ij} = \left(\frac{H_{i,j-1} + H_{i,j+1}}{2}\right) - H_{ij} \tag{7}$$

The subscript i represents the mass bin and the subscript j represents the replicate spectrum. Using Poisson data, if the set of all δ_{ij} residuals is calculated and plotted, and the power–law model fitted, values of $a = \frac{3}{2}$ and b equal to unity are expected.

2.6 χ^2 and Pull analysis: total signal behaviour

The spectra used in the following experiments are all replicates. As such, they should only differ in terms of specific patterns of noise. This knowledge is used to check how well signal is preserved from one processing step to the next in terms of both overall shape and total integral.

The known average spectrum shape, R_i, (determined by the Monte Carlo parameters) is fitted to each spectrum. A χ^2 per degree of freedom statistic can then be computed to perform a consistency check for each individual spectrum.

$$\chi^2_D = \frac{1}{D}\sum_i \frac{(\sqrt{H_i} - \sqrt{sR_i})^2}{\frac{1}{4}} \tag{8}$$

where D is the number of bins in a spectrum minus 1 (i.e. minus the fitted normalisation parameter) and s is the scaling required to fit the mean to the specific spectrum. If a spectrum is the *correct shape*, χ^2_D should be on average unity.

Since Monte Carlo data are generated with a known quantity of data (simulated ion counts), this ground truth can be compared to the total integral of each spectrum to ensure no signal has been lost or gained. The difference between fitted integral and ground truth can be divided by the expected error (the square root of the Poisson count of simulated ions):

$$\Delta_j = \frac{(\sum_i H_{ij}) - (s_j R_{ij})}{\sqrt{s_j R_{ij}}} \tag{9}$$

where j indicates a different spectrum. The distribution of these differences, Δ_j, is known as a Pull distribution. Assuming the spectra exhibit Poisson behaviour and the spectrum size is *unbiased*, the Pull distribution will have unit SD and mean of zero. If found to be non-zero, the mean indicates a net bias in signal. If found to be larger than unity, the SD indicates that errors are bigger than expected from Poisson counting.

2.7 Correlations

As a final check, the residuals between a fitted mean spectrum, $\delta_i = H_i - sR_i$, can be analysed for correlations. Residuals from adjacent mass bins can be used to compute a correlation coefficient, given by,

$$r = \frac{1}{N}\sum_i^{N-1} \frac{\delta_i \delta_{i+1}}{\frac{1}{4}} \tag{10}$$

which should be consistent with zero for independent Poisson spectrum bins. Correlations significantly above zero indicate that either raw data or some pre-processing steps have introduced unintended dependencies between mass bins.

3 Experiments

A range of challenges were presented to the preprocessing methods, with results measured using the six noted metrics of success. These challenges included:

1. *Ideal Monte Carlo spectra*, containing only independent Poisson data with no misalignment and no additional background. 500 spectra produced with exponentially decaying signal over time. Preprocessing should not be required for such data, and if applied should not corrupt the data. Results in Fig. 5.
2. *Misaligned Monte Carlo spectra*, with no additional background, but with misalignment of peaks within ± 2 mass bins. Alignment is hence necessary, but additional steps should not corrupt the data. Results in Fig. 7.
3. *Misaligned Monte Carlo spectra with background*, containing a non–zero baseline (random Gaussian addition with mean of 5 and sigma of 1.6), i.e. the most realistic simulation of data that our Monte Carlo can produce, requiring all preprocessing steps to be applied. Results in Fig. 8.
4. *RELAX air calibartion data*, containing genuine spectra, used for comparing calibration results against a conventional calibration method. Results in Fig. 10 and 11

Our preprocessing methods were applied to each dataset, in various combinations. As our methods are predominatly driven by the properties of the data, there is only 1 genuine free parameter, that being the width, w, of the baseline correction smoothing Gaussian kernel. A width, SD of 40 bins in the $\frac{m}{z}$ axis was used by default in experiments. Additionally, the results of an experiment varying the smoothing SD with the best performing combination of preprocessing steps is shown in Fig. 6.

The data was also processed using SpecAlign, with results in Fig. 9. Baseline subtraction in SpecAlign is performed by using a line–picking algorithm, whereby the user defines a $\frac{m}{z}$ window, and values beneath a dynamic local–average are utilised to deduce a global moving–average, in order to ascertain the baseline function. Various method are provided in SpecAlign for alignment of mass spectra: Peak alignment by FFT (PAFFT) correlation, and recursive alignment by FFT (RAFFT) correlation; the former relies upon a FFT and a segmentation model utilising equally–sized segments, the latter upon a FFT utilising a recursive segmentation model from global to local–level, in order to refine the alignment. Each of these three methods allow the order of the polynomial interpolation to be assigned as either 1 (linear), 2 (square) or 3 (cubic). Details are described elsewhere.[33, 34, 43]

The primary focus of this current work is that of assessing the properties of raw and pre-processed data, and therefore does not extend to any detailed investigation or explination of any subsequent analysis that may be undertaken using the pre-processed outputs. However, to show applicability of the pre-processing, an air calibation is performed with the help of a statistical modelling method which requires independent Poisson behaviour of histograms.

For the air calibation tests, conventional data reduction was performed by subtraction of a mean 'blank' spectrum (for baseline and contamination correction), followed by a manual identification of peaks.[3] Blank spectra were aquired, i.e. spectra with no sample present in the sample chamber, to estimate the spectra of baseline and equipment contaminants. The peak-ratios of manually segmented peaks were then compared to known ratios for the Xenon present in the air calibration samples. The mean and variance of these ratios were estimated by using 15 repeat datasets. In a our alternative analysis, the preprocessing techniques described in this paper were applied in order to obtain Poisson independent noise characteristics. Achieving these statistical properties allowed us to apply an LPM[25] (independent component analysis with an assumption of Poisson noise), to model the contamination within 'blank' spectra. The modelled contaimation was then subtracted from the spectra, and repeated ratio calculation were again computed using the same 15 repeat data sets. The key objective of this particular experiment was to provide an example of how an analysis requiring our 'ideal' spectra properties could be applied. Additional examples of applying LPMs to more complex spectra which were preprocessed with our methods can also be found.[30]

4 Results and Discussion

Overall, our processing pipeline successfully corrects for misalignment of peaks and also raised backgrounds. The criteria for success was defined as being the ability to produce spectra histograms that behave in a way consistent with independent Poisson bins, whilst maintaining the shape and total integral of the original data (i.e. reaching the dotted target lines in the associated figures). In constrast, the benchmark SpecAlign software failed to produce these desirable statistical properties under most conditions. After applying our preprocessing methods to real RELAX data, we were able to apply Linear Poisson Modelling to aid in determining calibration values, which were achieved within levels of up to a quarter of the variance achieved using a more conventional approach.

4.1 Ideal Spectra

With and without applying our preprocessing methods, ideal spectra produced Bland-Altamn error model parameters consistent with predicted ideal values, showing Poisson growth in errors as a function of intensity and no additional upward scaling of errors. Measures of signal shape and total signal integral also produced expected properties consistent with ideal spectra. Finally, there were no significant correlations between adjacent residuals, with unprocessed and integrated peaks both giving correlations consistent with zero. These results show that, a) our Monte Carlo data generator was correctly generating ideal spectra, b) that our preprocessing pipeline did not introduce any unwanted effects, and c) that our six figures of merit are capable of spotting ideal spectra when they occur. Fig. 5 summarises these results.

4.2 Misaligned Spectra

If no preprocessing is applied to misaligned spectra then most figures of merit fail to reach target values. The lack of alignment pushes the error scaling to such a high level (greater than 70) that the data point is omitted from Fig. 7. The power-law growth in errors is also more linear than square-root dependent. Signal shape is adversly affected, with poor χ^2_D observed. Simply integrating the peaks improves the statistical properties, as the reduction in binning mitigates against the shifting peaks.

When whole-bin and sub-bin alignment is applied, results are similar to one another, with an overall improvement in shape and integral preservation and reduction in error correlations. However, Unless spectra are aligned with sub-bin precision, there are significant scalings of errors observed. This can be attributed to the added variability of ions at bin boundaries being randomly counted in alternative bins.

These results show that a) when spectra are misaligned, preprocessing is essential, b) integrating peaks alone can solve many of the problems with the statistical properties, c) that for overall best results, sub-bin alignment should be performed (with integration of peaks performed afterwards if desired).

4.3 Misaligned Spectra with Background

When spectra are misaligned and have additional non-zero baseline, then applying no processing produces very poor statistical properties (points A in Fig. 8). When background is present, applying only peak integration also results in poor spectra, with biased total integral and larger than expected spread of total integral (points B). The ideal combination of preprocessing involves applying all three steps: sub-bin alignment, baseline correction and peak integration (points D). However, the addition of Gaussian noise superimposed upon the Possion signal still moves error scaling slightly away from unity.

These results show that a) baselines clearly bias signal (which is perhaps obvious) and also make total signal measurements less repeatable, b) that the additional Gaussian noise, on top of Poisson signal, changes the overall error model, so a pure Poisson assumption on real data should be made with causion, and c) that producing close to ideal spectra requires a complete preprocessing pipeline.

4.4 Background Smoothing Parameter

The optimal smoothing width, w, will be dataset dependent, and it is reasonable to assume that the optimal smoothing width is a function of peak width. A given replicated spectrum consists of peaks of different widths (see Table 1). Taking the best performing combination of processing (FT align, followed by baseline correction, then peak integration), and scanning over the smoothing parameter shows the optimum smoothing SD to be between 40

and 50 bins, i.e. ≈2.5× the average peak width. There is a clear minimum at this point in the BA plot parameters (Fig. 6).

4.5 Preprocessing using SpecAlign

Spectra that were misaligned with additional background were additionally tested against benchmark software. The standard SpecAlign baseline correction was applied, followed by all combinations of peak alignment, including the most basic peak matching, and more complex interpolation schemes, with results shown in Fig. 9. Overall, statistical properties atainable with SpecAlign fall short of our criteria for ideal spectra.

The interpolation schemes (FT, PAFT and RAFT), did produce the correct Poisson power-law behaviour, i.e. errors grew with square-root of intensity. Also, the overall signal shape, as assessed using χ^2_D, approaches a respectable level of 2 (in comparision to the no-processing equivelent of point A in Fig. 8). However, the repeatability of total integral (Pull distribution SD) appears to improve (i.e. smaller spread), at the expense of introducing systematic bias (Pull distribution mean). Larger correlations are also observed in comparision to our own methods.

In contrast, the simpler peak matching (PM) method did not bias the total integral, and gave improved bin independence, yet significantly failed other tests.

These results show that a) polynomial interpolation schemes add statistical instabilities, b) poorly designed preprocessing methods can trade off statistical repeatability with systematic errors, giving false impressions regarding levels of accuracy (as in real data, only the stability of a signal is observed if no knowledge of true mean behaviour is avaiable, i.e. bias can be hidden). Finally, simple peak matching methods behave much differently with regards to statistical effects than more complex approaches.

4.6 Preprocessing of real data

Statistical analysis methods typically make assumptions regarding the properties of data, for example, least-square methods assume Gaussian residuals, PCA assumes orthogonal modes of variation, and KS-tests assume Poisson noise within histograms. The analysis of a data set will be most efficient when the properties of that data are a good match to the assumptions made by the analysis approach. Similarly, methods applied to mass spectra will be more efficient if the properties of the spectra are well matched to the assumptions made. The application of LPMs, for instance, is best applied to histograms with independent Poisson bins. Our air calibration demonstrates this point.

Fig.10 and 11 allow us to compare the results of using our preprocessing methods to provide calibration values for Xenon in standard air. Achieving ideal spectra was key to the application of Linear Poisson Models (an independent component analysis for histograms) for modelling hydrocarbon contamination. The ability to model contamination supports a factor of approximately two improvement in accuracy of xenon peak measurement over the original method for the same dataset. The new estimates still conform to the behaviour expected for the equipment, but with a closer conformity, consistent with the predicted errors. The observed decrease in variance of calibration points is equivalent to having increased the original sample volume by a factor of four. Given that the main challenge with using xenon spectra in scientific analyses is the need to work with small samples, this represents a significant improvement in performance.

The observed variations were computed using 15 attempts at end-to-end calibration, including the process of extrapolating isotope ratios back to time zero, so that all sources of possible error were present. The level of calibration precision attainable is limited by the Poisson sampling of ions detected. There are $\approx 100,000$ atoms in an air calibration sample, with a $\approx 50\%$ detection rate. Using the total quantity of data (i.e. 15 repeat air cal. spectra), and knowledge of the relative quantities within each peak (Fig. 3), we find that our observed variances (Fig. 11) are consistent with reaching this limit.

The results show that a) our preprocessing can be a valuable tool in producing histograms with properties suitable for subsequent analysis by methods that rely upon independent Poisson histograms, and b) the independent Poisson behaviour attainable, coupled with the conservation of signal shape and total integral, permits statistical modelling to be performed to improve the calibation of RELAX data.

5 Conclusions

A great many algorithms are assessed using simple metrics, such as signal-to-noise, or true/false detection rates on such things as peak identifications. We have shown that a range of alternative figures of merit also have value, and

can identify problems with statistical properties of data that could otherwise be hidden. For example, χ^2 statistics can be used to compare actual verses expected distributions of spectra, which only give expected values if all peaks are present and are the correct shape, whereas a simple detection rate would only determine how many peaks were detected. In addition, the use of Pull distributions, Bland-Altman and correlation analyses can successfully quantify the behaviour of noise and bias, providing information about errors as a function of signal, and assess the effects of noise upon adjacent bins.

Our pre-processing approach can successfully correct spectra that has been misaligned and have non-zero baselines, in order to achieve idealised histograms. The subsequent analysis of spectra by methods that assume idependent Poisson histogram behaviour can therefore be used with greater confidence. In constrast, methods that were not designed with such properties in mind (e.g. SpecAlign) may superficially improve the alignment and baselines of spectra, but may compromise other properties. By following our quantitative approach, we have shown that calibration curves for air samples can be estimated with higher precision than previously obtained.[3]. The methods described herein can also be applicable to mass spectra acquired using other instrumentation e.g. MALDI (matrix–assisted laser desorption/ionisation), and different conditions e.g. laser intensity or different and more complicated spectra [30], but results must be assessed on a case-by-case basis.

Acknowledgements

Financial support is gratefully acknowledged from the Leverhulme Trust (grant no. RPG–2014–019) and the Science and Technology Facilities Council (STFC) (grant no. ST/M001253/1). The authors thank Dr Fiona Henderson and Dr Adam McMahon (both of the University of Manchester) for their kind instruction, guidance and assitance with mass spectrometry instrumentation, and Dr John Cowpe and Dr Sarah A Crowther (both of the University of Manchester) for acquiring RELAX data and maintaining the RELAX instrument.

References

[1] G. L. Glish and R. W. Vachet, *Nat. Rev. Drug Discov.*, 2003, **2,** 140–150.

[2] E. de Hoffmann and V. Stroobant, Mass Spectrometry: Principles and Applications, Third Edition, 2013. John Wiley and Sons.

[3] S. A. Crowther, R. K. Mohapatra, G. Turner, D. J. Blagburn, K. Kehm and J. D. Gilmour, *J. Anal. Atom Spectrom.*, 2008, **23,** 938–947.

[4] J. Hennig, K. D. M. Hennig and M. Sunnerhagen, *Bioinformatics*, 2008, **24,** 1310–1312.

[5] L. O. W. Wilson, A. Spriggs, J. M. Taylor and A. M. Fahrer, *Bioinformatics*, 2014, **30,** 151–156.

[6] I. Strashnov and J. Gilmour, *Hyperfine Interact.*, 2014, **227,** 259–270.

[7] M. Guilhaus, *J. Mass Spectrom.*, 1995, **30,** 1519–1532.

[8] T. Henkel, J. D. Gilmour, H. D. Holland (Ed.) and K. K. Turekian (Ed.), Treatise on Geochemistry (Second Edition), 2014, Oxford: Elsevier, 411–424.

[9] I. Strashnov, D. J. Blagburn and J. D. Gilmour, *J. Anal. At. Spectrom.*, 2011, **26,** 1763–1772.

[10] Y. C. Harn, M. J. Powers, E. A. Shank and V. Jojic, *Bioinformatics*, 2015, **31,** 142–50.

[11] P. D. Piehowski *et al.*, *Anal. Chem.*, 2009, **81,** 5593–5602.

[12] H. Shin, M. Mutlu, J. M. Koomen and M. K. Markey, *Cancer Inform.*, 2007, **3,** 219–230.

[13] X. Kong and C. Reilly, *Bioinformatics*, 2009, **25,** 3213–20.

[14] J. S. Morris, K. R. Coombes, J. Koomen, K. A. Baggerly and R. Kobayashi, *Bioinformatics*, 2005, **21,** 1764–75.

[15] A. Antoniadis, J. Bigot and S. Lambert–Lacroix, *J. Soc. Fr. Statistique*, 2010, **151,** 17–37.

[16] S. Gibb MALDIquant Quantitative Analysis of Mass Spectrometry Data.

[17] N. Jeffries, *Bioinformatics*, 2005, **21,** 3066–3073.

[18] W. Yu *et al.*, *Comp. Biol. Chem.*, 2006, **30,** 27–38.

[19] C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan and G. Siuzdak, *Anal. Chem.*, 2006, **78,** 779–787 .

[20] J. T. Halloran, J. A. Bilmes and W. S. Noble, *30th Conference on Uncertainty in Artificial Intelligence, UAI 2014; Quebec City; Canada; 23 July 2014 through 27 July 2014*, 2014, **78,** 320–329.

[21] R. L. Somorjai, B. Dolenko and R. Baumgartner, *Bioinformatics*, 2003, **19,** 1484–1491.

[22] D. Schneidman–Duhovny *et al.*, *Bioinformatics*, 2012, **28,** 3282–3289.

[23] Y. Yasui *et al.*, Profiling High–Dimensional Protein Expression Using MALDI–ToF Mass Spectrometry For Biomarker Discovery, 2006. Ed. by J. Crowley and D. P. Ankerst DP, Handbook of Statistics in Clinical Oncology 2. New York: Chapman–Hall/CRC.

[24] R. J. Barlow, 1999. Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences. John Wiley and Sons.

[25] P. D. Tar, N. A. Thacker, J. D. Gilmour and M. A. Jones, *Adv. Space Res.*, 2015, **56,** 92–105.

[26] P. D. Tar, R. Bugiolacchi, N. A. Thacker, J. D. Gilmour and MoonZoo Team, *Earth Moon Planets*, 2017, **119,** 47–63.

[27] P. D. Tar and N. A. Thacker, *Annals of the BMVA*, 2014, **1,** 1–22.

[28] H. N. B. Moseley, *Comput. Struct. Biotechnol. J.*, 2013, **4,** 1–12.

[29] LGC Limited 2003. Preparation of Calibration Curves – A Guide to Best Practice. LGC/VAM/2003/032.

[30] S. Deepaisarn, P. D. Tar, N. A. Thacker, A. Seepujak and A. Mcmahon. Linear Poisson Independent Component Analysis: A new method for quantifying biological mass spectrometry data. Submitted to *Bioinformatics*, 2017.

[31] I. Shah, A. Petroczi and D. P. Naughton, *Chem. Cent. J.*, 2012, **6**.

[32] J. M. Bland and D. G. Altman, *Lancet*, 1986, **327,** 307–310.

[33] J. W. H. Wong, G. Cagney and H. M. Cartwright, *Bioinformatics*, 2005, **21,** 2088–2090.

[34] J. W. H. Wong, C. Durante and H. M. Cartwright, *Anal. Chem.*, 2005, **77,** 5655–5661.

[35] G. W. Snedecor and W. G. Cochran, 1989. Statistical Methods, Eighth Edition, Iowa State University Press.

[36] T. P. Pridmore, *Lecture Notes in Computer Science*, 2002, **2390,** 310–319.

[37] F. J. Anscombe, *Biometrika*, 1948, **35,** 246–254.

[38] Advances in Signal Transforms: Theory and Applications, 2007. Ed. by J. Astola and L. Yaroslavsky, Ch. 8, Hindawi Publishing Corp. (New York).

[39] D. W. Marquardt, *SIAM J. Appl. Math.*, 1963, **11,** 431–441.

[40] W. H. Press, B. P. Flannery, S. A. Teukolsky and W. T. Vetterling, 1991. Numerical Recipes In C, Cambridge University Press.

[41] N. Bandeira, J. V. Olsen, M. Mann and P. A. Pevzner, *Bioinformatics*, 2008, **24,** i416–i423.

[42] J. D. Gilmour, G. Holland, A. B. Verchovsky, A. V. Fisenko, S. Crowther and G. Turner, *Geochim. Cosmochim. Ac.*, 2016, **177,** 78–93.

[43] T. N. Vu and K. Laukens, *Metabolites*, 2013, **3,** 259–276.

[44] D. Luc and L. Lyons, *CDF note*, 2002, **43**.

[45] *http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics/Peak–Alignment/.*

[46] J. A. Nelder and R. Mead, *Comput. J.*, 1965, **7,** 308–313.

[47] D. G. Krige. A Statistical Approach to Some Mine Valuatio ns and Allied Problems at the Witwatersrand. MSc thesis, University of Witwatersrand (1951).

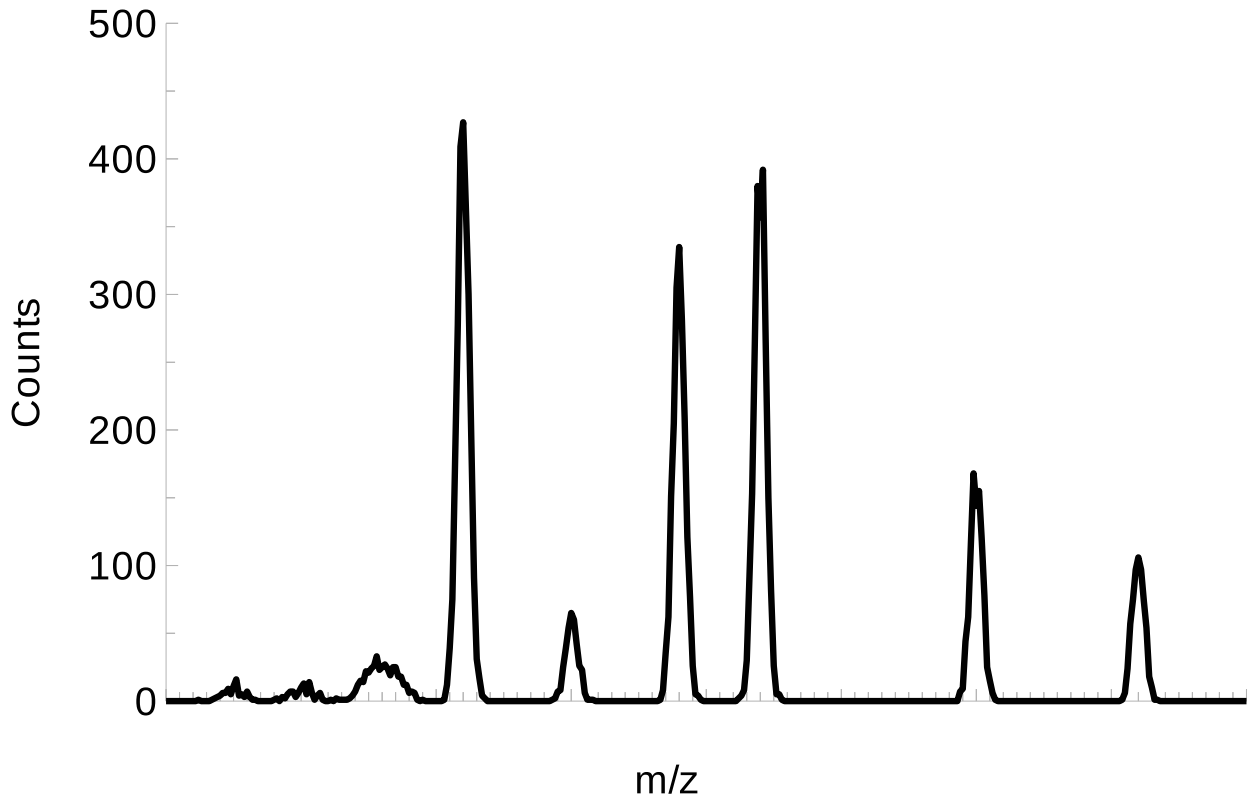


Figure 1: Example of a synthetic air calibration spectrum generated using Monte Carlo simulations. Peaks have a Gaussian profile, with bin frequencies varying with Poisson noise. Peaks correspond to isotopes ^{124}Xe to ^{136}Xe.

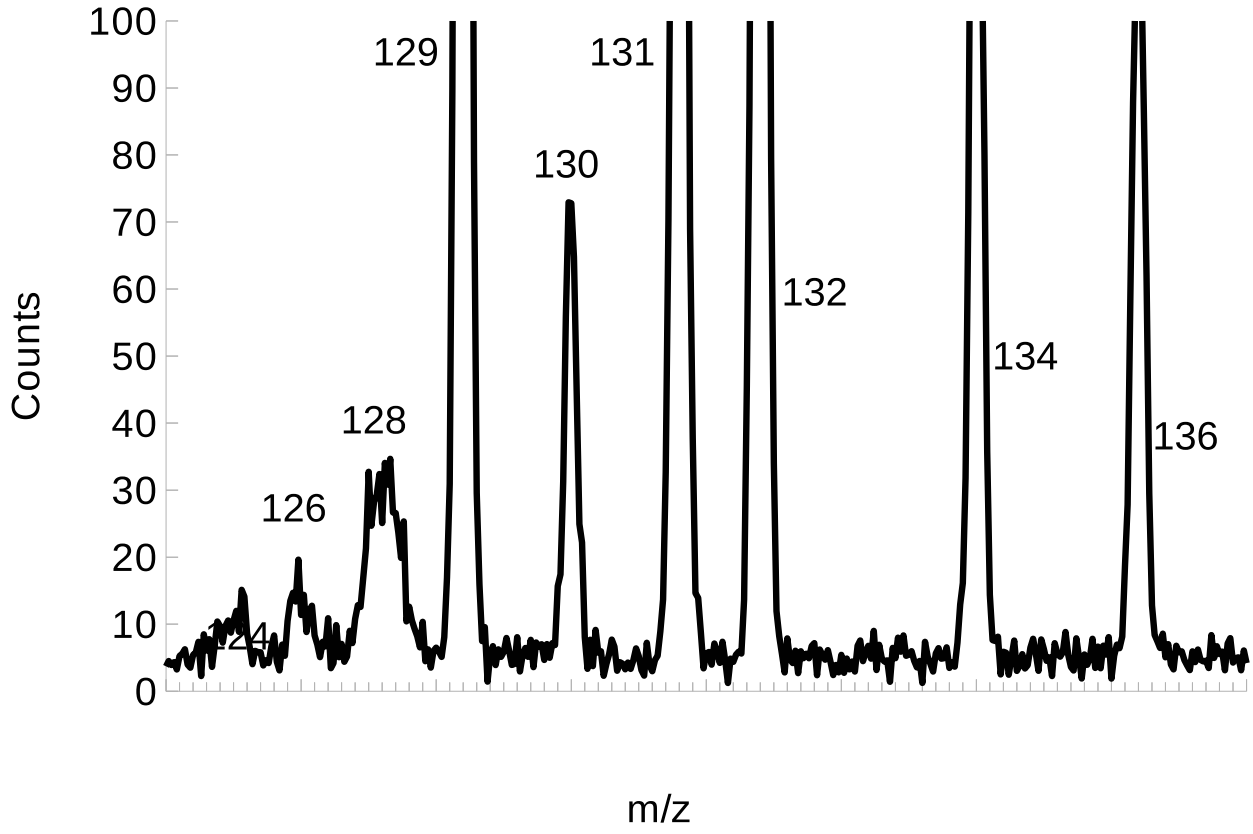


Figure 2: Example of simulated raised baseline using non–zero mean Gaussian noise background. Spread of the between-peak noise and level above zero is set to match that observed in real data.

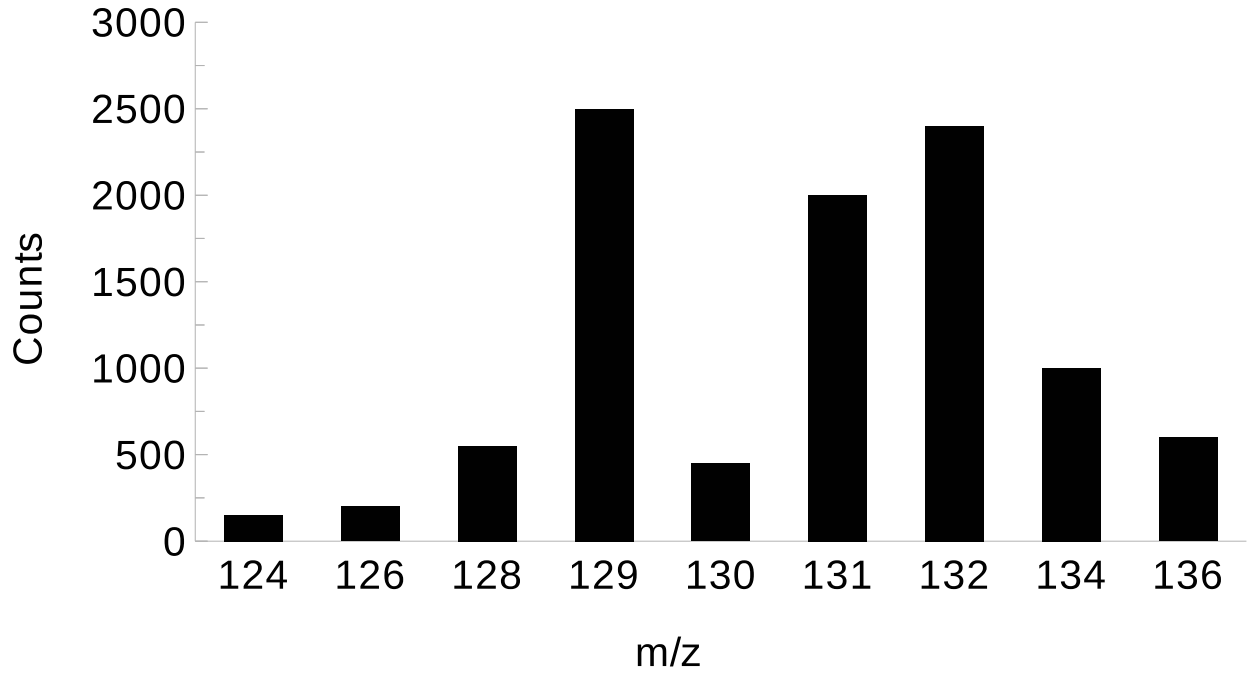
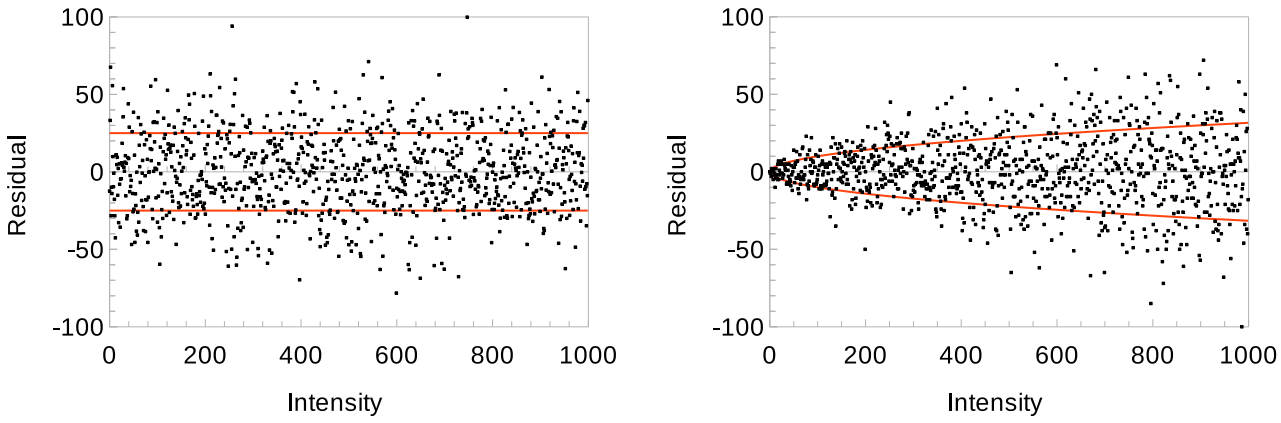


Figure 3: The results of applying the peak integration tool. Here, each peak in the spectrum has been identified and integrated into its own histogram bin.



(a) Gaussian distribution

(b) Poisson distribution

Figure 4: Bland–Altman (BA) plots for commonly–occurring distributions. For each distribution, a line of the mean residual value versus the intensity is shown. For the Gaussian function, the mean residual value is uniform with intensity; a Poisson distribution has square-root growth in errors with intensity.

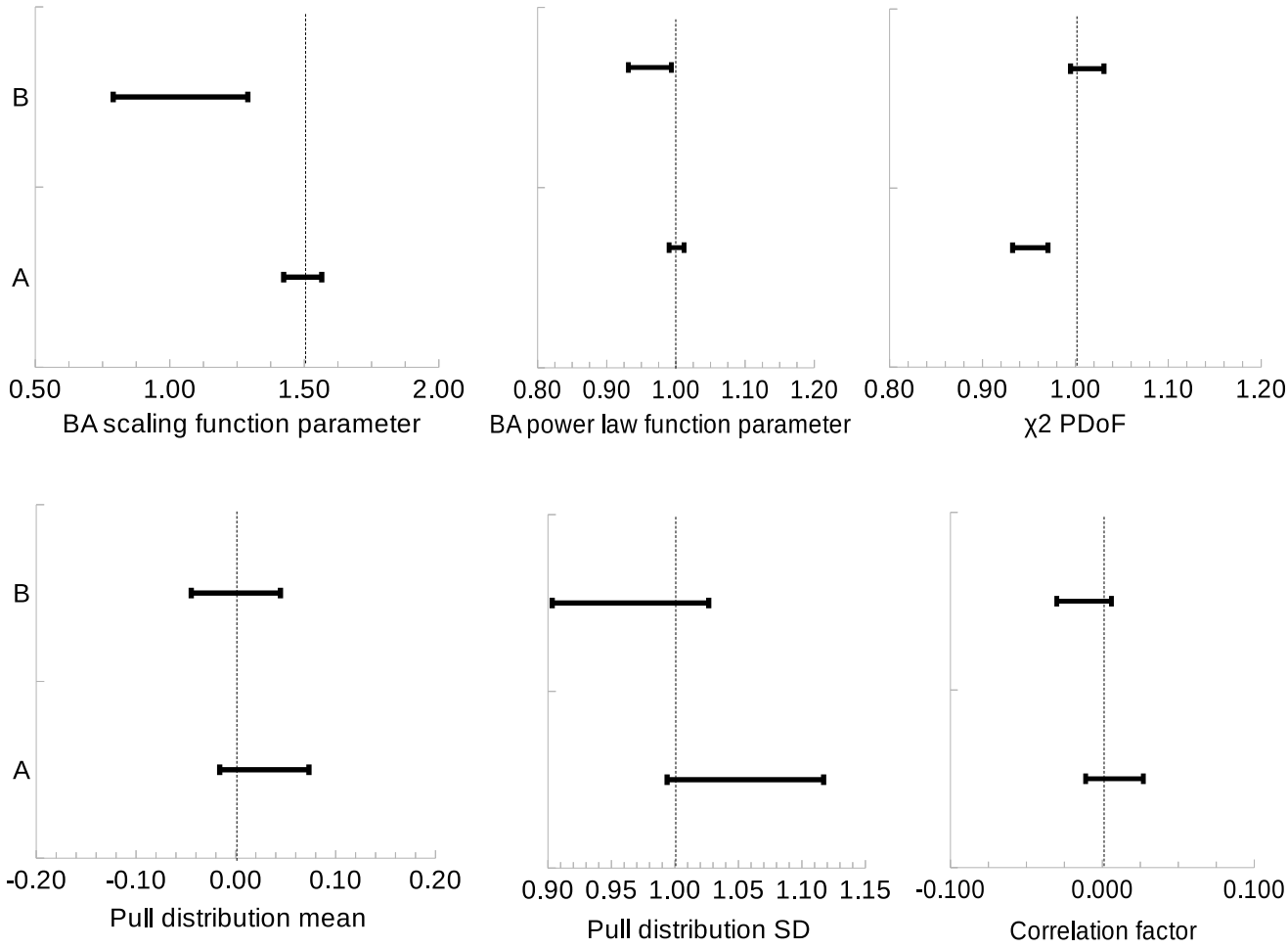
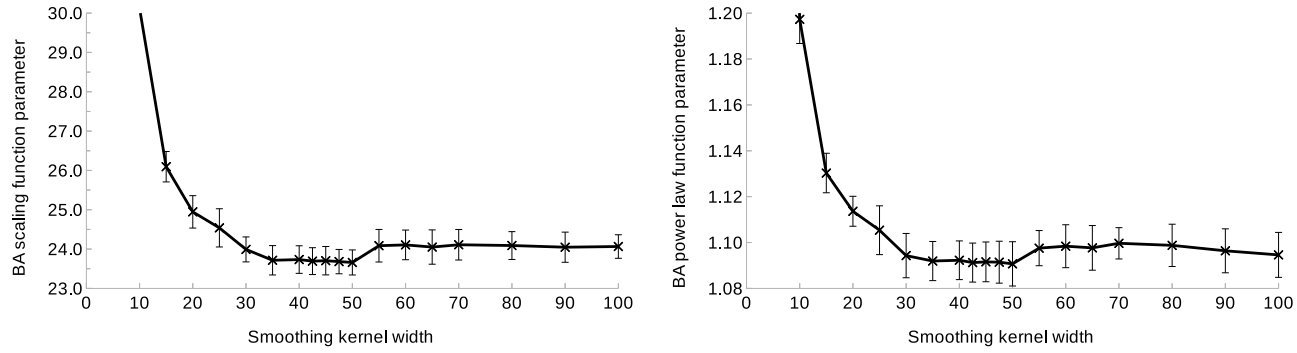


Figure 5: Ideal Monte Carlo spectra results. Bland–Altman (BA) plot error parameters, χ^2 PDoF (the χ^2_n), Pull distribution parameters and correlations. All values are dimensionless. Methods: (A) No processing; (B) Integrate peaks. Ideal parameter values are shown using dashed lines. These results thus confirm the expected properties of the ideal Monte Carlo spectra.



(a) (b)

Figure 6: Variation of (a) the BA scaling function parameter, and (b) the BA power –law function, as a function of smoothing level on baseline correction. Data used were real RELAX spectra. A minimum is seen at 40, after which both parameters continue to rise again with a kernel smoothing width of greater that 50. For each plot, the mean of the plateau region is greater than the mean of the minimum region.

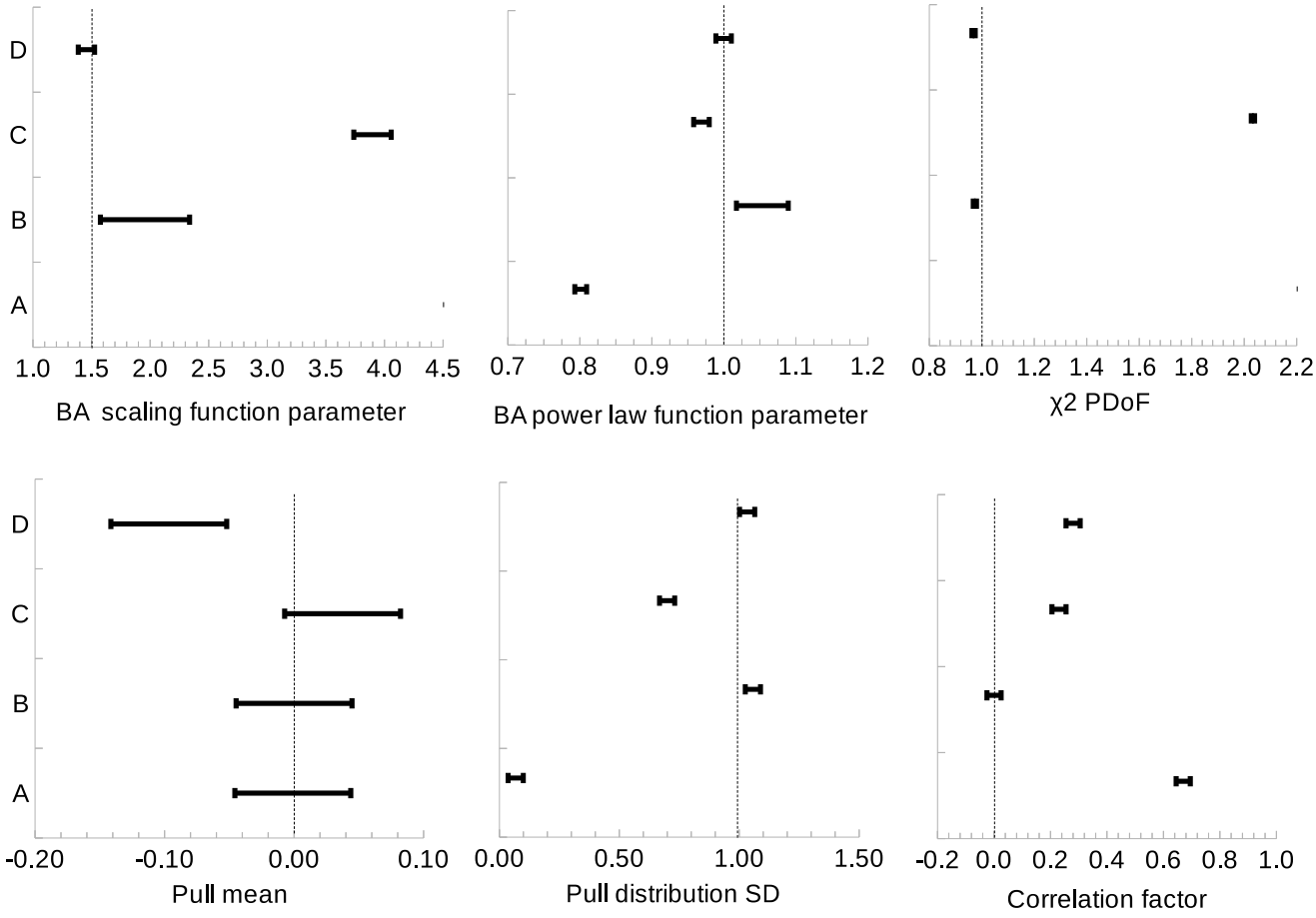


Figure 7: Misaligned Monte Carlo spectra results. BA plot error parameters, χ^2 PDoF (the χ^2_n), Pull distribution parameters and correlations. Methods: (A) No processing; (B) Integrated peaks; (C) Whole–bin alignment; (D) Sub–bin alignment. Ideal parameter values are shown using dashed lines. Failure of procesing method A has resulted in outlier values (not shown) for the BA scaling function parameter and χ^2 PDoF.

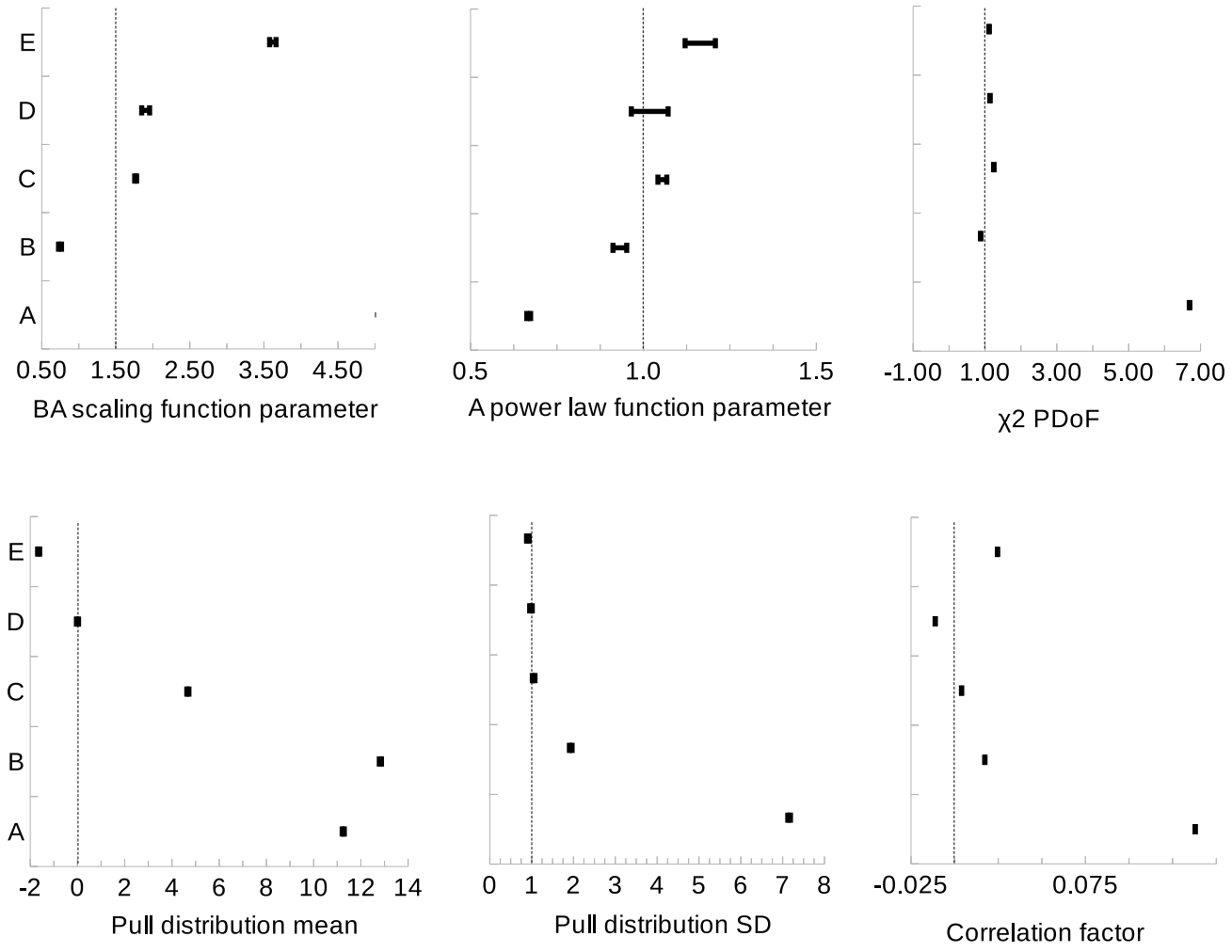


Figure 8: Misaligned Monte Carlo spectra with added background. Bland–Altman (BA) plot error parameters, χ^2 PDoF (the χ^2 per degree of freedom), Pull distribution parameters and correlations. All values are dimensionless. Methods: (A) No processing; (B) No alignment, no baseline correction and integrated peaks; (C) Sub–bin alignment and baseline corrected; (D) Sub–bin alignment and baseline corrected and integrated peaks; (E) No alignment, baseline correction and integrated peaks . Ideal parameter values are shown using dashed lines. Failure of processing method A has resulted in an outlier value (not shown) for the BA power–law function parameter.

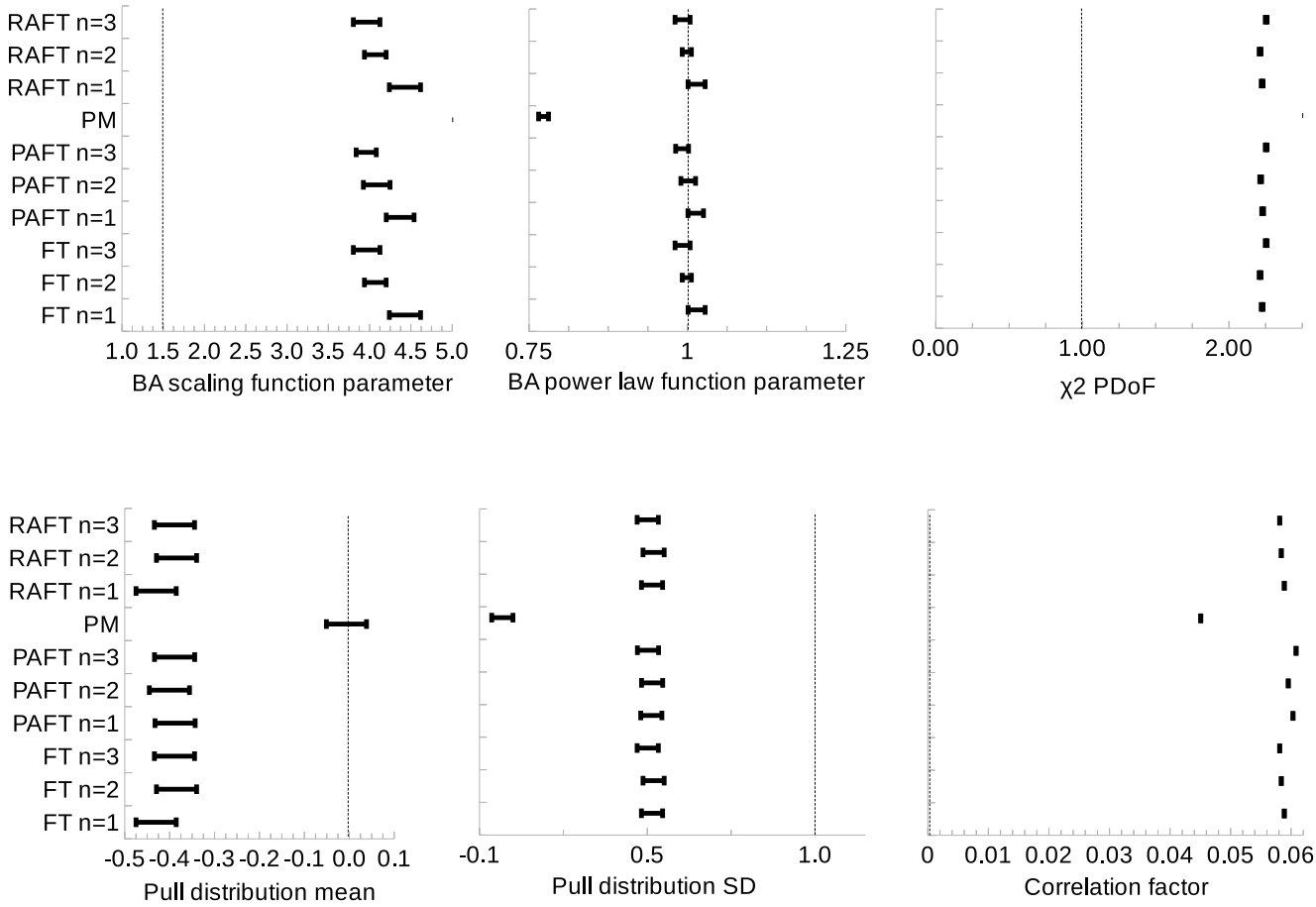


Figure 9: Bland–Altman (BA) plot error parameters, χ^2 PDoF (the χ^2_n), Pull distribution parameters and correlations for misaligned Monte Carlo spectra with background, after preprocessing using SpecAlign. The SpecAlign methods: Fourier transform (FT), peak alignment by Fourier transform (PAFT), peak matching (PM) and recursive alignment by Fourier transform (RAFT). The n refers the the polynomial order of the sub-bin interpolation function. Ideal parameter values are shown using dashed lines. Failure of the PM and PAFT (n=1) processing methods has resulted in outlier values (not shown) for the BA scaling function parameter, the χ^2 PDoF and correlation factor.

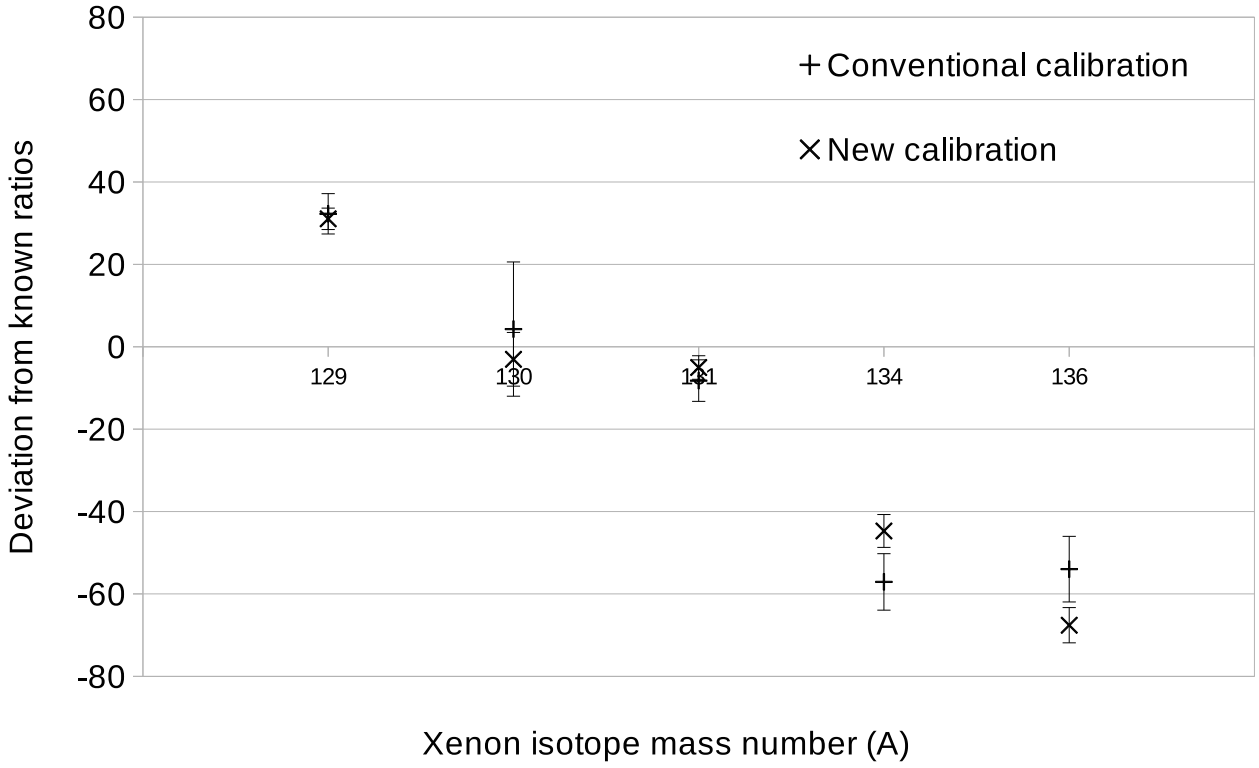


Figure 10: A comparison of deviation (bars) from a standard (crosses, representing an air mass spectrum), for a range of masses. Deviations are shown in parts per 1000. Real data is used, of the form used in the Monte Carlo replicated spectra.

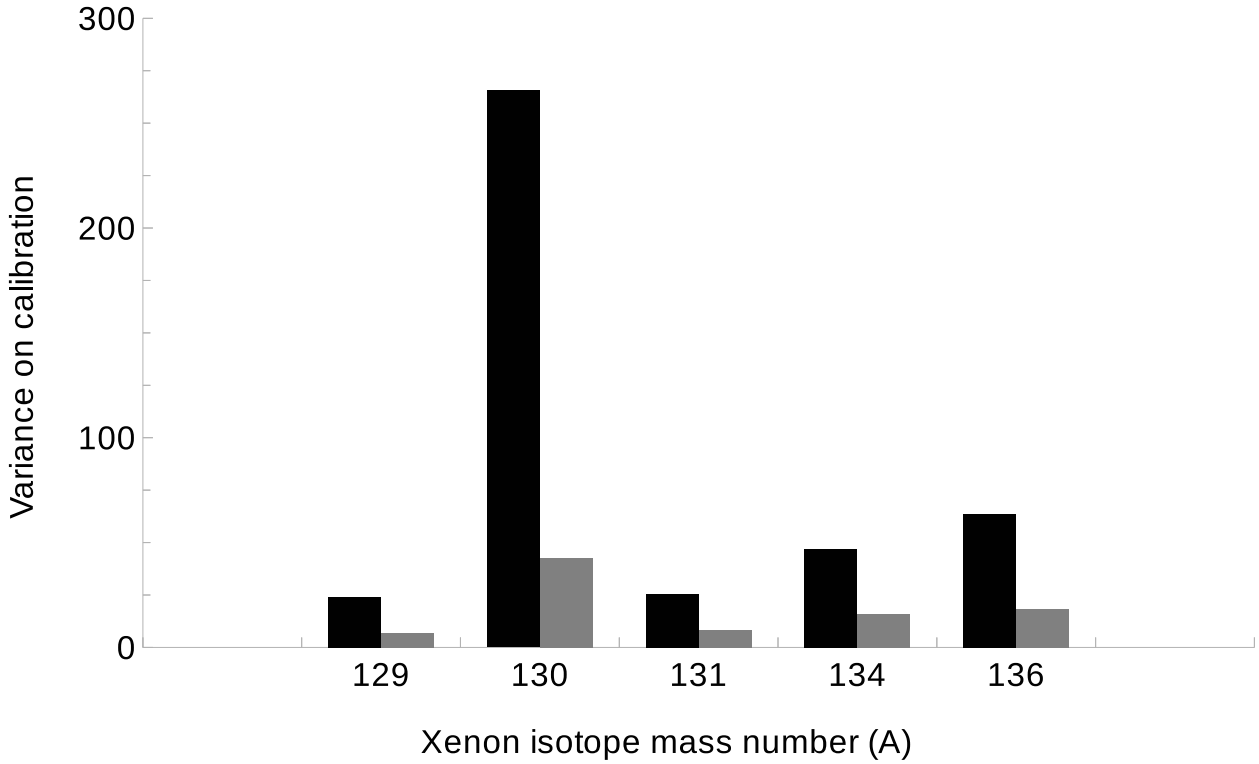


Figure 11: Variance of calibration points in parts per 1000. Comparison of variance (i.e. error bars from Fig.10) of calibration points for a range of masses, using the conventional and the new calibration methods. On average across the data points analysed, the new variances are equivalent to a factor of four increase in data quantity. (Black bars) Conventional calibration variance; (grey bars) new calibration variance.