ACS Editors' Choice

analytical chemistry

Article

pubs.acs.org/ac

Derivation from First Principles of the Statistical Distribution of the Mass Peak Intensities of MS Data

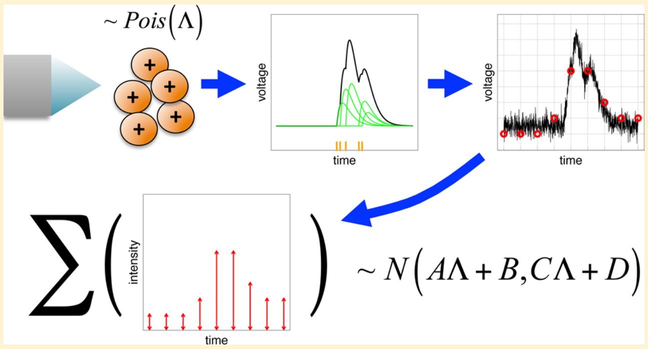
Andreas Ipsen*

Institute of Mass Spectrometry, College of Medicine, Swansea University, Swansea SA2 8PP, United Kingdom

Biological Sciences Division and Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, 3335 Innovation Ave. (K8-98), P.O. Box 999, Richland, Washington 99352, United States

Supporting Information

ABSTRACT: Despite the widespread use of mass spectrometry (MS) in a broad range of disciplines, the nature of MS data remains very poorly understood, and this places important constraints on the quality of MS data analysis as well as on the effectiveness of MS instrument design. In the following, a procedure for calculating the statistical distribution of the mass peak intensity for MS instruments that use analog-to-digital converters (ADCs) and electron multipliers is presented. It is demonstrated that the physical processes underlying the data-generation process, from the generation of the ions to the signal induced at the detector, and on to the digitization of the resulting voltage pulse, result in data that can be well-approximated by a Gaussian distribution whose mean and variance are determined by physically meaningful instrumental parameters. This allows for a very precise understanding of the signal-to-noise ratio of mass peak intensities and suggests novel ways of improving it. Moreover, it is a prerequisite for being able to address virtually all data analytical problems in downstream analyses in a statistically rigorous manner. The model is validated with experimental data.



Because several of the physical processes that are involved in the generation of MS data are essentially random in nature, the output data may be regarded as a (high-dimensional) random variable. An elementary first step in making sense of any random variable is to determine its probability distribution. But determining the probability distribution of MS data is a challenging task and it is noteworthy that despite the widespread use of the technology, it is one that apparently has not yet been fully accomplished for any instrument. In the most demanding scenario, a detailed mathematical model of each stage of the mass spectrometer is required, of the ionization process, ion optics, signal detection and amplification, and digitization. However, as will be demonstrated, various mathematical approximations may be employed to significantly simplify such modeling tasks, and through careful application of the central limit theorem (CLT hereafter), the highly complex distribution of the mass peak intensity may be well-approximated by a Gaussian distribution.

Although numerous noise models have been proposed for MS data, these generally employ heuristic models for the intensity distribution that do not attempt to account in detail for the effects of the instrumental architecture.[1−7] Two important exceptions to this are an application note by Gedcke[8] and a study by Harris et al.,[9] both of which model the uncertainty associated with ADC-digitized data acquired via mass spectrometers incorporating electron multipliers. Although both of these studies do discuss the instrumental architecture in substantial detail, neither provides a rigorous derivation of the statistical distribution of the data, as the models developed nevertheless leave out fundamental features of the data generation process. A rather different approach to noise modeling is provided by Shin et al.[10] who use parametric spectral density estimation to analyze time-of-flight MS data, although again without explicitly accounting for the effects of the various components of the instrument's data acquisition system.

For mass spectrometers that are based on ion counting, the data generation process is somewhat simpler and the models of the output data are arguably more mature.[11−13] For such instruments, models that take the more straightforward approach of treating the recorded ion count as being entirely Poissonian can also be used[14−16] provided the rate of ion arrivals is sufficiently low that detector saturation is negligible. However, ion counting mass spectrometers are currently less widely used, in part due to their more limited dynamic range.

The following study provides a detailed derivation and validation of the statistical distribution of the mass peak intensity (defined as the sum of the intensities recorded across

Received: September 22, 2014
Accepted: December 28, 2014
Published: January 24, 2015

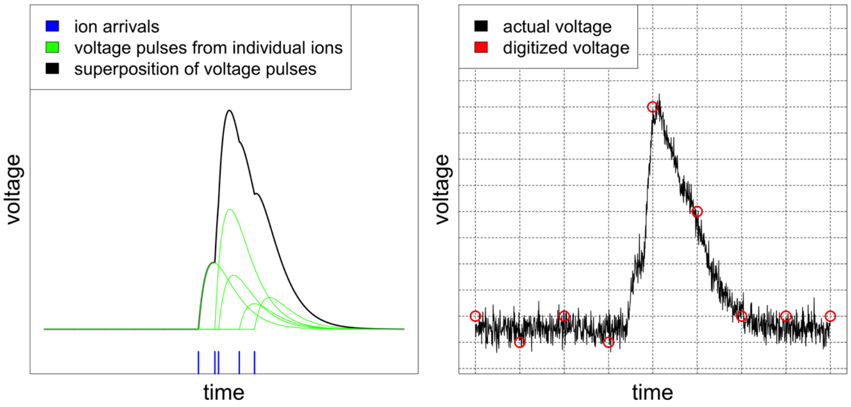


Figure 1. Simulated data illustrating the generation and digitization of the voltage signal induced over the course of a mass peak with 5 ion arrivals. On the left, the individual voltage pulses of each of the incoming ions are shown, along with their superposition. On the right, where the electronic noise is also included, the digitizer's sampling points and voltage discretization levels are indicated by the vertical and horizontal lines of the grid respectively, and the digitized voltages are indicated by the red circles.

a mass peak) of ADC-based mass spectrometers that detect ions via an electron multiplier. The model is therefore potentially applicable to a broad range of modern time-of-flight (TOF) instruments, as well as to many sector and quadrupole instruments, though not to Fourier transform ion cyclotron resonance (FTICR) mass spectrometers or Orbitraps, which employ image current detection. Although this study does not determine the full probability distribution of MS data, it does demonstrate the feasibility of the enterprise and it helps formalize the mathematical treatment of fundamental aspects of MS data that have not previously received much attention in the MS literature.

■ THEORY

In this section, a summary of the physical processes involved in the generation of mass peak intensities is provided. This is followed by the statistical modeling of each of the steps in this process.

System Summary. A random effect that affects all MS platforms is the fluctuating number of ions produced in the source. It is well-known that this process is governed by the Poisson distribution, so that if Λ is the average number of ions produced over a period of time, then the probability of k ions being produced is

$$P(k) = \frac{\Lambda^k e^{-\Lambda}}{k!} \qquad (1)$$

Some of these ions are likely to be lost prior to reaching the detector; however, the number of ions striking the detector will still be governed by the Poisson distribution, only with a reduced rate Λ.[13]

Upon striking the detector (which is some form of electron multiplier, such as a microchannel plate or a discrete dynode multiplier), an ion sets off a cascade of electrons that results in the initial electric charge being amplified by many orders of magnitude. The average factor by which the signal is amplified is called the gain of the multiplier and it can be adjusted by altering the electric potential across the multiplier. However, large deviations from this average value are common, as the precise number of electrons released is governed by the so-called "pulse height distribution",[17,18] which adds an additional stochastic element to the system. The electron cascade resulting from multiple near-simultaneous ion arrivals can typically be regarded as a superposition of multiple outcomes of the pulse height distribution, but nonlinear effects can become pronounced if very high voltages are used or if the number of ion arrivals is very high.

The signal output by the electron multiplier is further amplified by a preamplifier, which also helps shape the resulting voltage pulse in the time dimension. Careful impedance matching is required at this point to minimize signal reflection. Both the gain of the preamplifier and the shape of the resulting voltage signal tend to be very stable, provided the current signal is within the standard nonsaturated operating range. The effects of the preamplifier may therefore be regarded as deterministic, aside from the electronic noise, whose nature and severity depends on a number of factors, including the preamplifier circuit design, the temperature, the quality of the electromagnetic shielding used, and the presence of unwanted ground loops.

The voltage output by the preamplifier is passed to an ADC, which digitizes it, in some cases after additional pulse shaping. For modern TOF mass spectrometers, 8-bit Flash ADCs are normally used that are capable of discretizing the observed voltage signal into one of 256 levels (typically assigned the values 0−255) by comparing the observed voltage signal with 255 reference voltages. In the present study, it is assumed that the voltage signal passed to the ADC remains within the range of the reference voltages, and that these are evenly spaced. The digitization of the voltage signal is also impacted by the finite time resolution of the digitizer, as indicated in Figure 1 where the origin of the analog voltage pulse and its eventual digitization is illustrated. It should be noted that because a digitizer requires a finite time interval to sample from a voltage signal, all of the voltage pulses considered in the following should be interpreted as indicating the intensities obtained by initiating the sampling at each possible time point.

To keep the amount of data produced manageable, the ADC does not normally write the digitized spectra directly to disk. Rather, it accumulates a specified number, N_p, of spectra (often ranging from 100s to 1000s) for which the intensities of matching sampling points are summed in memory. The intensities resulting from this "histogramming", which is illustrated in Figure 2, are then written to disk if the data are being acquired in profile mode.

In this study, a further reduction will be made to the data, as it is not the intensities observed at each of the sampling points

DOI: 10.1021/ac503554u
Anal. Chem. 2015, 87, 1726−1734

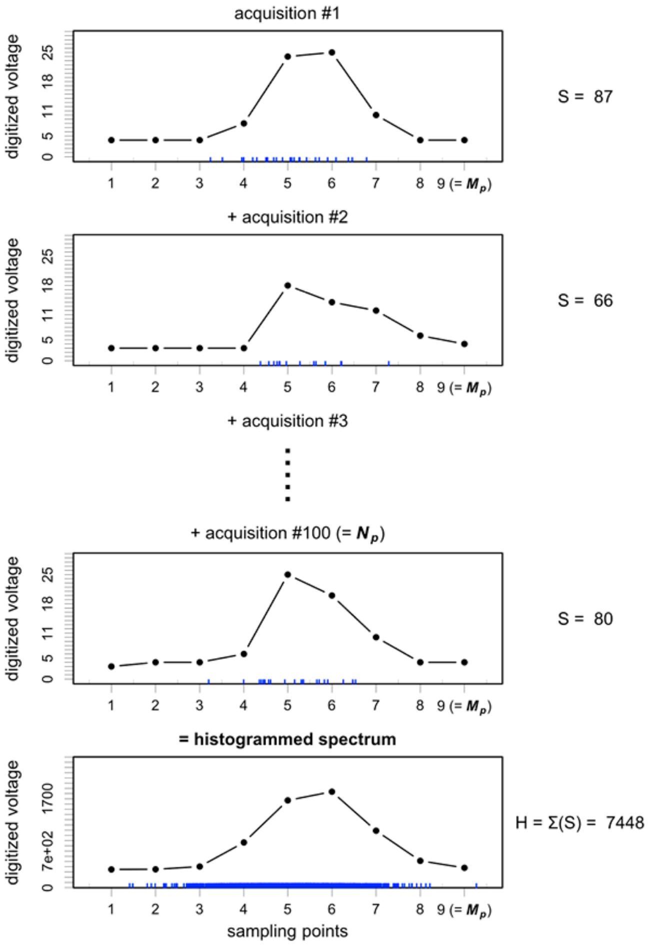


Figure 2. Histogramming of the top N_p = 100 mass peaks that are only recorded to memory, results in the spectrum at the bottom, which is written to disk if the data are acquired in profile mode. The sums of the digitized voltages across the individual mass peaks are listed to the right (the variable S), as is the equivalent sum for the histogrammed mass peak (H).

spanning a mass peak that are being modeled, but rather the sum of these intensities. This corresponds to the intensity that will typically be assigned to a centroided mass peak and this random variable will be referred to as H, while the corresponding random variable obtained from the individual acquisitions will be termed S. Note that if the rate of ion arrivals is constant across the N_p acquisitions, the distribution of H will be equivalent to the distribution of the sum of N_p outcomes of S. The length (in terms of ADC sampling points) of the mass window over which the intensities are summed will be labeled M_p and it should be wide enough to fully contain the mass peak. The mass peak under consideration may be due to a single isotopic species, or, if these are not fully resolved, may be due to multiple ones provided the pulse height distribution does not vary substantially across them. However, the mass window should not overlap with any outside peaks.

System Modeling. The following derivation of the mass peak intensity distribution makes repeated appeal to the CLT to establish the Gaussian nature of the final distribution. Strong emphasis is therefore placed on determining the mean and variance of the various stochastic effects that affect the distribution of the data. The order in which these effects are discussed is not the order in which they occur, but rather is chosen to simplify the mathematical treatment. For simplicity, we also begin with a highly idealized system wherein (1) all voltage pulses produced by the electron multiplier in response to incoming ions are of the same magnitude, (2) there is exactly one ion arrival per mass peak, (3) there is no electronic noise, (4) the ADC has infinite voltage resolution, and (5) no histogramming is performed. Each of these assumptions will be removed in turn as the model is developed.

Discretization of a Normalized Voltage Pulse. We begin the discussion by considering the intensity distribution caused by the time discretization of a single normalized voltage pulse, whose arrival time, Z, at the digitizer is determined primarily by the ion optics but also by a number of other factors relating to the electronics and to the detector characteristics.[19] The intensity of the voltage pulse will be labeled Y_t and referred to as the time discretization factor. Due to the assumptions of the idealized system considered here, calculating Y_t amounts to adding up the values on the normalized voltage pulse that happen to fall on the digitizer's M_p sampling points. Adjacent sampling points are separated by a time period Δt, which is determined by the time resolution of the digitizer, so that a smaller Δt will result in a larger value for Y_t, as illustrated in Figure 3. The mathematical function that describes the normalized voltage pulse, including any ringing, will be referred to as $f(t)$, and it will by definition have an area of 1.

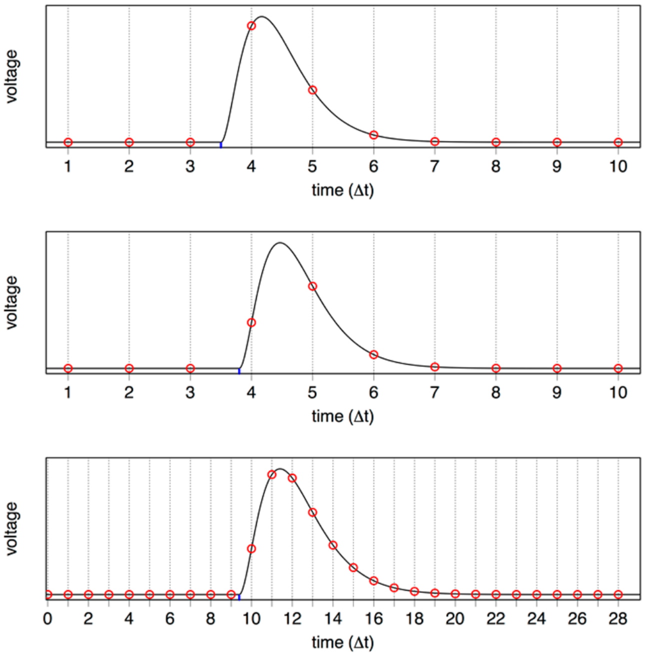


Figure 3. Normalized voltage pulses being sampled by a digitizer with infinite voltage resolution but finite time resolution. The time discretization factor is obtained by adding up the voltages indicated by the red circles. The distinct arrival times of the top two pulses result in different values for Y_t, as do the distinct time resolutions of the digitizers used for the bottom two pulses.

If the normalized voltage pulse is sampled at times t_0, $t_0 + \Delta t$, ..., $t_0 + \Delta t(M_p - 1)$, then Y_t is formally defined as

$$Y_t = \sum_{i=0}^{M_p - 1} f(t_0 + i\Delta t - Z) \tag{2}$$

As demonstrated in the Supporting Information (eqs 3, 4, and 5) where Monte Carlo validation is also provided (Figure

S-1), it can be shown that if the standard deviation of Z is not significantly smaller than Δt, then the mean and variance of Y_t are given by

$$E(Y_t) = \Delta t^{-1} \tag{3}$$

and

$$\begin{aligned} \operatorname{var}(Y_t) &= \Delta t^{-1} \sum_{i=-\infty}^{\infty} \int_{u=0}^{\infty} [f(u)f(u + i\Delta t)]\mathrm{d}u - \Delta t^{-2} \\ &= \Delta t^{-1}\alpha(f, \Delta t) - \Delta t^{-2} \end{aligned} \tag{4}$$

where $\alpha(f, \Delta t)$ is the sum of the autocorrelation of $f(t)$ over all lags that are integer multiples of Δt. The latter quantity may have to be calculated numerically, but this can be done efficiently via the fast Fourier transform. Moreover, because the expression only depends on $f(t)$ and Δt, this only needs to be done once, as those are fixed quantities. Note then that the mean and variance of Y_t are independent of the distribution of Z, aside from the requirement on the latter's width relative to Δt. As a consequence, it will not be necessary to consider the ion optics in detail in order to determine the distribution of the mass peak intensities, which simplifies the modeling problem significantly.

Pulse Height Distribution. The assumption that all voltage pulses are of the same magnitude can now be relaxed. The total area of a voltage pulse resulting from a single ion will be labeled Y_p and its mean and variance will be written as

$$E(Y_\mathrm{p}) = \mu_\mathrm{p} \text{ and } \operatorname{var}(Y_\mathrm{p}) = \sigma_\mathrm{p}^2 \tag{5}$$

The time discretization and intensity summation of a voltage pulse whose area is Y_p, rather than 1, as was the case in the previous subsection, can be written as the sum

$$\sum_{i=0}^{M_\mathrm{p}-1} Y_\mathrm{p} f(t_0 + i\Delta t - Z) = Y_\mathrm{p} \sum_{i=0}^{M_\mathrm{p}-1} f(t_0 + i\Delta t - Z) = Y_\mathrm{p} Y_t \tag{6}$$

that is, it is the product of the time discretization factor and the pulse height distribution. Because Y_p is clearly independent of Y_t, the mean and variance of the product $Y_\mathrm{p}Y_t$ can easily be obtained from standard formulas for the product of independent random variables as

$$E(Y_\mathrm{p}Y_t) = E(Y_\mathrm{p})E(Y_t) = \Delta t^{-1}\mu_\mathrm{p} \tag{7}$$

$$\begin{aligned} \operatorname{var}(Y_\mathrm{p}Y_t) &= E(Y_\mathrm{p})^2\operatorname{var}(Y_t) + E(Y_t)^2\operatorname{var}(Y_\mathrm{p}) + \operatorname{var}(Y_\mathrm{p}) \\ &\quad \operatorname{var}(Y_t) \\ &= \Delta t^{-1}\alpha(f, \Delta t)(\mu_\mathrm{p}^2 + \sigma_\mathrm{p}^2) - \Delta t^{-2}\mu_\mathrm{p}^2 \end{aligned} \tag{8}$$

Multiple Ion Arrivals. Equations 7 and 8 provide the mean and variance of the random variable describing the time discretization and intensity summation of the voltage pulse induced by a single ion arrival. If there are instead k ion arrivals over the summation period and their time discretization factors are $Y_t^{(1)}, Y_t^{(2)}, ..., Y_t^{(k)}$ and their pulse heights are $Y_\mathrm{p}^{(1)}, Y_\mathrm{p}^{(2)}, ..., Y_\mathrm{p}^{(k)}$, then the corresponding random variable for their superposition may be written

$$Y_\mathrm{g} = \sum_{j=1}^{j=k} Y_\mathrm{p}^{(j)} Y_t^{(j)} \tag{9}$$

This is because the voltage resolution of the digitizer is still assumed to be infinite, so that the time discretization and intensity summation of the superposition of multiple voltage pulses is a linear operation (see Figure 4).

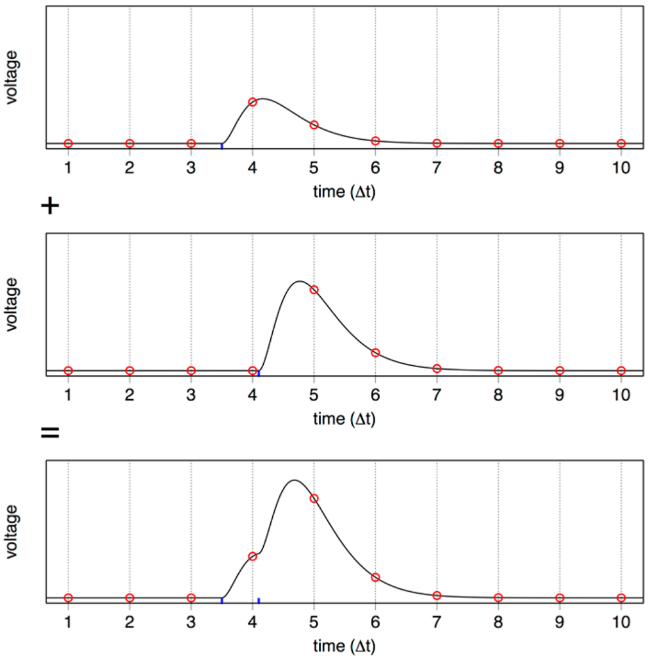


Figure 4. Illustration of the linearity of eq 9. The top two pulses have different starting times and different heights but summing all of the sampling points across both pulses gives the same result as summing the sampling points across their superposition (bottom). This generalizes for the superposition of an arbitrary number of pulses.

The mean and variance of the $Y_\mathrm{p}^{(j)}Y_t^{(j)}$ are finite and well-defined, and distinct terms can be assumed to be independent. We may therefore appeal to the CLT when k is large and approximate the distribution of Y_g by the Gaussian distribution; however, this has to be done with care. The conventional application of the CLT would suggest that if μ and σ^2 are the mean and variance of the $Y_\mathrm{p}Y_t$, then Y_g can be approximated by $N(k\mu, k\sigma^2)$. However, this is not appropriate in the present case, because k is not a fixed constant, but rather it is a random variable governed by the Poisson distribution with rate Λ, so that we are in effect adding up a random number of random variables. For this compound Poisson distribution,[20] the appropriate approximating Gaussian distribution for large Λ is $N(\mu\Lambda, (\mu^2 + \sigma^2)\Lambda)$, so that asymptotically Y_g will have the distribution

$$Y_\mathrm{g} \sim N(\Delta t^{-1}\mu_\mathrm{p}\Lambda, \Delta t^{-1}\alpha(f, \Delta t)(\mu_\mathrm{p}^2 + \sigma_\mathrm{p}^2)\Lambda) \tag{10}$$

Note that the requirement of Λ being large precludes the use of this model for many low-abundance compounds; however, this requirement will be relaxed considerably in a later subsection.

Electronic Noise. A two-component model can be used to account for the electronic noise encountered in this study. One component, Y_B, determines the noise baseline over the mass peak, and the other, Y_s, describes the high-frequency deviations from that baseline across the M_p sampling points. The high-frequency noise will primarily be due to the random thermal motion of the electrons and for the sampling rates encountered in mass spectrometry these terms will be approximately Gaussian and uncorrelated[21] and independent of other noise terms. It is difficult to develop a fully general model for the baseline electronic noise as it can be affected by electromagnetic interference from nearby electronics and potentially also by "flicker noise"[10] whose origin is not fully understood and for which general physical models have been relatively unsuccessful.[22] Ideally, electromagnetic interference should be eliminated through more careful shielding of the signal or through removal of the sources of the interference. Alternatively, if the interfering signals are composed of frequencies that are sufficiently distinct from those of the signals of interest that the latter are not significantly distorted, filtering may be employed. However, in cases where the interference cannot be removed by such means, its effects on the mass peak intensity may be accounted for reasonably well via a relatively simple model that assumes the Y_B to be independent, and having a well-defined mean and variance. In this case, the sum of the electronic noise terms over the summation period, Y_N, can be written

$$\begin{aligned} Y_N &= \sum_{i=0}^{M_p-1} (Y_B + Y_s^{(i)}) \\ &= M_p Y_B + \sum_{i=0}^{M_p-1} Y_s^{(i)} \end{aligned} \tag{11}$$

and if Y_B and Y_s have means μ_B and 0, and variances ${\sigma_B}^2$ and ${\sigma_s}^2$, respectively, then

$$E(Y_N) = E(M_p Y_B) + E(\sum_{i=0}^{M_p-1} Y_s^{(i)}) = M_p \mu_B \tag{12}$$

$$\text{var}(Y_N) = \text{var}(M_p Y_B) + \text{var}(\sum_{i=0}^{M_p-1} Y_s^{(i)}) = M_p^2 \sigma_B^2 + M_p \sigma_s^2 \tag{13}$$

In addition, if the baseline noise is Gaussian, then the sum of all these noise terms will also be Gaussian and have the mean and variance listed above. This last requirement will be relaxed later.

Quantization and Dithering. The final feature of the data acquisition system that must be modeled in order to determine the distribution of S is the quantization of the discretized voltage signal by an ADC with finite voltage resolution. Although the quantization error is fundamentally deterministic, it is possible to treat it as a further random variable, Y_q. A critical aspect of this step of the modeling problem is to ensure that the Y_q can be treated as being uncorrelated with the voltage signal as this greatly simplifies the mathematical treatment.

Let ΔV be the quantization step size by which the ADC can discriminate voltage levels. Then, provided the ADC is not saturated, the quantization error will range from $-\Delta V/2$ to $+\Delta V/2$, depending on how far the sampled voltage is from the ADC reference voltages at the sampling time. If the voltage signal varies rapidly (and aperiodically) relative to ΔV between all adjacent sampling points in the summation period, then the quantization errors can be considered uncorrelated with the voltage signal and independent of one another, and may be described by the Uniform distribution, ranging from $-\Delta V/2$ to $+\Delta V/2$.[23−25] By the properties of this uniform distribution, we therefore have

$$E(Y_q) = 0 \text{ and } \text{var}(Y_q) = \frac{1}{12}\Delta V^2 \tag{14}$$

and the statistic obtained by adding up the M_p outcomes of Y_q across the summation period has

$$E(\sum_{i=0}^{M_p-1} Y_q^{(i)}) = 0 \text{ and } \text{var}(\sum_{i=0}^{M_p-1} Y_q^{(i)}) = \frac{M_p}{12}\Delta V^2 \tag{15}$$

Because the summation period can be expected to include sampling points where the only signal present is due to the electronic noise, the requirement of a rapidly varying signal will entail a high value for σ_s. One rule of thumb[26] states that we must have $\sigma_s > \Delta V$ in order for the above model to apply.

In the case where σ_s is small relative to ΔV, several of the quantization errors are likely to be strongly correlated, as shown in Figure S-2 in the Supporting Information, and this greatly complicates the mathematical treatment of the resulting distribution. However, a method known as dithering[27] may be applied to break up the correlation in the quantization errors by adding further noise to the voltage signal prior to its quantization at each of the M_p sampling points. This generally increases the variance associated with Y_q, which may be written more generally as $d\Delta V^2$ where the value of d depends on whether and what type of dithering is applied. For example, we would have $d = 1/12$ if no dither were needed, but $d = 1/4$ if, say, a triangular dither were used.[28] Consequently, if M_p is large and the CLT can be applied, the sum of the quantization errors may be described by the Gaussian distribution $N(0, M_p d\Delta V^2)$.

Histogramming. Three types of random variables have been described so far, one due to the superposition of voltage pulses, a second due to the electronic noise, and a third due to the quantization errors. Each one of them may be regarded as Gaussian under the rather strict assumptions that Λ and M_p are large and that the baseline electronic noise is also Gaussian. If these assumptions are met, the mass peak intensity S will be the sum of three Gaussian distributions and will therefore also be Gaussian itself:

$$\begin{aligned} S &= [\sum_{j=1}^{k} Y_p^{(j)} Y_t^{(j)}] + [M_p Y_B + \sum_{i=0}^{M_p-1} Y_s^{(i)}] + [\sum_{i=0}^{M_p-1} Y_q^{(i)}] \\ &\sim N(\Delta t^{-1} \mu_p \Lambda, \Delta t^{-1} \alpha(f, \Delta t)({\mu_p}^2 + {\sigma_p}^2)\Lambda) \\ &\quad + N(M_p \mu_B, {M_p}^2 {\sigma_B}^2 + M_p {\sigma_s}^2) + N(0, M_p d\Delta V^2) \\ &\sim N(\Lambda \mu_p \Delta t^{-1} + M_p \mu_B, \Delta t^{-1} \alpha(f, \Delta t)({\mu_p}^2 + {\sigma_p}^2)\Lambda \\ &\quad + {M_p}^2 {\sigma_B}^2 + M_p {\sigma_s}^2 + M_p d\Delta V^2) \end{aligned} \tag{16}$$

The above assumptions may be relaxed considerably when determining the Gaussian approximation for H. H is obtained by summing N_p distinct outcomes of S, and this is equivalent to summing N_p outcomes of each of the three distinct types of random variables. Consequently, the CLT can be applied to the sum of voltage pulses, the sum of the electronic noise, and the sum of the quantization errors that are obtained over all N_p acquisitions. Rather than requiring that Λ and M_p are large, it will therefore suffice that ΛN_p and $N_p M_p$ are large. Because N_p will typically be at least 100, these requirements are rather mild and will usually be satisfied. If N_p is large, the CLT may also be

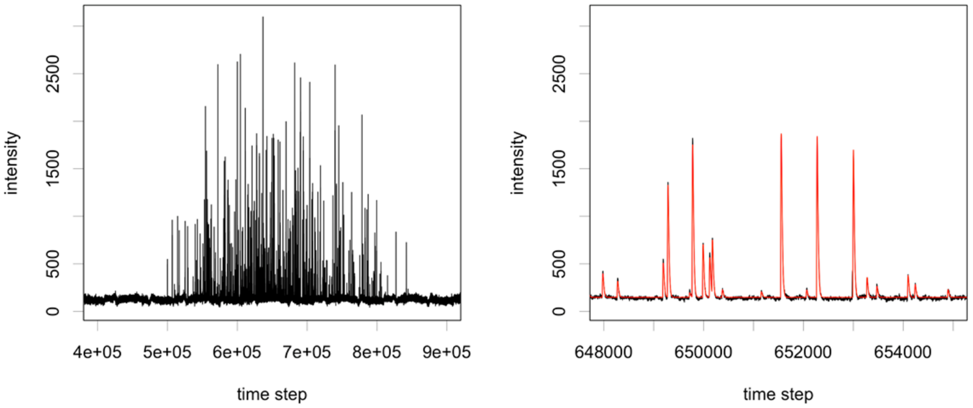


Figure 5. Left: ^{36}Ar mass peak observed via the oscilloscope, where many of the individual ion arrivals can be distinguished and the effects of the pulse height distribution are evident. Right: subsection of the same mass peak where the red segments indicate ion arrivals that were identified through an R script and to which the known functional form of $f(t)$ was fitted.

applied to the baseline noise so that its sum may be treated as Gaussian even if the individual outcomes are not. Consequently, if these assumptions are satisfied and if K is the total number of ion counts over the N_p acquisitions, the distribution of H may be written as

$$\begin{aligned} H &= \sum_{h=1}^{N_p} S^{(h)} \\ &= \sum_{h=1}^{N_p} \{[\sum_{j=1}^{k} Y_p^{(j)} Y_t^{(j)}] + [M_p Y_B + \sum_{i=0}^{M_p-1} Y_s^{(i)}] + [\sum_{i=0}^{M_p-1} Y_q^{(i)}]\}^{(h)} \\ &= [\sum_{j=1}^{K} Y_p^{(j)} Y_t^{(j)}] + [M_p \sum_{h=1}^{N_p} Y_B^{(h)} + \sum_{i=0}^{N_p M_p-1} Y_s^{(i)}] + [\sum_{i=0}^{N_p M_p-1} Y_q^{(i)}] \\ &\sim N(N_p \Lambda \mu_p \Delta t^{-1} + N_p M_p \mu_B, N_p \Delta t^{-1} \alpha(f, \Delta t)(\mu_p{}^2 + \sigma_p{}^2)\Lambda \\ &\quad + N_p(M_p{}^2 \sigma_B{}^2 + M_p \sigma_s{}^2 + M_p d \Delta V^2)) \end{aligned} \tag{17}$$

The final expression can be rewritten $N(A\Lambda + B, C\Lambda + D)$, where A, B, C, and D are constants that are determined by underlying instrumental parameters, whereas Λ is typically an unknown parameter. The model can be applied with minor modifications if Λ varies across the N_p acquisitions. Depending on the desired use of the model, some form of continuity correction may be employed because S and H can, strictly speaking, only result in values that are separated by multiples of ΔV. It should also be noted that even if the CLT is not applicable, the mean and variance used in eq 17 will still be valid, so long as the different terms can be assumed to be uncorrelated.

■ EXPERIMENTAL SECTION

Experimental data were acquired by monitoring the mass peaks induced by ^{36}Ar ions on a MAT95 sector mass spectrometer (Finnigan MAT GmbH, Bremen, Germany) operated in electron ionization positive ion mode and incorporating a discrete dynode multiplier for signal detection. ^{36}Ar ions were chosen due to the relatively low degree of interference from distinct species of ions over the required m/z range. The mass spectrometer's preamplifier was coupled to a 12-bit 1 GHz LeCroy HD04000 oscilloscope in order to allow for detailed monitoring of the voltage signals induced. 901 spectra were initially acquired, but in order to guarantee the model assumption that the mass windows should fully contain the mass peaks, 340 spectra whose mass peaks were not completely covered in the digitization period were discarded, leaving 561 mass peaks, one of which is shown in Figure 5. It is noted that for modern TOF instruments, the widths of the voltage pulses would be much narrower relative to Δt and much broader relative to the spread of the voltage pulse arrival times. The data were standardized so that $\Delta t = 1$ and $\Delta V = 1$. There was evidence of electromagnetic interference in the signal, but this could not be filtered out without significantly distorting the voltage pulses induced by incoming ions.

Due to the high speed of the oscilloscope's digitizer relative to the width of the voltage pulse shape, $f(t)$, the latter could be sampled in detail and it was determined that the functional form of the probability density function of the gamma distribution provided a very close fit to it. An R software program was developed, which was capable of automatically identifying voltage pulses induced by the incoming ions and fitting a scaled $f(t)$ to them, even allowing for some double-pulsing (see Figure 5). As the rate of ion arrivals was kept sufficiently low that the induced voltage pulses of individual ions could be distinguished, the total number of ions could be determined with a very high degree of accuracy, averaging 538.84 for the 561 recorded mass peaks. The empirical distribution of the fitted rate and shape parameters of $f(t)$ had extremely low variance, suggesting that the normalized shape of $f(t)$ was indeed constant. Therefore, the same program was run a second time with the rate and shape parameters fixed to the median values obtained from the first run, in order to minimize any effect their variability might have on the pulse height distribution. This yielded a distribution of over 3 × 10^5 recorded voltage pulse heights that could be used to approximate the true pulse height distribution.

The 561 mass peaks were randomly divided into two groups. One was used to calculate a set of S-statistics by summing the recorded intensities across the mass peaks. The other group was used to estimate the parameters of the Gaussian model. This involved estimating Λ from the identified ion counts, estimating the mean and variance of the pulse height distribution (μ_p and $\sigma_p{}^2$), as well as the electronic noise parameters μ_B, $\sigma_B{}^2$, and $\sigma_s{}^2$. $\alpha(f, \Delta t)$ was calculated numerically. Because both M_p and Λ were large and the distribution of baselines was approximately Gaussian, the Gaussian model should apply to the calculated S-statistics, meaning that if the model is correct, then applying the estimated parameters to eq

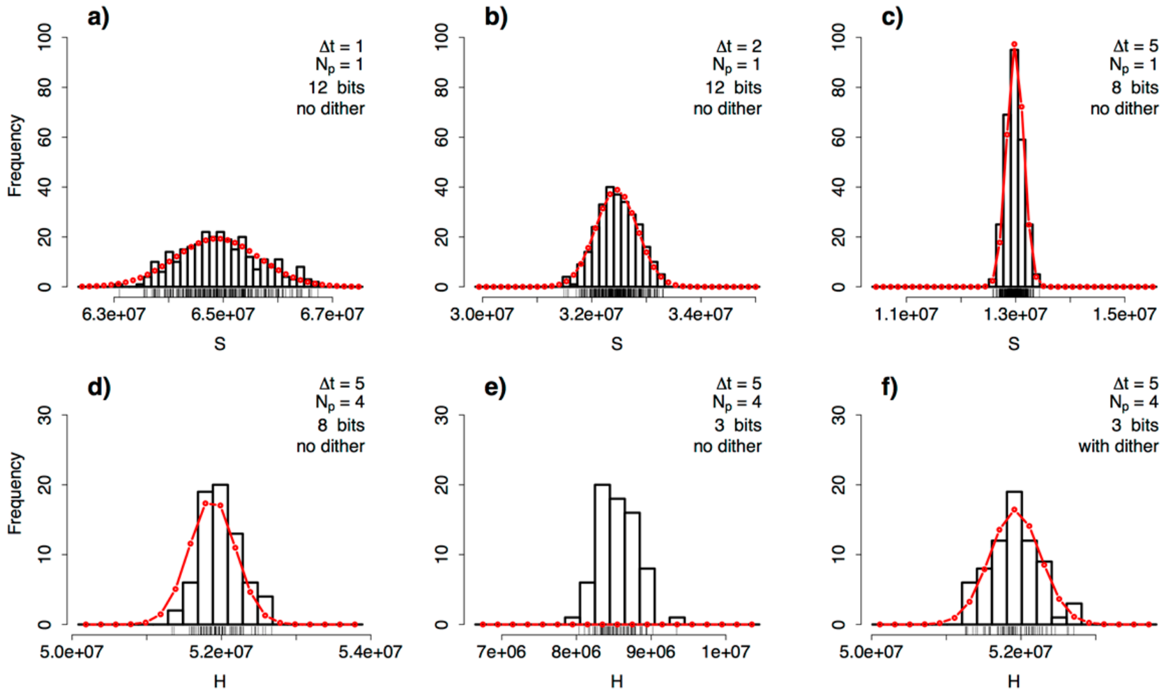


Figure 6. Distributions of the empirical mass peak intensities (black) along with the distributions predicted by the Gaussian model (red) for the six experimental setups listed in the top right corners. The different setups give rise to very distinct empirical distributions, but the Gaussian model is able to predict these with good accuracy. When the voltage resolution is reduced to 3 bits in plot e, the undithered model breaks down, but if a dither is applied as in plot f, the model is again applicable.

16, should result in a Gaussian distribution that approximates the empirical distribution of the *S*-statistics. This empirical distribution is plotted alongside the predicted theoretical distribution in the top left segment of Figure 6. Despite the parameters used in the Gaussian model being estimates, rather than the true values, the fit is very close.

To assess more broadly whether the Gaussian model properly accounts for the effects of the time discretization, voltage quantization, dithering and histogramming, the above procedure was repeated for data sets that were altered to reflect changes to the relevant instrumental parameters. For example, the data set obtained by discarding the recorded voltage of every other time-step corresponds to what would be observed on an instrument with $\Delta t = 2$, i.e., with half the time resolution of the oscilloscope used. Similarly, discarding periodically every two out of three time-steps corresponds to $\Delta t = 3$, etc. It is also straightforward to determine the data that would have been obtained for an instrument with 1, ..., 11 bit voltage resolution, by rounding the recorded voltages appropriately. It is to be expected that an undithered Gaussian model will break down as the voltage resolution is lowered, however the above theory predicts that it can be restored through the application of a triangular dither. This can easily be simulated by adding pseudorandomly generated numbers to the 12 bit data, prior to reducing the voltage resolution. *H*-statistics were also obtained by summing sets of *S*-statistics. The resulting empirical distributions are shown in Figure 6 along with the Gaussian distributions predicted by eqs 16 and 17. Additional examples are provided in the Supporting Information (Figure S-3) where quantile-quantile plots against the predicted Gaussian distributions are also shown along with the outcomes of Kolmogorov—Smirnov tests (Figures S-4 and S-5, Supporting Information). Despite the diverse distributions produced by the altered instrumental settings, the theoretical distributions predicted by the Gaussian model remain in close agreement with empirical ones, provided a dither is applied when the voltage resolution is low, which is encouraging given the high level of detail with which the model has been formulated.

DISCUSSION

The Gaussian model can provide some useful insights into the performance of the mass spectrometers to which it is applicable. If Λ is large enough that the electronic noise and quantization errors are negligible, the signal-to-noise ratio can be approximated as

$$\frac{\mu_H}{\sigma_H} \approx \frac{N_p \Lambda \mu_p \Delta t^{-1}}{\sqrt{N_p \Delta t^{-1} \alpha(f, \Delta t)(\mu_p^{\ 2} + \sigma_p^{\ 2})\Lambda}} = \sqrt{\left[\frac{1}{\Delta t \alpha(f, \Delta t)}\right]\left[\frac{\mu_p^{\ 2}}{\mu_p^{\ 2} + \sigma_p^{\ 2}}\right][N_p \Lambda]} \tag{18}$$

The terms in the first two square brackets tend to 1 from below as Δt and σ_p tend to zero, i.e., as the digitizer speed tends to infinity and the pulse height distribution approaches a δ-function. Consequently, because $N_p\Lambda$ is the total rate of ion arrivals over all N_p acquisitions, the characteristic Poissonian signal-to-noise ratio that can be encountered in the analysis of unsaturated pulse counting data is recovered in this limit, as might be expected. More generally, the above signal-to-noise ratio does increase in proportion to the square root of the total rate of ion arrivals, but only as a fraction of the Poissonian signal-to-noise ratio, as determined by the square root of the first two terms.

To verify this prediction, the signal-to-noise ratios of the *H*-statistics suggested by eq 18 were calculated along with those obtained from the experimental data for N_p = 1−40 (thereby covering a broad range of ion arrival rates) and for Δt = 1, 100, and 1000, thus altering the value of $\Delta t\alpha(f, \Delta t)$. No dither was applied, as the original 12-bit voltage resolution was used. To minimize the electronic noise, the baselines of the spectra used to calculate the H-statistics were first subtracted, and in order to increase the available sample size for higher values of N_p, the full set of 561 spectra were used both for the calculation of the empirical signal-to-noise ratios and for the estimation of the parameters from which the theoretical signal-to-noise ratios were derived. The resulting theoretical and empirical signal-to-noise ratios are in close agreement, as illustrated in Figure 7.

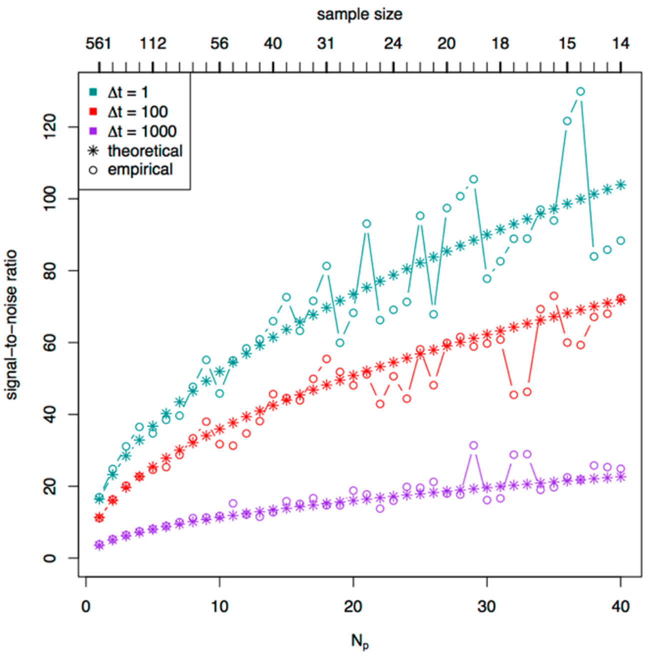


Figure 7. Empirical signal-to-noise ratios for H-statistics along with the values predicted by eq 18. The sample size is lower for H-statistics with higher values of N_p since these require more spectra each.

The most pragmatic manner in which convergence to the Poissonian limit on the signal-to-noise ratios may be sought can be assessed in detail based on the mean and variance in eq 17, and doing so may be very instructive in planning instrument design. Clearly, it is desirable to minimize the variance by reducing the electronic noise, although reductions in σ_s below ΔV will not be helpful if the Gaussian model is to be employed. Not surprisingly, the variance of the pulse height distribution should be reduced as far as possible while the mean should be increased although this must be balanced against the risk of saturating the ADC. It should be noted that simply increasing the electron multiplier gain may be counterproductive depending on the relationship between μ_p and σ_p inherent to the electron multiplier. If we have $\sigma_p \propto (\mu_p)^r$, then according to eq 18, increasing μ_p will have no effect on the signal-to-noise ratio for $r = 1$, whereas it will decrease it for $r > 1$ and increase it for $r < 1$.

The signal-to-noise ratio might also be improved by reducing M_p, e.g., by reducing the width of the underlying mass peak or the width of $f(t)$, but again this must be balanced against the risk of saturating the ADC. Although the time discretization factor's contribution to the variance will often be modest compared with that of the pulse height distribution, methods of minimizing $\Delta t\alpha(f, \Delta t)$ by appropriate pulse shaping should also be examined. For example, if $f(t)$ is a rectangular function of length Δt, the variance of Y_t will be zero, and $\Delta t\alpha(f, \Delta t)$ will equal 1 in eq 18. Smoother, but broader functions with this property can also easily be devised. It is noteworthy that so far as the signal-to-noise ratio of the mass peak intensity is concerned, such pulse shapes would obviate the need for minimizing Δt through the use of faster (and more expensive) digitizers. It is also worth noting that substantial sharpening of $f(t)$ risks reducing the signal-to-noise ratio.

Before relying on these predictions, whether for the purpose of MS data analysis or for planning instrument design, engineers should take steps to ensure that the Gaussian model is indeed applicable. The model requires that the induced currents and voltages are within the linear range of the electron multiplier and the subsequent circuitry which must be designed with care to prevent any substantial dependence between $f(t)$ and the signal strength. Such consideration may require particular attention for TOF instruments, whose high-speed circuitry can make them more susceptible to problems such as ringing and crosstalk. Furthermore, as noted earlier, the model employed for the baseline electronic noise does not fully account for interfering frequencies. Ideally, the sources of such interference should be identified and eliminated, or the circuitry shielded from them. Alternatively, it may be possible to account for the effects of the interferences through more elaborate signal processing. Finally, it is important to keep the voltage signal within the range of the ADC as far as possible, including the noise baseline, which is in contrast to the commonly used technique of suppressing the electronic noise by keeping the lowest level of the digitizer slightly above the baseline.[29] While the inclusion of the baseline would make it difficult to distinguish signals with low ion counts from the electronic noise, baseline suppression is also problematic as it is an inherently nonlinear operation that is especially distortive to weak signals.

Aside from its utility to instrument design, the Gaussian model should prove useful to informaticians as methods of data analysis that incorporate precise knowledge of the data's statistical distribution can make far better use of the information in the data and can attain a far greater degree of statistical rigor than heuristic alternatives. Such methods of data analysis will require the development of protocols for determining the parameters of the Gaussian model, and ideally this will include methods for determining the mean and variance of the pulse height distribution over a broad range of ion masses and charge states. The development of such protocols and associated methods of data analysis will be examined in future studies, as will be the extension of the model to the analysis of mass peak centroids. Quite clearly, the Gaussian model also provides an efficient and accurate method of simulating mass peak intensities. Furthermore, the model may prove applicable beyond mass spectrometry, because many scientific instruments employ a somewhat similar setup incorporating electron multipliers and ADCs.

■ CONCLUSION

The Gaussian model appears to be the first detailed model for the distribution of mass peak intensities of ADC-digitized MS data acquired via electron multipliers to have been derived essentially from first principles. This is somewhat surprising given the widespread use of such instruments and the heavy reliance that modern biology places on the platform. There may have been a presumption that it was only feasible to model the MS data in broad strokes based on their general qualitative features due to the complexity of the instruments. This view would also justify the widespread use of heuristic methods of data analysis that are based on rules of thumb, rather than first principles, and the associated tendency of treating the data generation process as something of a black box. But although our understanding of MS data remains far from complete, even with the Gaussian model, its derivation does demonstrate the feasibility of taking a more rigorous approach to MS data modeling, with potentially significant implications for both instrument design and data analysis.

■ ASSOCIATED CONTENT

S Supporting Information

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author

*A. Ipsen. E-mail: a.ipsen@swansea.ac.uk.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The author thanks Gareth Brenton for valuable discussions, and for assistance with the experimental work. The author is supported by an MRC fellowship in biomedical informatics (Grant No. MR/J013994/1).

■ REFERENCES

(1) Anderle, M.; Roy, S.; Lin, H.; Becker, C.; Joho, K. *Bioinformatics* **2004**, *20*, 3575−3582.

(2) Schulz-Trieglaff, O.; Pfeifer, N.; Gröpl, C.; Kohlbacher, O.; Reinert, K. *BMC Bioinf.* **2008**, *9*, 423.

(3) Du, P.; Stolovitzky, G.; Horvatovich, P.; Bischoff, R.; Lim, J.; Suits, F. *Bioinformatics* **2008**, *24*, 1070−1077.

(4) Hundertmark, C.; Fischer, R.; Reinl, T.; May, S.; Klawonn, F.; Jänsch, L. *Bioinformatics* **2009**, *25*, 1004−1011.

(5) Morris, J. S.; Coombes, K. R.; Koomen, J.; Baggerly, K. A.; Kobayashi, R. *Bioinformatics* **2005**, *21*, 1764−1775.

(6) Rocke, D. M.; Lorenzato, S. *Technometrics* **1995**, *37*, 176−184.

(7) Yang, C.; Yu, W. In 2011 *Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, Pacific Grove, CA, November 6−9, 2011; pp 1036−1040.

(8) Gedcke, D. A. *How Counting Statistics and the ADC Sampling Interval Control Mass Accuracy in Time of Flight Mass Spectrometry*; Application Note AN61; ORTEC: Oak Ridge, TN.

(9) Harris, F. M.; Trott, G. W.; Morgan, T. G.; Brenton, A. G.; Kingston, E. E.; Beynon, J. H. *Mass Spectrom. Rev.* **1984**, *3*, 209−229.

(10) Shin, H.; Mutlu, M.; Koomen, J. M.; Markey, M. K. *Cancer Inform.* **2007**, *3*, 219−230.

(11) Coates, P. B. *Rev. Sci. Instrum.* **1992**, *63*, 2084−2088.

(12) Stephan, T.; Zehnpfenning, J.; Benninghoven, A. *J. Vac. Sci. Technol., A* **1994**, *12*, 405−410.

(13) Ipsen, A.; Ebbels, T. M. D. *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 779−791.

(14) Lee, H.-N.; Marshall, A. G. *Anal. Chem.* **2000**, *72*, 2256−2260.

(15) Ipsen, A.; Want, E. J.; Lindon, J. C.; Ebbels, T. M. D. *Anal. Chem.* **2010**, *82*, 1766−1778.

(16) Kimmel, J. R.; Yoon, O. K.; Zuleta, I. A.; Trapp, O.; Zare, R. N. *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 1117−1130.

(17) Lombard, F. J.; Martin, F. *Rev. Sci. Instrum.* **1961**, *32*, 200−201.

(18) Dietz, L. A. *Rev. Sci. Instrum.* **1965**, *36*, 1763−1770.

(19) Guilhaus, M. *J. Mass Spectrom.* **1995**, *30*, 1519−1532.

(20) Boland, P. J. *Statistical and Probabilistic Methods in Actuarial Science*; CRC Press: Boca Raton, FL, 2007.

(21) Gardner, W. A. *Introduction to Random Processes, With Applications to Signals and Systems*; Macmillan Publishing Co.: New York, 1986.

(22) Dutta, P.; Horn, P. M. *Rev. Mod. Phys.* **1981**, *53*, 497−516.

(23) Widrow, B. *Trans. AIEE, Part II: Appl. Ind.* **1961**, *79*, 555−568.

(24) Sripad, A. B.; Snyder, D. *IEEE Trans. Acoust., Speech, Signal Process.* **1977**, *25*, 442−448.

(25) Marco, D.; Neuhoff, D. L. *IEEE Trans. Inf. Theory* **2005**, *51*, 1739−1755.

(26) Widrow, B.; Kollar, I.; Liu, M.-C. *IEEE Trans. Instrum. Meas.* **1996**, *45*, 353−361.

(27) Lipshitz, S. P.; Wannamaker, R. A.; Vanderkooy, J. *J. Audio Eng. Soc.* **1992**, *40*, 355−375.

(28) Wannamaker, R. A.; Lipshitz, S.; Vanderkooy, J.; Wright, J. N. *IEEE Trans. Signal Process.* **2000**, *48*, 499−516.

(29) *Diving Deep into Single Ion Counting with FastFlight Digital Signal Averager*; Application Note AN53; ORTEC: Oak Ridge, TN.

DOI: 10.1021/ac503554u